REQUIREMENTS

- Omitting the names of team members will not be resolved
- Late submissions will result in a score deduction
- Each team elective from at least 8 (eight) to a maximum 12 (twelve) students
- *Exercise* on file Microsoft Word. Each question including request and solution by bilinguals English and Vietnamese.
- Number of team is set by the teacher

Team number 1 do exercise in order 1, 11, 21, 31, 41, …

Team number 2 do exercise in order 2, 12, 22, 32, 42, …

Team number 3 do exercise in order 3, 13, 23, 33, 43, …

INSTRUCTIONS FOR EXPRESSING PAGE 1

SCHOOL OF MEDICINE – VIETNAM NATIONAL UNIVERSITY – HCMC

TEAM - HOMEWORK

SUBJECT PROBABILITY & STATICTISC

Class ………., School year: …………

Team members : (*follow ABC*)

1. Nguyễn Văn A
2. Lê Thị B

………..………..

Probability & Statistics in Medicine
TEAM HOMEWORK

Data and probability

1.1 Empirical distributions

Let $(x_1, \dots, x_n)$ be a sample, *i.e.* a series of numerical values for a certain variable in a set of n individuals.

- The *modalities* are the different values.
- The *empirical mean* is $\overline{x} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i$.
- The *empirical variance* is $s_x^2 = \left(\dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i^2 \right) - \overline{x}^2$.
- The *empirical standard deviation* is the square root of the empirical variance.
- A sample is *centered and reduced* if its mean is 0 and its variance 1. In order to *center and reduce* a sample, substract the mean from each modality, then divide by the standard deviation.
- The *empirical frequency* of an interval is the ratio of the number of values in that interval, to the total number of individuals.
- The *median* is the smallest modality such that at least 50% of the values are smaller or equal.
- The *lower quartile* is the smallest modality such that at least 25% of the values are smaller or equal.
- The *upper quartile* is the smallest modality such that at least 75% of the values are smaller or equal.
- A statistical character is considered as *continuous* when (almost) all values are different. When for most modalities, several individuals have the same value, the character is *discrete.*

Example. Here are numbers by age of non-smoking mothers at delivery.

age	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
number	7	8	9	10	12	3	2	5	4	5	2	4	2	0	1

1. What are the modalities ?

   The modalities are the whole numbers between 21 *and* 35*.*

2. Is this a discrete or a continuous variable ?

   Given the precision of the data, several individuals have the same modality (are considered as having the same age). Thus it is a discrete variable.

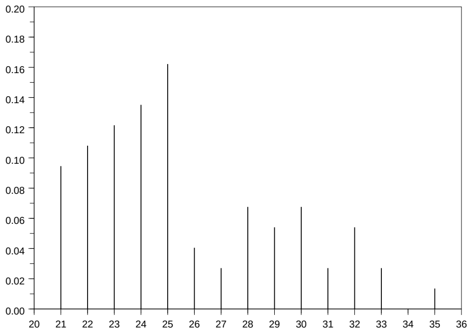
3. Find the empirical frequencies of the modalities.

   To get the empirical frequencies, divide the numbers by the total number of individuals, which is 74.

age	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
frequency	$\frac{7}{74}$	$\frac{8}{74}$	$\frac{9}{74}$	$\frac{10}{74}$	$\frac{12}{74}$	$\frac{3}{74}$	$\frac{2}{74}$	$\frac{5}{74}$	$\frac{4}{74}$	$\frac{5}{74}$	$\frac{2}{74}$	$\frac{4}{74}$	$\frac{2}{74}$	$\frac{0}{74}$	$\frac{1}{74}$
rounded freq.	0.095	0.108	0.122	0.135	0.162	0.041	0.027	0.068	0.054	0.068	0.027	0.054	0.027	0	0.014

4. Represent the empirical frequencies on a bar chart.

   The bar chart consists of drawing a vertical segment above each modality, with height proportional to the number or to the empirical frequency.



5. Find the empirical mean, variance, and standard deviation of the sample.

   For the empirical mean:

$$\overline{x} = \frac{1}{74}\Big(7\times 21 + 8\times 22 + \cdots + 0\times 34 + 1\times 35\Big) = 25.662\ .$$

   The average age in this sample is approximately 25 *years and* 8 *months.*

   For the empirical variance:

$$s_x^2 = \frac{1}{74}\Big(7\times 21^2 + 8\times 22^2 + \cdots + 0\times 34^2 + 1\times 35^2\Big) - (25.662)^2 = 12.683\ .$$

The standard deviation is the square root of the variance:

$$s_x = \sqrt{12.683} = 3.561\ ,$$

that is approximately 3 *years and* 7 *months.*

6. Find the values of the empirical distribution function.

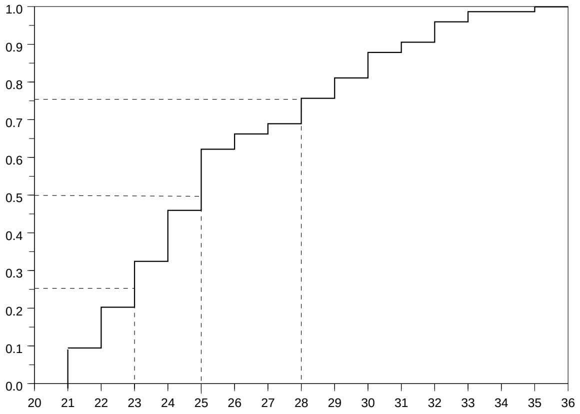
   The values of the empirical distribution function are the cumulated sums of frequencies.

age	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
cum. freq.	$\frac{7}{74}$	$\frac{15}{74}$	$\frac{24}{74}$	$\frac{34}{74}$	$\frac{46}{74}$	$\frac{49}{74}$	$\frac{51}{74}$	$\frac{56}{74}$	$\frac{60}{74}$	$\frac{65}{74}$	$\frac{67}{74}$	$\frac{71}{74}$	$\frac{73}{74}$	$\frac{73}{74}$	$\frac{74}{74}$
rounded	0.095	0.203	0.324	0.459	0.622	0.662	0.689	0.757	0.811	0.878	0.905	0.959	0.986	0.986	1

7. What is the empirical frequency of the interval [22 ; 25]?

   It is the sum of empirical frequencies for the modalities 22, 23, 24, 25, *or else the increment of the empirical distribution function* $F(25) - F(21)$, *that is* $39/74 \simeq 0.527$. *More than half of the women in the sample are between* 22 *and* 25 *years old.*

8. Draw a graphical representation of the empirical distribution function. Determine from the graph the median and the quartiles of the sample.



The median is 25 *years; the first quartile is* 23 *years, the last quartile is* 28 *years.*

9. Compare the mean with the median, then the standard deviation with the distances between the median and the quartiles.

   The mean is larger than the median, which is normal for a distribution skewed to the right. For the same reason, the gap between the last quartile and the median is larger than that between the median and the first quartile. Both are lower than the standard deviation: this is the case for most distributions, whether they are symmetrical or skewed.

1. Here are numbers by age of smoking birth mothers at delivery.

age	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
number	5	5	4	3	3	5	1	4	3	2	3	2	1	1	1

1. What are the modalities ?

2. Is this a discrete or a continuous variable?

3. Find the empirical frequencies of the modalities.

4. Represent the empirical frequencies on a bar chart.

5. Find the empirical mean, variance, and standard deviation of the sample.

6. Find the values of the empirical distribution function.

7. What is the empirical frequency of the interval [22 ; 25] ?

8. Draw a graphical representation of the empirical distribution function. Determine from the graph the median and the quartiles of the sample.

9. Compare the mean with the median, then the standard deviation with the distances between the median and the quartiles.

2. Consider the sample $(1, 0, 2, 1, 1, 0, 1, 0, 0)$.

1. What is its empirical mean?

2. What is its empirical variance?

3. Center and reduce this sample.

4. If you had to propose a model for these data: would you choose a discrete or a continuous model?

3. Consider the sample

$$(1.2, 0.2, 1.6, 1.1, 0.9, 0.3, 0.7, 0.1, 0.4)\ .$$

1. What is its empirical mean?

2. What is its empirical variance?

3. Center and reduce this sample.

4. If you had to propose a model for these data: would you choose a discrete or a continuous model?

1.2 Probabilities and conditional probabilities

- The *probability of an event* in a population is the proportion of individuals for which the event is true.

- The *conditional probability of A knowing B*, is the proportion of individuals for which A is true *among those for which B is also true.* It is the ratio of the probability of "A and B" to the probability of B:

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \text{ and } B]}{\mathbb{P}[B]} \; .$$

- The *total probability formula* gives the probability of an event A as a function of the conditional probabilities knowing another event B and its contrary $\overline{B}$:

$$\mathbb{P}[A] = \mathbb{P}[A \mid B]\,\mathbb{P}[B] + \mathbb{P}[A \mid \overline{B}]\,\mathbb{P}[\overline{B}] \; .$$

- The *Bayes formula* exchanges the order of conditional probabilities:

$$\mathbb{P}[B \mid A] = \frac{\mathbb{P}[A \mid B]\,\mathbb{P}[B]}{\mathbb{P}[A \mid B]\,\mathbb{P}[B] + \mathbb{P}[A \mid \overline{B}]\,\mathbb{P}[\overline{B}]} \; .$$

Example. In a sheep breeding farm, an estimated 30% of the sheep suffer from a certain disease. A test for this disease is available. If a sheep is not ill, it has 9 chances out of 10 to react negatively to the test; if it is ill, it has 8 chances out of 10 to have a positive reaction. All the sheep in the farm are submitted to the test.

Throughout the exercise, the event "the sheep is ill" will be denoted by M, and the event "the sheep has a positive reaction" by T. The text gives:

$$\mathbb{P}[M] = 0.3 \; , \quad \mathbb{P}[\overline{T} \mid \overline{M}] = 0.9 \; , \quad \mathbb{P}[T \mid M] = 0.8 \; .$$

1. What is the probability for a sheep in that farm not to be ill?

$$\mathbb{P}[\overline{M}] = 1 - \mathbb{P}[M] = 1 - 0.3 = 0.7 \; .$$

2. What is the conditional probability for a sheep to have a positive reaction knowing that it is not ill?

$$\mathbb{P}[T \mid \overline{M}] = 1 - \mathbb{P}[\overline{T} \mid \overline{M}] = 1 - 0.9 = 0.1 \ .$$

3. What is the probability for a sheep not to be ill and have a positive reaction?

$$\mathbb{P}[T \text{ and } \overline{M}] = \mathbb{P}[T \mid \overline{M}]\, \mathbb{P}[\overline{M}] = 0.1 \times 0.7 = 0.07 \ .$$

4. What proportion of the sheep have a positive reaction?

   Use the formula of total probabilities or compute it straight away, by distinguishing among those sheep reacting positively, those which are ill from those which are not.

$$\begin{aligned} \mathbb{P}[T] &= \mathbb{P}[T \text{ and } M] + \mathbb{P}[T \text{ and } \overline{M}] \\ &= \mathbb{P}[T \mid M]\, \mathbb{P}[M] + \mathbb{P}[T \mid \overline{M}]\, \mathbb{P}[\overline{M}] \\ &= 0.8 \times 0.3 + 0.1 \times 0.7 = 0.24 + 0.07 = 0.31 \ . \end{aligned}$$

5. What is the probability for a sheep to be ill, knowing that it has reacted positively?

   Use the Bayes formula or prove it again as follows.

$$\begin{aligned} \mathbb{P}[M \mid T] &= \frac{\mathbb{P}[T \text{ and } M]}{\mathbb{P}[T]} \\ &= \frac{\mathbb{P}[T \mid M]\, \mathbb{P}[M]}{\mathbb{P}[T \mid M]\, \mathbb{P}[M] + \mathbb{P}[T \mid \overline{M}]\, \mathbb{P}[\overline{M}]} \\ &= \frac{0.8 \times 0.3}{0.8 \times 0.3 + 0.1 \times 0.7} \simeq 0.774 \ . \end{aligned}$$

6. What is the probability for a sheep not to be ill, knowing it has reacted negatively?

   Use the Bayes formula or prove it again as follows.

$$\begin{aligned} \mathbb{P}[\overline{M} \mid \overline{T}] &= \frac{\mathbb{P}[\overline{T} \text{ and } \overline{M}]}{\mathbb{P}[\overline{T}]} \\ &= \frac{\mathbb{P}[\overline{T} \mid \overline{M}]\, \mathbb{P}[\overline{M}]}{\mathbb{P}[\overline{T} \mid \overline{M}]\, \mathbb{P}[\overline{M}] + \mathbb{P}[\overline{T} \mid M]\, \mathbb{P}[M]} \\ &= \frac{0.9 \times 0.7}{0.9 \times 0.7 + 0.2 \times 0.3} \simeq 0.913 \ . \end{aligned}$$

4. There are three sorts of a given plant: early, normal, and late. It can also be either dwarf or tall. In a sample of plants grown from 1000 seeds, there are 600 dwarf, 200 late, 300 early dwarf, 250 normal tall, 100 late tall. Consider the plant grown from a seed taken at random.

1. What is the probability that it is early? normal? late? dwarf? tall?
2. A dwarf plant is observed. What is the probability that it is early? normal? late?
3. A tall plant is observed. What is the probability that it is early? normal? late?
4. A late plant is observed. What is the probability that it is dwarf? tall?

5. In a batch of manufactured items, 5% are faulty. The items are checked, but the checking is not perfect. If the item is good, it is accepted with probability 0.96; if it is faulty, it is rejected with probability 0.98. An item is chosen at random, then checked.

1. What is the probability that this item is rejected?
2. What is the probability that it is good, knowing that it has been rejected?
3. What is the probability that it is faulty, knowing that it has been accepted?
4. What is the probability that there is an error in the checking (the item is good and rejected or bad and accepted)?

6. Here are the percentages of the different blood types in France.

Group	O	A	B	AB
Factor				
Rhesus +	37.0	38.1	6.2	2.8
Rhesus –	7.0	7.2	1.2	0.5

1. Determine the probability distribution of the four groups O, A, B, AB in the French population.
2. Determine the probability distribution of the four groups among the persons with positive rhesus
3. Determine the probability distribution of the four groups among the persons with negative rhesus
4. If a person of group O is chosen at random, what is the probability that he/she has a negative rhesus? Same question for a person of group B.

1.3 Binomial distribution

- When n *experiments are repeated independently*, the random variable X equal to the number of realizations of a given *event of probability* p, follows the *binomial* distribution with parameters n and p.
- The variable X may take all integer values between 0 and n.
- For any integer k between 0 and n, the variable X takes value k with probability:

  $$\mathbb{P}[X = k] = \binom{n}{k} p^k \, (1-p)^{n-k} \ ,$$

  where

  $$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n \times (n-1) \cdots \times (n-k+1)}{k \times (k-1) \cdots \times 3 \times 2 \times 1}$$

  is the number of ways to choose k objects among n.
- The expectation of X is np, its variance is $np(1-p)$.

Example. From past experience, it is known that a certain surgery has a 90% chance to succeed. This surgery is going to be performed on 5 patients. Let X be the random variable equal to the number of successes out of the 5 attempts.

1. What model do you propose for X?

   Assuming that the outcome (success or failure) of the 5 attempts are independent, the number of successes follows the binomial distribution with parameters 5 *and* 0.9*. The random variable X takes its values in the set* $\{0, 1, 2, 3, 4, 5\}$*, and for any integer k in that set:*

   $$\mathbb{P}[X = k] = \binom{5}{k} 0.9^k \, 0.1^{5-k} \ .$$

2. What is the probability that the surgery will fail all 5 times?

   $$\mathbb{P}[X = 0] = 0.1^5 = 0.00001 \ .$$

3. What is the probability for the surgery to fail exactly 3 times?

   $$\mathbb{P}[X = 2] = \binom{5}{2} 0.9^2 \, 0.1^3 = 0.0081 \ .$$

4. What is the probability for the surgery to succeed at least 3 times?

$$\begin{aligned}\mathbb{P}[X \geqslant 3] &= \mathbb{P}[X=3]+\mathbb{P}[X=4]+\mathbb{P}[X=5]\\ &= \binom{5}{3}0.9^3\,0.1^2+\binom{5}{4}0.9^4\,0.1^1+\binom{5}{5}0.9^5\,0.1^0\\ &= 0.0729+0.32805+0.59049=0.99144\;.\end{aligned}$$

7. When a hunter aims at a helpless rabbit, he has 1 chance out of 10 to hit it.

1. Two hunters aim independently at the same rabbit. Find the probability that:
   (a) neither of them hit;
   (b) only one of them hits;
   (c) both hunters hit.
2. Four hunters aim independently at the same rabbit.
   (a) What is the probability distribution of the number of shots suffered by the poor animal? Give the expectation and variance of that distribution.
   (b) What is the probability that the rabbit is hit at most twice?
   (c) What is the probability that the rabbit is hit at least twice?
3. Ten hunters aim independently at the same rabbit.
   (a) What is the probability for the rabbit not to be hit?
   (b) What is the probability that the rabbit becomes inedible (if it has received at least 5 shots).

8. At an identification session, 6 witnesses are asked to identify a murderer among 4 suspects, including yourself.

1. If each one of the 6 witnesses chooses at random, what are your chances:
   (a) of not being pointed out?
   (b) of being pointed out exactly once?
   (c) of being pointed out twice or more?
2. It turns out that 2 of the 6 witnesses have identified you as the murderer. Referring to 1 (c), do you expect that the judge will think that this may be due to chance?
3. What if 4 of the 6 witnesses have identified you?

1.4 Hypergeometric distribution

- In a set of N elements, among which m have been marked, n distinct elements are selected at random. The random variable X, equal to the number of marked elements among the selected n, follows the *hypergeometric distribution with parameters N, m, n*.

- In the case where $n \leq m$ and $n \leq N - m$, X may take all integer values between 0 and n.

- For any integer k between 0 and n, X takes value k with probability:

$$\mathbb{P}[X = k] = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} .$$

- The expectation of X is nm/N.

Example. There are 18 girls and 11 boys in a certain group of students. A sample of 5 persons is chosen at random in that group. Let X be the random variable equal to the number of girls in that sample.

1. What model do you propose for X?

   The distribution of X is the hypergeometric distribution with parameters $N = 29$ (total number of persons), $m = 18$ (the "marked" individuals are the girls), and $n = 5$ (the size of the sample). The values are the integers between 0 and 5. For any integer $k = 0, 1 \ldots, 5$:

$$\mathbb{P}[X = k] = \frac{\binom{18}{k}\binom{11}{5-k}}{\binom{29}{5}} .$$

2. Give the expectation of X.

   The expectation of X is $5 \times 18/29 \simeq 3.1$. It is the size of the sample, multiplied by the proportion of girls in the group.

3. Find the probability of having only girls in the sample.

$$\mathbb{P}[X = 5] = \frac{\binom{18}{5}}{\binom{29}{5}} \simeq 0.072 .$$

4. Find the probability of having at least one girl in the sample.

   The value of $\mathbb{P}[X \geqslant 1]$ must be calculated. It could be done as $\mathbb{P}[X = 1] + \mathbb{P}[X =$

$2] + \mathbb{P}[X = 3] + \mathbb{P}[X = 4] + \mathbb{P}[X = 5]$, *but it is quicker to calculate* $1 - \mathbb{P}[X = 0]$, *which is the same:*

$$\mathbb{P}[X \geqslant 1] = 1 - \mathbb{P}[X = 0] = 1 - \frac{\binom{11}{5}}{\binom{29}{5}} \simeq 0.996 \ .$$

5. Find the probability for the sample to have exactly 3 girls.

$$\mathbb{P}[X = 3] = \frac{\binom{18}{3}\binom{11}{2}}{\binom{29}{5}} \simeq 0.378 \ .$$

9. In each of the following situations, give the probability distribution of the random variable X and its expectation. Find the probability for X to be 0, then the probability for X to be 2 or more.

1. At a card table, 8 cards are delt to each of the 4 players, out of a deck of 32. Let X be the number of aces received by a given player.

2. At a belote table, the four players make teams of two. Let X be the number of diamonds of a given team.

3. At a bridge table, thirteen cards are handed out to each of the four players. Let X be the number of figures (jack, queen, or king) of a given player.

4. On a loto card, you ticked 6 numbers out of an array of 49. Let X be the number of good numbers ticked on your card.

1.5 Normal distribution

- If no software is available, the following data for the normal distribution with mean 0 and variance 1, denoted by $\mathcal{N}(0, 1)$, are given in the tables:

  ⋆ the values of the distribution function F: for a value of x, the table gives the probability $p = P[X \leqslant x] = F(x)$.

  ⋆ the values of the quantile function $F^{-1}(p)$: for a probability p the table gives the value $x = F^{-1}(p)$ such that $p = \mathbb{P}[X \leqslant x]$.

- The density of the $\mathcal{N}(0, 1)$ distribution is symmetric:

$$\mathbb{P}[X \leqslant -x] = \mathbb{P}[X \geqslant x] \ .$$

- If a random variable X follows the $\mathcal{N}(\mu, \sigma^2)$ distribution, then $(X-\mu)/\sqrt{\sigma^2}$ follows the $\mathcal{N}(0,1)$ distribution. Thus:

$$\begin{aligned}\mathbb{P}[a \leqslant X \leqslant b] &= P\left[\frac{a-\mu}{\sqrt{\sigma^2}} \leqslant \frac{X-\mu}{\sqrt{\sigma^2}} \leqslant \frac{b-\mu}{\sqrt{\sigma^2}}\right] \\ &= F\left(\frac{b-\mu}{\sqrt{\sigma^2}}\right) - F\left(\frac{a-\mu}{\sqrt{\sigma^2}}\right)\,,\end{aligned}$$

  where F is the distribution function of the $\mathcal{N}(0,1)$.

- If X and Y are two independent random variables, with respective distributions $\mathcal{N}(\mu_x, \sigma_x^2)$ and $\mathcal{N}(\mu_y, \sigma_y^2)$, then $X+Y$ follows the $\mathcal{N}(\mu_x+\mu_y, \sigma_x^2+\sigma_y^2)$ and $X-Y$ follows the $\mathcal{N}(\mu_x-\mu_y, \sigma_x^2+\sigma_y^2)$.

Exeample. The height X of men in France is modeled by a normal distribution $\mathcal{N}$ $(172, 196)$ (unit: cm).

1. What proportion of French men are less than 160 cm tall?

$$\mathbb{P}[X < 160] = \mathbb{P}\left[\frac{X-172}{\sqrt{196}} < \frac{160-172}{\sqrt{196}}\right] = F(-0.857) = 1 - F(0.857) = 0.1957\,,$$

   where F denotes the distribution function of the $\mathcal{N}(0,1)$ distribution.

2. What proportion of French men are more than two meters tall?

$$\mathbb{P}[X > 200] = \mathbb{P}\left[\frac{X-172}{\sqrt{196}} > \frac{200-172}{\sqrt{196}}\right] = 1 - F(2) = 0.02275\;.$$

3. What proportion of French men are between 165 and 185 centimeters tall?

$$\begin{aligned}\mathbb{P}[165 < X < 185] &= \mathbb{P}\left[\frac{165-172}{\sqrt{196}} < \frac{X-172}{\sqrt{196}} < \frac{185-172}{\sqrt{196}}\right] \\ &= F(0.928) - F(-0.5) = 0.8234 - 0.3085 = 0.5149\;.\end{aligned}$$

4. If ten thousand French men chosen at random were ranked by increasing size, how tall would be the 9000-th?

   The question amounts to finding the size such that 90% of the French are smaller, i.e. the 90-th quantile of the ninth decile. Let x be that size.

$$\mathbb{P}[X < x] = \mathbb{P}\left[\frac{X-172}{\sqrt{196}} < \frac{x-172}{\sqrt{196}}\right] = 0.9$$

   Thus $\frac{x-172}{\sqrt{196}}$ is the value of the quantile function of the $\mathcal{N}(0,1)$ distribution for $p = 0.9$, that is 1.2816. Therefore:

$$x = 172 + 1.2816 \times \sqrt{196} \simeq 190 \;\textit{cm}.$$

5. The height of French women is modeled by a normal distribution $\mathcal{N}(162, 144)$ (in centimeters). What is the probability for a French man chosen at random to be taller than a French woman chosen at random?

   Let X denote the size of the man and Y that of the woman, and suppose they are independent. Then $X - Y$ follows the normal distribution $\mathcal{N}(10, 340)$. The probability for X to be larger than Y is the probability for $X - Y$ to be positive:

$$\mathbb{P}\left[X - Y > 0\right] = \mathbb{P}\left[\frac{(X-Y)-10}{\sqrt{340}} > \frac{0-10}{\sqrt{340}}\right] = 1 - F(-0.5423) = 0.7062\ .$$

10. Let X be a random variable with $\mathcal{N}(0, 1)$ distribution.

1. Express with the distribution function of X, then compute using the table the following probabilities.

   (a) $\mathbb{P}[X > 1.45]$

   (b) $\mathbb{P}[-1.65 \leqslant X \leqslant 1.34]$

   (c) $\mathbb{P}[|X| < 2.05]$

2. Find the value of u in the following cases.

   (a) $\mathbb{P}[X < u] = 0.63$

   (b) $\mathbb{P}[X \geqslant u] = 0.63$

   (c) $\mathbb{P}[|X| < u] = 0.63$

11. Let X be a random variable with $\mathcal{N}(0, 1)$ distribution. Let $Y = 2X - 3$.

1. What is the distribution of Y?

2. Find $\mathbb{P}[Y < -4]$.

3. Find $\mathbb{P}[-2 < Y < 3]$.

12. Let X be a random variable with $\mathcal{N}(3, 25)$ distribution.

1. Express with the distribution function of the $\mathcal{N}(0, 1)$ distribution, then compute using the table the following probabilities.

   (a) $\mathbb{P}[X < 6]$

   (b) $\mathbb{P}[X > -2]$

   (c) $\mathbb{P}[-1 \leqslant X \leqslant 1.5]$

2. Find the value of u in the following cases.

(a) $\mathbb{P}[X < u] = 0.63$

(b) $\mathbb{P}[X > u] = 0.63$

(c) $\mathbb{P}[|X - 3| \leqslant u] = 0.63$

13. In a given country, the cholesterol concentration of a person taken at random is modeled by a normal distribution with mean 200 mg/100 mL and standard deviation 20 mg/100 mL.

1. What is the probability that a person taken at random in that country has a cholesterol rate below 160 mg/100 mL?

2. What proportion of the population has a cholesterol rate between 170 and 230 mg/100 mL?

3. In another country, the mean cholesterol rate is 190 mg/100 mL, for the same standard deviation as before. Answer the previous questions for this other country.

4. A person is taken at random in each country. What is the probability for the person from the first country to have a higher cholesterol rate than the person from the second country?

14. The size of an ear of wheat in a field is modeled by a random variable X with normal distribution $\mathcal{N}(15, 36)$ (unit: cm).

1. What is the probability for an ear to be smaller than 16 cm?

2. There are about 15 million ears in the field. Give an estimate of the number of ears larger than 20 cm.

3. A sample of 10 ears is picked up in the field. What is the probability that all of them have sizes in the interval [16 ; 20]?

4. In another field, the size of an ear of wheat is modeled by a random variable Y with normal distribution $\mathcal{N}(10, 16)$. What is the probability for an ear from the first field to be larger than an ear from the second field?

1.6 Approximation of a binomial to a normal distribution

- If n is large enough, the binomial distribution $\mathcal{B}(n, p)$ can be approximated to the normal distribution $\mathcal{N}(np, np(1-p))$, having the same expectation and variance.

- In that case, if X follows the $\mathcal{B}(n,p)$ distribution, one computes the probability for X to be in the interval $[a,b]$ by:

$$\begin{aligned}\mathbb{P}[a \leqslant X \leqslant b] &= P\left[\frac{a-np}{\sqrt{np(1-p)}} \leqslant \frac{X-np}{\sqrt{np(1-p)}} \leqslant \frac{b-np}{\sqrt{np(1-p)}}\right] \\ &\simeq F\left(\frac{b-np}{\sqrt{np(1-p)}}\right) - F\left(\frac{a-np}{\sqrt{np(1-p)}}\right) ,\end{aligned}$$

where F is the distribution function of the $\mathcal{N}(0,1)$.

Example. From past experience, it is known that a certain surgery has a 90% chance to succeed. This surgery is performed by a certain clinic 400 times each year. Let N be the number of successes next year. The normal approximation will be used for N.

1. Find the expectation and variance of N.

   The expectation is $400 \times 0.9 = 360$, *the variance is* $400 \times 0.9 \times 0.1 = 36$.

2. Find the probability for the clinic to perform successfully the surgery at least 345 times.

$$\begin{aligned}\mathbb{P}[N \geqslant 345] &= \mathbb{P}\left[\frac{N-360}{\sqrt{36}} \geqslant \frac{345-360}{\sqrt{36}}\right] \\ &= 1 - F(-2.5) = F(2.5) = 0.9938 .\end{aligned}$$

3. Find the probability that the surgery fails in that clinic more than 28 times in the year.

$$\begin{aligned}\mathbb{P}[N \leqslant 372] &= \mathbb{P}\left[\frac{N-360}{\sqrt{36}} \leqslant \frac{372-360}{\sqrt{36}}\right] \\ &= F(2) = 0.9772 .\end{aligned}$$

4. The insurance accepts to cover a certain number of failed surgeries: that number has only a 1% chance to be exceeded. What number is it?

   Let n be the number of failed surgeries that must be determined. The corresponding number of successes is $400-n$. *Therefore* $\mathbb{P}[N \leqslant 400-n] = 0.01$. *Now:*

$$\begin{aligned}\mathbb{P}[N \leqslant 400-n] &= \mathbb{P}\left[\frac{N-360}{\sqrt{36}} \leqslant \frac{400-n-360}{\sqrt{36}}\right] \\ &= F\left(\frac{40-n}{\sqrt{36}}\right) = 0.01 .\end{aligned}$$

The number $\frac{40-n}{\sqrt{36}}$ *is the quantile at* 0.01 *of the* $\mathcal{N}(0,1)$*, that is* -2.3236*. Thus:*

$$\frac{40-n}{\sqrt{36}} = -2.3263 \implies n = 40 + 2.3263\sqrt{36} \simeq 54 \; .$$

The reasoning could also be applied to the number of failed surgeries $R = 400 - N$*. It follows the binomial distribution* $\mathcal{B}(400, 0.1)$*, that can be approximated by the normal* $\mathcal{N}(40, 36)$*. The desired number is such that* $\mathbb{P}[R > n] = 0.01$*.*

$$\begin{aligned}
\mathbb{P}[R > n] &= \mathbb{P}\left[\frac{R-40}{\sqrt{36}} > \frac{n-40}{\sqrt{36}}\right] \\
&= 1 - F\left(\frac{n-40}{\sqrt{36}}\right) \\
&= F\left(\frac{40-n}{\sqrt{36}}\right) = 0.01 \; .
\end{aligned}$$

Of course the result is the same.

15. Among people old enough to receive an injection against the flu, 40% of them ask for it. In a population of 150000 persons old enough to receive the injection, let N be the number of those that will ask for it.

1. What model would you propose for N?

2. If 60500 syringes are prepared, what is the probability that these will not suffice?

3. Find the number of syringes that should be prepared to ensure that there will be enough with 90% probability at least.

16. A restaurant, serving only upon reservation, has 50 seats. The proba-bility that a someone with a reservation does not show up is 1/5. Let N be the number of meals served on a given day. The normal approximation will be used for N.

1. If the chef accepts 50 reservations, what is the probability he will serve more than 45 meals?

2. If he accepts 55 reservations, what is the probability he will find himself in an embarassing situation?

17. Suppose there is probability 0.1 of being controlled in the tramway. Mr A. makes 700 trips per year. The normal approximation will be used for the number of fraud checks.

1. Find the probability that Mr. A will be controled between 60 and 80 times in the year.

2. Mr A. always travels without paying. Knowing that the price of a ticket is 1 euro, what minimal fine should the transportation company charge if they wanted Mr. A to have, over a 1 year period, a probability 0.75 of spending more than he would were he honest.

18. Between Grenoble and Valence TGV, two buses of 50 seats depart on fridays at 4:10 pm. The number of travellers showing up for the trip has a mean of 80 and a standard deviation of 10. The normal approximation will be used for that number.

1. Find the probability for the two buses to be full.
2. One of the buses departs from the station, the other from Victor Hugo square. The passengers choose one or the other at random, but they cannot change if the bus is full. Assume 90 passengers want to go from Grenoble to Valence. What is the probability that at least one of them cannot ?
3. With the same hypotheses as in the previous questions, what should the size of the buses be in order to ensure that the probability of turning down a passenger is lower than 0.05?

19. On average, a passenger that has bought a plane ticket shows up at registration with probability 0.9. A given plane has two hundred seats.

1. If the airline company accepts 220 reservations, what is the probability it will have to turn passengers away?
2. How many reservations should it accept at most to make sure that the probability of turning down at least one passenger is no larger than 0.01?

Parametric estimation

2.1 Estimating a parameter

- For an unknown parameter, an estimator is a function of the data, taking values close to that parameter. It is *unbiased* if its expectation is equal to the parameter. It is *convergent* if the probability for it to take a value at distance up to ε from the parameter tends to 1 as the size of the sample tends to infinity.

- The *empirical frequency* of an event is an unbiased convergent estimator of the probability of that event.

- The *empirical mean* of a sample is an unbiased convergent estimator of the theoretical expectation of the variables.

- The *empirical variance* of a sample is a convergent estimator of the theoretical variance of the variables. Un unbiased estimator is obtained by multiplying the empirical variance by $n/(n-1)$, where n is the size of the sample.

Example. Consider the statistical sample $(1, 0, 2, 1, 1, 0, 1, 0, 0)$.

1. Find its empirical mean and variance.

$$\overline{x} = \frac{6}{9} = \frac{2}{3} \quad \textit{and} \quad s_x^2 = \frac{4}{9} \;.$$

2. Supposing that the data are realizations of a variable with an unknown distribution, give unbiased estimates for the expectation and the variance of that distribution.

   The empirical mean $(2/3)$ *is an unbiased estimate of the expectation. An unbiased estimate of the variance is obtained, multiplying* s_x^2 *by* $9/8$*: this gives* $1/2$.

3. The data of the sample are modeled by a binomial distribution $\mathcal{B}(2, p)$. Use the empirical mean to propose an estimate for p.

   The expectation of the $\mathcal{B}(2, p)$ *distribution is* $2p$. *It is estimated by the empirical mean (here* $2/3$*). Thus* p *can be estimated by:*

$$\frac{2/3}{2} = \frac{1}{3} \;.$$

4. With the same model, use the empirical variance to propose another estimate for p.

   The variance of the $\mathcal{B}(2, p)$ *distribution is* $2p(1-p)$. *It is estimated by* $1/2$. *The value of* p *can be estimated by solving the equation* $2p(1-p) = 1/2$*, giving* $p = 1/2$.

5. The data of the sample are now modeled by a Poisson distribution $\mathcal{P}(\lambda)$, the expectation of which is λ. What estimate would you propose for λ?

   The parameter λ can be estimated by the empirical mean, $2/3$.

20. Consider the statistical sample $(1, 3, 2, 3, 2, 2, 0, 2, 3, 1)$.

1. Supposing that the variables are realizations of a variable with unknown distribution, give unbiased estimates for the expectation and the variance of that distribution.

2. The data of that sample are modeled by a $\mathcal{B}(3, p)$ distribution. Use the empirical mean to propose an estimate for p.

21. Consider the statistical sample $(1.2, 0.2, 1.6, 1.1, 0.9, 0.3, 0.7, 0.1, 0.4)$.

1. The data of that sample are modeled by a uniform distribution on the interval $[0, \theta]$. What estimate would you propose for θ?

2. The data of the sample are now modeled by a normal distribution $\mathcal{N}(\mu, \sigma^2)$. What estimates would you propose for μ and σ^2?

2.2 Confidence intervals for a Gaussian sample

A Gaussian sample is a n-tuple $(X_1, \ldots, X_n)$ of independent random variables with normal distribution $\mathcal{N}(\mu, \sigma^2)$. The empirical mean and variance of the sample are given by:

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

- If the theoretical variance σ^2 is *known*, a confidence interval at level $1-\alpha$ for μ is obtained by:

$$\left[\overline{X} - u_\alpha \frac{\sqrt{\sigma^2}}{\sqrt{n}} \ ; \ \overline{X} + u_\alpha \frac{\sqrt{\sigma^2}}{\sqrt{n}}\right] \ ,$$

  where u_α is the quantile of order $1-\alpha/2$ for the normal distribution $\mathcal{N}(0, 1)$.

- If the theoretical variance σ^2 is *unknown*, a confidence interval at level $1-\alpha$ for μ is obtained by:

$$\left[\overline{X} - t_\alpha \frac{\sqrt{S^2}}{\sqrt{n-1}} \ ; \ \overline{X} + t_\alpha \frac{\sqrt{S^2}}{\sqrt{n-1}}\right] \ ,$$

  where t_α is the quantile of order $1-\alpha/2$ for the Student distribution with parameter $n-1$.

- If the theoretical variance σ^2 is *unknown*, a confidence interval at level $1-\alpha$ for σ^2 is obtained by:

  $$\left[\frac{nS^2}{v_\alpha}\ ;\ \frac{nS^2}{u_\alpha}\right]\ ,$$

  where u_α is the quantile of order $\alpha/2$ for the chi-squared distribution with parameter $n-1$, and v_α is its quantile of order $1-\alpha/2$.

Example. The compression force of a certain type of concrete is modeled by a Gaussian random variable with expectation μ and variance σ^2. The measurement unit is the *psi* (pound per square inch). In questions 1. to 4., it will be supposed that the variance σ^2 is known and equal to 1000. An empirical mean of 3250 psi has been observed from a sample of 12 measurements.

1. Give a 95% confidence interval for μ.

   Here, $\alpha = 0.05$ and $1-\alpha/2 = 0.975$. The quantile of order 0.975 for the $\mathcal{N}(0,1)$ distribution is 1.96. The confidence interval is:

   $$\left[3250 - 1.96\frac{\sqrt{1000}}{\sqrt{12}}\ ;\ 3250 + 1.96\frac{\sqrt{1000}}{\sqrt{12}}\right] \;=\; [3232\,;\,3268]\ .$$

   There is no point in giving more digits than in the empirical mean. The lower bound is rounded to the left, and the upper bound to the right; thus the rounding can only enlarge the interval, ensuring that the confidence level remains higher than 0.95.

2. Give a 99% confidence interval for μ. Compare its width with that of the interval in the previous question.

   Here, $\alpha = 0.01$ and $1-\alpha/2 = 0.995$. The quantile of order 0.995 for the $\mathcal{N}(0,1)$ distribution is 2.5758. The confidence interval is:

   $$\left[3250 - 2.5758\frac{\sqrt{1000}}{\sqrt{12}}\ ;\ 3250 + 2.5758\frac{\sqrt{1000}}{\sqrt{12}}\right] \;=\; [3226\,;\,3274]\ .$$

   This interval is wider than the previous one. The higher the probability that the mean belong to the interval (0.99 instead of 0.95), the wider the interval must be. To get greater confidence, less precision must be accepted.

3. If using the same sample, a confidence interval of width 30 psi were given, what would its confidence level be?

   The width of a confidence interval at level $1-\alpha$ is:

   $$2u_\alpha\frac{\sqrt{1000}}{\sqrt{12}}\ .$$

If that width is 30*, then:*

$$u_\alpha = \frac{30\sqrt{12}}{2\sqrt{1000}} = 1.6432\ .$$

This value is the quantile of order 0.9498 = $1-\alpha/2$ *for the* $\mathcal{N}(0,1)$*. Thus* $\alpha =$ 0.1003 *and* $1-\alpha = 0.8997$*.*

4. What minimal number of trials would be necessary to estimate μ with a precision of ±15 psi, at confidence level 0.95?

   For n *trials, the precision of the confidence interval at level* 0.95 *is:*

   $$\pm 1.96\frac{\sqrt{1000}}{\sqrt{n}}\ .$$

   If it is ±15*, then:*

   $$n = \left(\frac{1.96\sqrt{1000}}{15}\right)^2 = 17.07\ .$$

   The sample size must be at least 18*.*

5. From now on the theoretical variance is supposed to be unknown. For the 12 trials mentioned above:

   $$\sum_{i=1}^{12} x_i^2 = 126761700\ .$$

   Give a 95% confidence interval for μ and compare it with that of question 1. Repeat the calculation for a 99% confidence interval and compare it with that of question 2.

   The estimated variance is:

   $$s^2 = \frac{1}{12} \times 126761700 - (3250)^2 = 975\ .$$

   The quantile of order 0.975 *for the Student* $\mathcal{T}(n-1)$ *distribution is* 2.201*, that of order* 0.995 *is* 3.106*. The* 95% *confidence interval is:*

   $$\left[3250 - 2.201\frac{\sqrt{975}}{\sqrt{11}}\ ;\ 3250 + 2.201\frac{\sqrt{975}}{\sqrt{11}}\right] = [3229\,;\,3271]\ .$$

   The 99% *confidence interval is:*

   $$\left[3250 - 3.106\frac{\sqrt{975}}{\sqrt{11}}\ ;\ 3250 + 3.106\frac{\sqrt{975}}{\sqrt{11}}\right] = [3220\,;\,3280]\ .$$

Even though the estimated variance is lower than the theoretical one in this case, the confidence intervals calculated with the Student distribution (unknown variance) are wider thus less precise than those computed with the normal distribution (known variance). This is due to the fact that Student distributions are more scattered than the $\mathcal{N}(0,1)$ *one: the interval containing* 95% *of values for the* $\mathcal{T}(11)$ *is* $[-2.201\ ;\ +2.201]$, *instead of* $[-1.96\ ;\ +1.96]$ *for the* $\mathcal{N}(0,1)$. *It is reasonable to expect less precision when less information is available on the model.*

6. Give a 95% confidence interval for the variance, and for the standard deviation.

   The quantile of order 0.025 *for the khi-squared distribution* $\mathcal{X}^2(11)$ *is* $u_\alpha = 3.816$. *The quantile of order* 0.975 *is* $v_\alpha = 21.92$. *The* 95% *confidence interval for the variance is:*

   $$\left[\frac{12\times 975}{21.92}\ ;\ \frac{12\times 975}{3.816}\right] = [533\,;\,3067]\ .$$

   By taking the square root of both bounds, a confidence interval for the standard deviation is obtained:

   $$\left[\sqrt{\frac{12\times 975}{21.92}}\ ;\ \sqrt{\frac{12\times 975}{3.816}}\right] = [23.1\,;\,55.4]\ .$$

   The confidence intervals for the variance or the standard deviations on small samples are usually very wide.

22. The weight of grapes produced per vine has been measured on 10 vines selected at random in a vineyard. The results in kilograms are the following:

$$2.4 \quad 3.4 \quad 3.6 \quad 4.1 \quad 4.3 \quad 4.7 \quad 5.4 \quad 5.9 \quad 6.5 \quad 6.9\ .$$

The weight of grapes produced by each vine is modeled by a $\mathcal{N}(\mu,\sigma^2)$ distribution.

1. Find the empirical mean and variance of the sample.

2. Give a 95% confidence interval for μ.

3. Give a 95% confidence interval for σ^2.

4. From now on, the standard deviation of productions per plant is supposed to be known and equal to 1.4. Give a 95% confidence interval for μ.

5. Find the minimal number of plants that should be taken to estimate μ with 99% confidence with a precision of ±500 grams.

23. A study on coronary blood flow velocity has lead to the following results in 18 people:

75, 77, 78, 77, 77, 72, 72, 72, 70, 71, 69, 69, 68, 66, 64, 66, 62, 61.

The values of that sample are modeled by a random variable with normal distribution $\mathcal{N}(\mu, \sigma^2)$, where μ and σ^2 are unknown.

1. Find the mean and variance of the sample.
2. Find the intervals of confidence for μ at levels 0.95, 0.98, and 0.99.
3. Find the intervals of confidence for σ^2 at levels 0.95, 0.98, and 0.99.
4. What would the confidence intervals for μ be, if the variance σ^2 was known and equal to 26?

24. A laboratory uses an optical device to measure the fluorescein concen-tration in solutions. The measurements are modeled by a Gaussian random variable, the expectation of which is equal to the true concentration of the solution, and the standard deviation, guaranteed by the company selling the device is known: $\sigma = 0.05$.

1. Nine measurements are made for the same solution. The empirical mean of the 9 results is 4.38 mg/l. Give a 99% confidence interval for the true concentration of the solution.
2. For that same sample, what is the confidence level of the interval [4.36 ; 4.40]?
3. What should the sample size be if the concentration of the solution had to be known at confidence level 0.99, with a precision of ±0.01 mg/l?
4. On the same sample of 9 measurements, a standard deviation of 0.08 mg/l has been observed. Give a 99% confidence interval for the theoretical standard deviation. What do you think of the company's guarantee?
5. Answer the first question again, this time supposing that the theoretical standard deviation is unknown, and estimated by the empirical one.

25. To study the rotting of potatoes, a researcher injects bacteria that induce decay, into 13 potatoes. He then measures the rotten area (in mm^2) in these 13 potatoes. He gets a mean of 7.84 mm^2 and an empirical variance of 14.13. The rotten area of a potato is modeled by a random variable with $\mathcal{N}(\mu, \sigma^2)$ distribution.

1. Find a confidence interval for μ at level 0.95, then 0.99.
2. Find a confidence interval for σ^2 at level 0.95, then 0.99.

26. The production of a new kind of apple tree has to be estimated. The production of one tree is modeled by a Gaussian random variable with expectation μ and standard deviation σ, both unknown.

1. A sample of 15 apple trees, has produced a mean crop of 52 kg with standard deviation 5 kg. Find a confidence interval for the expected production of the apple trees of the new species, at level 0.95, then 0.99.
2. Find a confidence interval for the standard deviation σ, at level 0.95.

2.3 Confidence interval for the expectation on a large sample

For a large sample, a confidence interval at approximate level $1-\alpha$ for the expectation is obtained by:

$$\left[\overline{X} - u_\alpha \frac{\sqrt{S^2}}{\sqrt{n}}\ ;\ \overline{X} + u_\alpha \frac{\sqrt{S^2}}{\sqrt{n}}\right] ,$$

where u_α is the quantile of order $1-\alpha/2$ for the normal distribution $\mathcal{N}(0,1)$.

Example. The fluorescein concentration of a given solution has been measured 90 times. An empirical mean of 4.38 mg/l and a standard deviation of 0.08 mg/l have been observed. Give a confidence interval for the true concentration of the solution, at confidence levels 0.95 and 0.99.

The quantile of order 0.975 *for the* $\mathcal{N}(0,1)$ *distribution is* 1.96*. The* 95% *confidence interval is:*

$$\left[4.38 - 1.96\frac{0.08}{\sqrt{90}}\ ;\ 4.38 + 1.96\frac{0.08}{\sqrt{90}}\right] \ =\ [4.363\,;\ 4.397]\ .$$

The quantile of order 0.995 *for the* $\mathcal{N}(0,1)$ *distribution is* 2.5758*. The* 99% *confidence interval is:*

$$\left[4.38 - 2.5758\frac{0.08}{\sqrt{90}}\ ;\ 4.38 + 2.5758\frac{0.08}{\sqrt{90}}\right] \ =\ [4.358\,;\ 4.402]\ .$$

27. The production of a new kind of apple trees has to be estimated. On a sample of 80 trees, a mean crop of 51.5 kg, with standard deviation 4.5 kg has been observed. Find a confidence interval for the expected production of apple trees of that species, at level 0.95, then 0.99.

28. The lengths in millimeters of 152 cuckoo eggs have been measured. The empirical mean was found to be 40.8 mm, with an empirical variance of 14.7 mm^2. Find a confidence interval for the expected length of a cuckoo egg, at levels 0.95, 0.98, and 0.99.

29. The lengths in millimeters of 150 walnut shells have been measured. The empirical mean is 27.6 mm, and the empirical standard deviation is 3.7 mm. Give a confidence interval for the expected length of a wallnut shell, at level 0.99, then 0.998.

30. Sleeping pills have been given to two groups of patients A and B. The 100 patient of group A have received a new sleeping pill whereas the 50 patient of group B have received an old one. The patients of group A have slept 7.82 hours on average, with a standard deviation of 0.24 h; the patients of group B have slept 6.75 hours on average, with a standard deviation of 0.30 h.

1. Find a confidence interval for the average sleeping time of patients receiving the new pill, at levels 0.90, 0.95, 0.99.

2. Same question for patients taking the old pill.

3. Do you think that the new sleeping pill is more efficient than the old one?

2.4 Confidence interval of a probability for a large sample

For a large binary sample, a confidence interval at level $1-\alpha$ for the probability of the event is obtained by:

$$\left[\overline{X} - u_\alpha \frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{n}}\;;\;\overline{X} + u_\alpha \frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{n}}\right]\;,$$

where n is the sample size, $\overline{X}$ is the empirical frequency of the event and u_α is the quantile of order $1-\alpha/2$ of the normal distribution $\mathcal{N}(0,1)$.

Example. In order to study the influence of X-rays on the spermatogenesis of Bombyx mori, males have been exposed to radiation on the second day and on the fourth day of the larval stage. These males have been mated with non exposed females, and the number of fertile eggs laid by the females have been counted: out of a total of 5646 eggs laid, 4998 were fertile. In a control group of non exposed males and females, 5834 fertile eggs out of 6221 were obtained.

1. Find a 95% confidence interval for the proportion of fertile eggs after radiation exposure of males.

   The empirical frequency of fertile eggs after exposure of males is:

   $$F = \frac{4998}{5646} = 0.885\;.$$

   The 95% *confidence interval is:*

   $$\begin{aligned} & \left[0.885 - 1.96\frac{\sqrt{0.885(1-0.885)}}{\sqrt{5646}}\;;\;0.885 + 1.96\frac{\sqrt{0.885(1-0.885)}}{\sqrt{5646}}\right] \\ = \quad & [0.876\,;\,0.894]\;. \end{aligned}$$

2. Find a 95% confidence interval for the proportion of fertile eggs of non exposed couples.

   The empirical frequency of fertile eggs for non exposed couples is:

   $$F = \frac{5834}{6221} = 0.938\;.$$

The 95% *confidence interval is:*

$$\left[0.938 - 1.96\frac{\sqrt{0.938(1-0.938)}}{\sqrt{6221}}\ ;\ 0.938 + 1.96\frac{\sqrt{0.938(1-0.938)}}{\sqrt{6221}}\right]$$

$$= \ [0.931\ ;\ 0.944]\ .$$

3. What do you think of the influence of radiation exposure on fertility?

   The two confidence intervals do not intersect; thus the proportion of fertile eggs is significantly lower for exposed males.

31. On a sample of $n = 500$ teenagers, 210 were found to be overweight. Let p be the proportion of overweight teenagers. Find a confidence interval for p, at levels 0.95 and 0.99.

32. A clinic has proposed a new surgery, and has had 40 failures out of 200 attempts. Let p be the probability of success of that new surgery.

1. What estimate would you propose for p?
2. Using the normal approximation, give a 95% confidence interval for p.
3. How many surgeries should the clinic perform to know the success probability with a precision of ±1% with 95% confidence level?

Statistical testing

3.1 Decision rule, threshold and p-value

- In a test, the *null hypothesis* $\mathcal{H}_0$ is the one for which the probability of rejecting it wrongly is controled. It is the most valuable one, the one it would be costly to reject wrongly.

- The *threshold* of the test, also called the *first kind risk* is the probability of rejecting $\mathcal{H}_0$ wrongly:
$$\mathbb{P}_{\mathcal{H}_0}[\text{ Reject } \mathcal{H}_0\,] = \alpha \;.$$

- The *test statistic* is a function of the data, for which we know the probability distribution under the null hypothesis $\mathcal{H}_0$.

- The *decision rule* specifies, as a function of the values taken by the test statistic, in which cases the hypothesis $\mathcal{H}_0$ should be rejected.

- A test may be:

  ⋆ *two-tailed* if the decision rule is:
  $$\text{Reject } \mathcal{H}_0 \iff T \notin [l, l']$$
  (reject too small or too large values). Usually, l and l' are chosen such that $\mathbb{P}_{\mathcal{H}_0}[T < l] = \mathbb{P}_{\mathcal{H}_0}[T > l'] = \alpha/2$.

  ⋆ *one-tailed* if the decision rule is:
  $$\text{Reject } \mathcal{H}_0 \iff T < l$$
  (reject too small values),
  or else:
  $$\text{Reject } \mathcal{H}_0 \iff T > l$$
  (reject too large values).

- The *p-value* is the threshold for which the observed value of the test statistic would be on the boundary of the rejection region. It is the probability under $\mathcal{H}_0$ that the test statistic be beyond the value that has been observed.

- The *second kind risk* is the probability of accepting $\mathcal{H}_0$ wrongly, *i.e.* the probability of accepting $\mathcal{H}_0$ when the *alternative hypothesis* $\mathcal{H}_1$ is true:
$$\mathbb{P}_{\mathcal{H}_1}[\text{ accept } \mathcal{H}_0\,] = \beta \;.$$
The *power* of the test is $1-\beta$. It is the probability of being right rejecting $\mathcal{H}_0$.

Example. For an adult, the logarithm of the D-dimer concentration, denoted by X, is modeled by a normal random variable with expectation μ and variance σ^2. The variable X is an indicator for the risk of thrombosis: it is considered that for healthy individuals, μ is -1, whereas for individuals at risk μ is 0. In both cases, the value of σ^2 is the same: 0.09.

1. Dr. House does not want to worry his patients if there is no need to. What hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ will he choose to test? Give the decision rule for his test, at threshold 1%, and at threshold 5%.

If Dr. House does not want to worry a patient, the hypothesis he considers as dangerous to reject wrongly is that the patient is not at risk, thus that his value of X (the test statistic) has expectation -1. His hypothesis $\mathcal{H}_0$ is $\mu = -1$ (the patient is not at risk), that he will test against $\mathcal{H}_1$: $\mu = 0$ (the patient is at risk). He will choose to reject too high values for X. The decision rule will be:

$$\textit{Reject } \mathcal{H}_0 \iff X > l\ ,$$

where:

$$\mathbb{P}_{\mathcal{H}_0}[X > l] = \alpha\ .$$

According to the null hypothesis $\mathcal{H}_0$, the test statistic X follows the $\mathcal{N}(-1, 0.09)$ distribution, hence $\frac{X-(-1)}{\sqrt{0.09}}$ follows the $\mathcal{N}(0,1)$ distribution. An equivalent decision rule is:

$$\textit{Reject } \mathcal{H}_0 \iff \frac{X-(-1)}{\sqrt{0.09}} > \frac{l-(-1)}{\sqrt{0.09}}\ .$$

Therefore $\frac{l-(-1)}{\sqrt{0.09}}$ is the value which has probability α to be overpassed for a $\mathcal{N}(0,1)$ distribution: 1.6449 *for $\alpha = 0.05$,* 2.3263 *for $\alpha = 0.01$. At threshold* 0.05 *the decision rule of the test is:*

$$\begin{aligned}\textit{Reject } \mathcal{H}_0 &\iff \frac{X-(-1)}{\sqrt{0.09}} > 1.6449\\ &\iff X > 1.6449\sqrt{0.09} + (-1) = -0.5065\ .\end{aligned}$$

Dr. House declares the patient is at risk if the logarithm of his D-dimer concentration is higher than -0.5065.

At threshold 0.01 *the decision rule of the test is:*

$$\begin{aligned}\textit{Reject } \mathcal{H}_0 &\iff \frac{X-(-1)}{\sqrt{0.09}} > 2.3263\\ &\iff X > 2.3263\sqrt{0.09} + (-1) = -0.3021\ .\end{aligned}$$

The lower the threshold, the less the decision rule rejects risky patients: what should happen to reject $\mu = -1$ at threshold 0.01 *is more unusual than at threshold* 0.05*.*

2. Find the second kind risk and the power of the tests of the previous question.

   The second kind risk is the probability of rejecting $\mathcal{H}_1$ wrongly. Under hypothesis $\mathcal{H}_1$, $\mu = 0$, and the variable X follows the $\mathcal{N}(0, 0.09)$ distribution.
   For the test at threshold 0.05, *the probability of accepting $\mathcal{H}_0$ wrongly (i.e. of declaring wrongly that a patient is not at risk) is:*

   $$\beta = \mathbb{P}_{\mathcal{H}_1}[X \leqslant -0.5065] = \mathbb{P}_{\mathcal{H}_1}\left[\frac{X-0}{\sqrt{0.09}} \leqslant \frac{-0.5065-0}{\sqrt{0.09}}\right]$$

   Under hypothesis $\mathcal{H}_1$, $\frac{X-0}{\sqrt{0.09}}$ follows the $\mathcal{N}(0,1)$ distribtuion. Therefore we must calculate, for a variable with $\mathcal{N}(0,1)$ distribution, the probability to fall below $\frac{-0.5065-0}{\sqrt{0.09}} = -1.6885$: this is the value of the distribution function of the $\mathcal{N}(0,1)$ at -1.6885, that is 0.0457. *The power is:*

   $$1 - \beta = 1 - 0.0457 = 0.9543\ .$$

   For the test at threshold 0.01, *the reasoning is the same, replacing the bound* -0.5065 *by* -0.3021. *A second kind risk equal to* 0.1570 *is found, with a power equal to* 0.8430.
   When the threshold is lower, the risk of wrongly rejecting $\mathcal{H}_0$ is lowered, but the risk of accepting it wrongly is higher and the power is lower. For the test at threshold 0.01, *the probability that Dr. House wrongly declaring that a patient is not at risk is about* 16%.

3. A patient has a value of X equal to -0.46. Find the p-value of Dr. House's test.

   The p-value is the threshold at which -0.46 would be the bound. Knowing the results of the first question, since -0.46 lies between -0.5065 and -0.3021, the p-value must be between 0.05 *and* 0.01. *It is the probability under $\mathcal{H}_0$, that the variable X is higher than -0.46.*

   $$\mathbb{P}_{\mathcal{H}_0}[X > -0.46] = \mathbb{P}_{\mathcal{H}_0}\left[\frac{X-(-1)}{\sqrt{0.09}} > \frac{-0.46-(-1)}{\sqrt{0.09}}\right] = \mathbb{P}_{\mathcal{H}_0}\left[\frac{X-(-1)}{\sqrt{0.09}} > 1.8\right]\ .$$

   Under $\mathcal{H}_0$, $\frac{X-(-1)}{\sqrt{0.09}}$ follows the $\mathcal{N}(0,1)$ distribution. The probability we are looking for is $1 - F(1.8)$, where F is the distribution function of the $\mathcal{N}(0,1)$, i.e. 0.0359.

4. Dr. Cuddy's point of view is that she'd rather worry a patient wrongly than not warn him of an actual risk. What hypotheses $\mathcal{H}'_0$ and $\mathcal{H}'_1$ will she choose to test? Give the decision rule for her test, at threshold 1%, and at threshold 5%.

   If Dr. Cuddy does not want to miss a patient at risk, the hypothesis she considers as dangerous to reject wrongly is that he is at risk, and that his variable X has expectation 0. *Her hypothesis $\mathcal{H}'_0$ is $\mu = 0$ (the patient is at risk), that she will* *test against $\mathcal{H}'_1$: $\mu = -1$ (the patient is not at risk). She will choose to reject lower values of X. The decision rule will be:*

$$\textit{Reject } \mathcal{H}'_0 \iff X < l' \ ,$$

where:

$$\mathbb{P}_{\mathcal{H}'_0}[X < l'] = \alpha \ .$$

Under hypothesis $\mathcal{H}'_0$, the test statistic X follows the $\mathcal{N}(0, 0.09)$ distribution, therefore $\frac{X-0}{\sqrt{0.09}}$ follows the $\mathcal{N}(0,1)$. An equivalent decision rule is:

$$\textit{Reject } \mathcal{H}'_0 \iff \frac{X-0}{\sqrt{0.09}} < \frac{l'-0}{\sqrt{0.09}} \ .$$

Thus $\frac{l'-0}{\sqrt{0.09}}$ is the value such that a variable with $\mathcal{N}(0,1)$ distribution falls below with probability α: -1.6449 for $\alpha = 0.05$, -2.3263 for $\alpha = 0.01$. At threshold 0.05 the decision rule is:

$$\begin{aligned} \textit{Reject } \mathcal{H}_0 &\iff \frac{X-0}{\sqrt{0.09}} < -1.6449 \\ &\iff X < -1.6449 \times \sqrt{0.09} + 0 = -0.4935 \ . \end{aligned}$$

Dr Cuddy declares that the patient is not at risk when his D-dimer variable is below -0.4935.
At threshold 0.01 the decision rule is:

$$\begin{aligned} \textit{Reject } \mathcal{H}'{}_0 &\iff \frac{X-0}{\sqrt{0.09}} < -2.3263 \\ &\iff X < -2.3263 \times \sqrt{0.09} + 0 = -0.6980 \ . \end{aligned}$$

5. Depending on the threshold, for what values of X will Drs. House and Cuddy's diagnoses agree?

   If $X < \min\{l, l'\}$, Dr. House accepts $\mathcal{H}_0$, Dr. Cuddy rejects $\mathcal{H}'_0$. In both cases, the conclusion for the patient is the same: he is not at risk. Conversely, if $X > \max\{l, l'\}$ Dr. House rejects $\mathcal{H}_0$, Dr. Cuddy accepts $\mathcal{H}'_0$ and the conclusion is the same: the patient is at risk.

   The conclusions differ for a patient whose value of X lies between l and l'. At threshold 0.05 the bounds are $l = -0.5065$ and $l' - 0.4935$. For a patient whose variable X is between -0.5065 and -0.4935, Dr. House declares him at risk (he rejects $\mathcal{H}_0$), Dr. Cuddy declares he is not at risk (she rejects $\mathcal{H}'_0$).

   At threshold 0.01, the bounds are $l = -0.3021$ and $l' = -0.6980$. For a patient whose variable X is between -0.6980 and -0.3021, Dr. House declares he is not at risk (he accepts $\mathcal{H}_0$), Dr. Cuddy declares he is at risk (she accepts $\mathcal{H}'_0$).

6. Give the decision rule of the test for the null hypothesis $\mathcal{H}_0'' : \mu = -1$ against the alternative one $\mathcal{H}_1'' : \mu \neq -1$.

   This is a two-tailed test. The decision rule will be:

   $$\textit{Reject } \mathcal{H}_0'' \iff X \notin [l_1, l_2] \ ,$$

   where:

   $$\mathbb{P}_{\mathcal{H}_0''}[\, X \notin [l_1, l_2]\,] = 0.05 \ .$$

   Under hypothesis $\mathcal{H}_0''$, the test statistic X follows the $\mathcal{N}(-1, 0.09)$ distribution, hence $\frac{X-(-1)}{\sqrt{0.09}}$ follows the $\mathcal{N}(0,1)$. An equivalent decision rule is:

   $$\textit{Reject } \mathcal{H}_0'' \iff \frac{X-(-1)}{\sqrt{0.09}} \notin \left[\frac{l_1-(-1)}{\sqrt{0.09}}\, ; \, \frac{l_2-(-1)}{\sqrt{0.09}}\right] \ .$$

   The interval $\left[\frac{l_1-(-1)}{\sqrt{0.09}}\, ; \, \frac{l_2-(-1)}{\sqrt{0.09}}\right]$ must contain 95% of the values taken by a variable with $\mathcal{N}(0,1)$ distribution. The interval centered at 0 is chosen: $[-1.96\, ; +1.96]$. Hence:

   $$\frac{l_1-(-1)}{\sqrt{0.09}} = -1.96 \implies l_1 = (-1) - 1.96\sqrt{0.09} = -1.588 \ ,$$

   and:

   $$\frac{l_2-(-1)}{\sqrt{0.09}} = +1.96 \implies l_2 = (-1) + 1.96\sqrt{0.09} = -0.412 \ ,$$

   At threshold 0.05 the decision rule of the two-tailed test is:

   $$\textit{Reject } \mathcal{H}_0 \iff X \notin [-1.588\, ; \, -0.412] \ .$$

   The patient is said to have logarithm of D-dimer concentration significantly different from -1 when his variable X is either lower than -1.588, or larger than -0.488.

7. A patient has a value of X equal to -0.46. Find the p-value for the test of the previous question.

   The p-value is the threshold for which the observed value would be a bound of the rejection region. That rejection region is centered at -1. The other bound should be $-1-(-0.46-(-1)) = -1.54$.

The p-value is the following probability.

$$\begin{aligned}
& \mathbb{P}_{\mathcal{H}_0''}[X \notin [-1.54\,;\,-0.46]\,] \\
= \;& \mathbb{P}_{\mathcal{H}_0''}\left[\frac{X-(-1)}{\sqrt{0.09}} \notin \left[\frac{-1.54-(-1)}{\sqrt{0.09}}\;;\;\frac{-0.46-(-1)}{\sqrt{0.09}}\right]\right] \\
= \;& \mathbb{P}_{\mathcal{H}_0''}\left[\frac{X-(-1)}{\sqrt{0.09}} \notin [-1.8\,;\,+1.8]\right] .
\end{aligned}$$

Under the hypothesis $\mathcal{H}_0''$, the variable $\frac{X-(-1)}{\sqrt{0.09}}$ follows the $\mathcal{N}(0,1)$ distribution: the desired probability is 0.0719*. The p-value is found to be twice that of the one-tailed test in question 3.*

33. A packaging machine is supposed to produce packs of 1 kg. The actual weight of a pack is modeled by a random variable following a normal distribution, with a standard deviation of 20 g. However, it is possible to tune the mean weight of the packs.

1. The production manager decides not to distribute packs with weight too far away from the prescribed value of 1 kg. What hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ should he test? Establish the decision rule for that test at thresholds 5% and 1%.

2. The company manager thinks that the packs going out on sale are too heavy, causing money loss. What hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ should the production manager use to answer the criticism? Establish the decision rule for that test at thresholds 5% and 1%.

3. A pack has been weighed at 1018 grams. What is the p-value for the test of the previous question? What is the p-value for the test of the first question?

4. A consumers' association sues the company for selling packs that are too light. What hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ should the production manager use to answer? Establish the decision rule for that test at thresholds 5% and 1%.

5. A pack has been weighed at 982 grams. What is the p-value for the test of the previous question? What is the p-value for the test of the first question?

34. A paracetamol concentration of more than 150 mg per kilogram body weight is considered as dangerous. The mesurements of paracetamol in blood tests are modelled by a random variable with normal distribution $\mathcal{N}\,(\mu,\,\sigma^2)$. The standard-deviation, linked to the testing method, is supposed to be known and equal to 5 mg.

1. Give the hypotheses and the decision rule for the test deciding, at threshold 5%, whether a patient is at risk, from the results of one blood test (you are a cautious doctor).
2. A patient arrives at the hospital with signs of paracetamol poisoning. A blood test is made and a concentration of 140 mg is found. Give the p-value for the test of the previous question. Should that patient be considered at risk?

35. Let X be the pollution index measured close to a chemical plant. It is modeled by a $\mathcal{N}(\mu, \sigma^2)$ distribution. The standard deviation is supposed to be known and equal to 4. The state regulations fix the maximal pollution index at 30.

1. The head manager wants to show that his plant complies with the regulations. What hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ should he test? Establish the decision rule for that test at thresholds 5% and 1%.
2. The Green party wants to prove that the pollution is higher than prescribed. What hypotheses $\mathcal{H}'_0$ and $\mathcal{H}'_1$ should they test? Establish the decision rule for that test at thresholds 5% and 1%.

3.2 Tests on a sample

Denote by:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2$$

the mean and standard deviation of the sample. The expectation of the unknown distribution is μ, its variance is σ^2. The test statistics and their probability distribution under the null hypothesis $\mathcal{H}_0$ are the following.

- Testing values of expectation, Gaussian sample, σ^2 known.

$$\mathcal{H}_0 : \mu = \mu_0 \quad ; \quad T = \sqrt{n}\left(\frac{\overline{X} - \mu_0}{\sqrt{\sigma^2}}\right) \text{ follows the } \mathcal{N}(0,1)\ .$$

- Testing values of the expectation, Gaussian sample, σ^2 unknown.

$$\mathcal{H}_0 : \mu = \mu_0 \quad ; \quad T = \sqrt{n-1}\left(\frac{\overline{X} - \mu_0}{\sqrt{S^2}}\right) \text{ follows the } \mathcal{T}(n-1)\ .$$

- Testing values of the variance, Gaussian sample, σ^2 unknown.

$$\mathcal{H}_0 : \sigma^2 = \sigma_0^2 \quad ; \quad T = n\left(\frac{S^2}{\sigma_0^2}\right) \text{ follows the } \mathcal{X}^2(n-1)\ .$$

- Testing values of the expectation, large sample, σ^2 known or not.

$$\mathcal{H}_0 : \mu = \mu_0 \quad ; \quad T = \sqrt{n}\left(\frac{\overline{X} - \mu_0}{\sqrt{S^2}}\right) \text{ follows the } \mathcal{N}(0,1) \; .$$

- Testing values of a probability, large binary sample.

$$\mathcal{H}_0 : p = p_0 \quad ; \quad T = \sqrt{n}\left(\frac{\overline{X} - p_0}{\sqrt{p_0(1-p_0)}}\right) \text{ follows the } \mathcal{N}(0,1) \, .$$

Example. For an adult, the logarithm of the D-dimer concentration, denoted by X, is modeled by a normal random variable with expectation μ and variance σ^2. The variable X is an indicator for the risk of thrombosis: it is considered that for healthy individuals, μ is -1, whereas for individuals at risk μ is 0. The influence of olive oil on thrombosis risk must be evaluated.

1. A group of 13 patients, previously considered as being at risk, had an olive oil enriched diet. After the diet, their value of X was measured, and this gave an empirical mean of -0.15. The variance σ^2 is supposed to be known and equal to 0.09. Give the decision rule for the test of $\mathcal{H}_0 : \mu = 0$ against $\mathcal{H}_1 : \mu = -1$, at threshold 5%. What p-value corresponds to -0.15? What is your conclusion? Find the second kind risk and the power of the test.

   We have a Gaussian sample with known variance, and we build a test on the value of the expectation. The test statistic is:

$$T = \sqrt{13}\frac{\overline{X} - 0}{\sqrt{0.09}} \; .$$

   According to the null hypothesis $\mathcal{H}_0$, T follows the normal distribution $\mathcal{N}(0,1)$. The hypothesis $\mathcal{H}_0$ is rejected when T takes low values. At threshold 5%, the bound is -1.6449. The decision rule is:

$$\textit{Reject } \mathcal{H}_0 \iff T < -1.6449 \iff \overline{X} < -0.1369 \; .$$

   For $\overline{X} = -0.15$, the test statistic takes a value of -1.8028, the corresponding p-value is 0.0357. At threshold 5%, the hypothesis $\mathcal{H}_0$ is rejected, thus the decision is that there has been a significant improvement. But at any threshold smaller than 3.57%, $\mathcal{H}_0$ is not rejected: the olive oil has not made a significant improvement. Under hypothesis $\mathcal{H}_1$, $\sqrt{13}\frac{\overline{X}-(-1)}{\sqrt{0.09}}$ follows the $\mathcal{N}(0,1)$ distribution. The second *kind risk is the probability of accepting $\mathcal{H}_0$ wrongly, i.e.:*

$$\begin{aligned}
\beta &= \mathbb{P}_{\mathcal{H}_1}[\overline{X} > -0.1369] \\
&= \mathbb{P}_{\mathcal{H}_1}\left[\sqrt{13}\frac{\overline{X}-(-1)}{\sqrt{0.09}} > \sqrt{13}\frac{-0.1369-(-1)}{\sqrt{0.09}}\right] \\
&= \mathbb{P}_{\mathcal{H}_1}\left[\sqrt{13}\frac{\overline{X}-(-1)}{\sqrt{0.09}} > 10.3732\right] \\
&\simeq 0\ .
\end{aligned}$$

The second kind risk is very close to 0 *(lower than* 10^{-20}*), and the power is very close to* 1*.*

2. For the same group of 13 patients, an empirical standard deviation of 0.37 has been observed. Give the decision rule for the test of $\mathcal{H}_0 : \sigma^2 = 0.09$, against $\mathcal{H}_1 : \sigma^2 \neq 0.09$, at threshold 5%. What is your conclusion?

   We must test a value of the variance for a Gaussian sample. The test statistic is:

   $$T = 13\frac{S^2}{0.09}\ .$$

   Under hypothesis $\mathcal{H}_0$, it follows the chi-squared distribution with parameter 12*. We want a two-tailed test, hence a decision rule rejecting values euther too low or too high.*

   $$\textit{Reject } \mathcal{H}_0 \iff T \notin [l, l']\ ,$$

   where l and l' are the quantiles of order 0.025 *and* 0.975 *for the chi-squared distribution with parameter* 12*:* $l = 4.4038$ *and* $l' = 23.3367$*. Here the test statistic takes a value of* 19.7744*. It is a high value, but not high enough to reject the hypothesis that the theoretical variance is* 0.09*.*

3. Assuming that the variance is unknown and using the estimate of the previous question, give the decision rule for the test of $\mathcal{H}_0 : \mu = 0$, against $\mathcal{H}_1 : \mu < 0$, at threshold 5%. What is your conclusion?

   We have a Gaussian sample with unknown variance and we must build a test on the value of the expectation. The test statistic is:

   $$T = \sqrt{12}\frac{\overline{X}-0}{\sqrt{S^2}}\ .$$

   Under hypothesis $\mathcal{H}_0$, T follows the Student distribution $\mathcal{T}(12)$. The hypothesis $\mathcal{H}_0$ is rejected when T takes low values. At threshold 5% *the bound is* -1.7823*. The decision rule is:*

   $$\textit{Reject } \mathcal{H}_0 \iff T < -1.7823\ .$$

For $\overline{X} = -0.15$ and $\sqrt{S^2} = 0.37$, the test statistic T takes a value of -1.4044, thus $\mathcal{H}_0$ cannot be rejected (the corresponding p-value is 0.0928). It can be said that there has not been a significant improvement.

4. The same experiment is repeated on a group of 130 patients, for whom an empirical mean of -0.12 and a standard deviation of 0.32 are observed. Give the decision rule for the test of $\mathcal{H}_0 : \mu = 0$ against $\mathcal{H}_1 : \mu < 0$, at threshold 5%. What p-value corresponds to -0.12? What is your conclusion?

   Now we must test a value of the expectation for a large sample. The test statistic is:

   $$T = \sqrt{130}\frac{\overline{X} - 0}{\sqrt{S^2}} \; .$$

   Under hypothesis $\mathcal{H}_0$, T follows the normal distribution $\mathcal{N}(0, 1)$. The hypothesis $\mathcal{H}_0$ is rejected when the values of T are too low values. At threshold 5% the bound is -1.6449. The decision rule is:

   $$\textit{Reject } \mathcal{H}_0 \iff T < -1.6449 \; .$$

   For $\overline{X} = -0.12$ and $\sqrt{S^2} = 0.32$, the test statistic takes a value of -4.2757, the corresponding p-value is close to 10^{-5}. Thus it can be concluded without doubt that the expected value of X is significantly lower than 0.

5. The D-dimer concentration of the 130 patients was measured before the diet. After the diet, the concentration was found to be lower for 78 patients, higher for 52 patients. Build a test to decide whether the olive oil regime has improved the condition for a significant proportion of the patients. With the observations you have, what is the p-value for that test, what is your conclusion?

   Let p be the probability of improvement (lower value of X). If the regime had no effect, fluctuations in measurements would be merely random and there would be as many improvements as worsenings: the proportion of improvements would be $1/2$. We must test, for a large binary sample, the hypothesis $\mathcal{H}_0 : p = 0.5$, against $\mathcal{H}_1 : p > 0.5$. The test statistic is:

   $$T = \sqrt{n}\frac{\overline{X} - 0.5}{\sqrt{0.5(1 - 0.5)}} \; .$$

   Under hypothesis $\mathcal{H}_0$, the test statistic follows the normal distribution $\mathcal{N}(0, 1)$. Here $\overline{X}$ is the observed proportion of improvements, that is $78/130$. The test statistic takes a value of 2.2804, the corresponding p-value (probability that a $\mathcal{N}(0, 1)$ variable be higher than 2.2804) is 0.0113. It can be concluded that the improvement is significant at threshold 5%, but not quite at threshold 1%.

36. A packaging machine is supposed to produce 1 kg packs. The actual weight of a pack is modeled by a random variable following a normal distribution with a standard deviation of 20 g. It is possible to tune the mean weight of the packs. In order to check that the tuning is correct, a sample of 10 packs is weighed.

1. Let $\mathcal{H}_0$ be the hypothesis: “the mean weight is 1 kg”. Build a test at threshold 1%, of $\mathcal{H}_0$ against hypothesis $\mathcal{H}_1$: “the mean weight is different from 1 kg”. Find the p-value of that test, for a sample of average weight 1011 grams.
2. Same question for hypothesis $\mathcal{H}_1$: “the mean weight is larger than 1 kg”.
3. Answer again the two previous questions for a sample of 100 packs, with mean weight 1005 g.
4. On a sample of 10 packs, a mean weight of 1011 g has been observed, with an empirical standard deviation of 32 g. At threshold 1%, is this observation compatible with the value of 20 g for the theoretical standard deviation?
5. For the sample of the previous question, supposing the variance is unknown, can it be said that the packs are significantly too heavy on average at threshold 1%?

37. A paracetamol concentration of more than 150 mg per kilogram body weight is considered as dangerous. The mesurements of paracetamol in blood tests are modelled by a random variable with normal distribution $\mathcal{N}(\mu, \sigma^2)$. The standard-deviation, linked to the testing method, is supposed to be known and equal to 5 mg. For better assessment, 4 blood tests are usually made. The results are assumed to be independent realisations of the same normal distribution $\mathcal{N}(\mu, \sigma^2)$.

1. Give the hypotheses and the decision rule for the test deciding, at threshold 5%, whether a patient is at risk, on view of 4 blood tests. (you are a cautious doctor).
2. On a given patient, the 4 blood tests gave concentrations of 140, 133, 148, 144. Give the p-value for the test of the previous question. Is he at risk?
3. From now on, the standard-deviation is supposed *unknown*. Give the test statistic and the decision rule for the test deciding, at threshold 5%, whether a patient is at risk, on view of 4 blood tests.
4. For the patient of question 2, give an interval containing the p-value for the test of the previous question. What is your conclusion?

38. For a given population, the weight of newborn babies is modeled by a normal distribution. In the whole population, the standard deviation of newborn weights is 380 g. The average weight of a newborn whose mother does not smoke is 3400 g. In order to study the effect of tobacco, the babies of 30 mothers who smoked during pregnancy have been weighed, giving an empirical mean of 3240 g, with standard deviation 426 g.

1. Assuming that the standard deviation of the sample is known and equal to that of the whole population, calculate the p-value of the test deciding whether the newborns of the sample are significantly lighter on average. What is your conclusion, at threshold 5%?
2. Assuming that the standard deviation is unknown, give a test statistic and a decision rule, to test the same hypotheses as in the previous question. What is your conclusion?
3. Is the observed standard deviation significantly higher than that of the whole population?
4. Answer question 1. for a sample of 300 newborns, for which a mean weight of 3340 g has been observed.

39. The 15 lengths of 15 cuckoo eggs (expressed in millimeters) are as follows:

19.8, 22.1, 21.5, 20.9, 22.0, 21.0, 22.3, 21.0, 20.3, 20.9, 22.0, 22.0, 20.8, 21.2, 21.0 .

The following values are given:

$$\sum x_i = 318.8 \quad \text{and} \quad \sum x_i^2 = 6782.78 \ .$$

The length of a cuckoo egg is modeled by a random variable with distribution $\mathcal{N}(\mu, \sigma^2)$.

1. Find the empirical mean and variance of the sample.
2. Test the hypothesis $\mathcal{H}_0 : \sigma^2 = 0.4$ against $\mathcal{H}_1 : \sigma^2 > 0.4$, at threshold 5%.
3. Test the hypothesis $\mathcal{H}_0 : \mu = 21$" against $\mathcal{H}_1 : \mu > 21$, at threshold 5%.
4. Give an interval containing the p-value of the test in the previous question.

40. After a treatment of a certain species of rodents, a sample of 10 animals is selected and weighed. The weights in grams are the following

83 , 81 , 84 , 80 , 85 , 87 , 89 , 84 , 82 , 80 .

The following values are given:

$$\sum x_i = 835 \quad \text{and} \quad \sum x_i^2 = 69801 \ .$$

It is known that untreated rodents have an average weight of 87.6 g. The weight of a rodent is modeled by a normal random variable.

1. At threshold 5%, test the hypothesis that "the treatment has no effect on the mean weight" against "the treatment decreases the mean weight".

2. Give an interval containing the p-value of the test in the previous question.

41. A renting car company makes an experiment to decide between two types of tyres. Eleven cars are driven on a circuit with type A tyres. The tyres are then replaced by type B, and the cars are driven again on the same circuit. The consumption in liters per 100 km of these cars are modeled by Gaussian random variables. Here are the observations:

Car	1	2	3	4	5	6	7	8	9	10	11
Tyres A	4.2	4.7	6.6	7	6.7	4.5	5.7	6	7.4	4.9	6.1
Tyres B	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8	6.9	4.9	6

1. Assuming that the observed differences follow a normal distribution, what test statistic would you propose?

2. What hypotheses would you test to decide whether the tyres have an effect on gas consumption?

3. What hypotheses would you test to decide whether type B tyres are significantly better on average?

4. At threshold 5% what are your conclusions?

42. Nine patients with anxiety symptoms receive a sedative. The condition of the patient before and after treatment is evaluated by an index that the doctor calculates according to the answers to a series of questions. If the treatment is efficient, the index must decrease. The observed values of the index on the 9 patients are the following:

Before	1.83	0.5	1.62	2.48	1.68	1.88	1.55	3.06	1.3
After	0.88	0.65	0.59	2.05	1.06	1.29	1.06	3.14	1.29

1. What modeling assumption would you make and what hypotheses would you test?

2. Give an interval containing the p-value of the test deciding whether the sedative significantly improves the condition of the patients on average. What is your conclusion?

43. A factory must deliver rods the length of which is modeled by a normal distribution with expectation 40 mm. The rods cannot be used if they are smaller than 39 mm or longer than 41 mm, and the factory guarantees that less than 1% must be discarded.

1. Assuming that the machine produces rods having the right length on average, what should the standard deviation be in order to ensure that only 1% of the rods should be discarded?

2. On a sample of 15 rods, an empirical mean of 40.3 mm has been observed, with a standard deviation of 0.6 mm. Is the observed standard deviation significantly higher than the theoretical one of the previous question?
3. Are the rods significantly too long on average?
4. A customer claims he has received 112 useless rods out of a batch of 10000. Is he right to complain?

44. The percentage of 35-year-old women having wrinkles is 25%. Out of 200 women having followed an anti-wrinkle treatment, 40 of them still had wrinkles. At threshold 5%, can it be said that the treatment is efficient?

45. For a certain disease, there exists a treatment that cures 70% of the cases. A laboratory proposes a new treatment claiming that it is better than the previous one. Out of 200 patients having received the new treatment, 148 of them have been cured. As the expert in charge of deciding whether the new treatment should be authorized, what are your conclusions?

46. Here is the table of blood group frequencies in France:

Group Factor	O	A	B	AB
Rhesus +	0.370	0.381	0.062	0.028
Rhesus −	0.070	0.072	0.012	0.005

The blood transfusion center in Pau has observed the following distribution on 5000 blood donors.

Group Factor	O	A	B	AB
Rhesus +	2291	1631	282	79
Rhesus −	325	332	48	12

One wishes to answer statistically the questions below. For each question, the test statistic and the p-value should be calculated, and the conclusion given.

1. Is type O+ significantly more frequent in Pau?
2. Among people with positive rhesus, is the frequency of group O significantly different in Pau?
3. Among people with group O, is the frequency of the positive rhesus significantly higher in Pau?

3.3 Comparison of two independent samples

For the first sample:

$$\overline{X} = \frac{1}{n_x}\sum_{i=1}^{n_x} X_i \quad \text{and} \quad S_x^2 = \left(\frac{1}{n_x}\sum_{i=1}^{n_x} X_i^2\right) - \overline{X}^2 \ ,$$

the expectation of the unknown distribution is μ_x, its variance is σ_x^2.

For the second sample:

$$\overline{Y} = \frac{1}{n_y}\sum_{j=1}^{n_y} Y_j \quad \text{and} \quad S_Y^2 = \left(\frac{1}{n_y}\sum_{j=1}^{n_y} Y_j^2\right) - \overline{Y}^2 \ ,$$

the expectation of the unknown distribution is μ_y, its variance is σ_y^2.

The test statistics to be used and their distribution under the null hypothesis $\mathcal{H}_0$ are the following.

- Fisher test: comparison of variances, Gaussian sample.

  $$\mathcal{H}_0 : \sigma_x^2 = \sigma_y^2 \quad ; \quad T = \frac{\frac{n_x}{n_x-1}S_x^2}{\frac{n_y}{n_y-1}S_y^2} \text{ follows the Fisher distribution } \mathcal{F}(n_x-1, n_y-1) \ .$$

  If $T < 1$, swap X and Y (i.e. replace T by $1/T$) and compare to the quantile of order $1-\alpha/2$ of the Fisher distribution $\mathcal{F}(n_y-1, n_x-1)$.

- Student test: comparison of expectations, Gaussian sample.

  $$\mathcal{H}_0 : \mu_x = \mu_y \quad ; \quad T = \frac{\sqrt{n_x+n_y-2}}{\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}} \frac{\overline{X}-\overline{Y}}{\sqrt{n_x S_x^2 + n_y S_y^2}} \ ,$$

  follows the Student distribution $\mathcal{T}(n_x+n_y-2)$, si $\sigma_x = \sigma_y$.

- Comparison of expectations, large samples.

  $$\mathcal{H}_0 : \mu_x = \mu_y \quad ; \quad T = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{S_x^2}{n_x}+\frac{S_y^2}{n_y}}} \text{ follows the normal distribution } \mathcal{N}(0,1) \ .$$

Example. The question whether cooking with olive oil reduces risk of thrombosis is studied. For this, the logarithm of the D-dimer concentration, modeled by a normal distribution, is considered. A sample of 9 people regularly using sunflower oil, gives a mean of -0.78, with standard deviation 0.27. A sample of 13 people regularly using olive oil, gives a mean of -0.97, with standard deviation 0.32.

1. Test the equality of variances at threshold 0.05.

   The Fisher test has to be applied to see if the difference between the observed variances of the two samples is significant or not. The statistic of the test is calculated, first putting the the lower variance on the numerator:

   $$T = \frac{\frac{9}{8}0.27^2}{\frac{13}{12}0.32^2} = 0.7393\ .$$

   The hypothesis $\mathcal{H}_0 : \sigma_x = \sigma_y$ *must be tested against* $\mathcal{H}_1 : \sigma_x \neq \sigma_y$*. This is a two-tailed test, that rejects those values outside the interval* $[l, l']$*, where* l *and* l' *are the quantiles of order* 0.025 *and* 0.975 *of the distribution of* T *under* $\mathcal{H}_0$*, that is the Fisher distribution* $\mathcal{F}(8, 12)$*. However the quantile of order* 0.025 *for* $\mathcal{F}(8, 12)$ *is the inverse of the quantile of order* 0.975 *for the* $\mathcal{F}(12, 8)$*. Therefore it is simpler to swap* X *and* Y*, which amounts to computing* $1/T = 1.3526$*. This value must be compared to the quantile of order* 0.975 *for the Fisher distribution with parameters* 12 *and* 8 *(and not* 8 *and* 12 *since* X *and* Y *have been swapped). That bound is* 4.1997*. The observed value* 1.3526 *is lower, thus the hypothesis of equality of variances has to be accepted at threshold* 5%*.*

2. At threshold 0.05, what test would you propose to decide whether olive oil lowers significantly the risk of thrombosis? What is your conclusion? Give an interval containing the p-value.

   Having accepted the equality of variances, the application of the Student test for equality of expectations is justified. Denoting by X *the variable "logarithm of the D-dimer concentration for a person having sunflower oil", and* Y *the same variable for persons having olive oil, the following hypotheses must be tested:*

   $$\mathcal{H}_0 : \mu_x = \mu_y \quad \textit{against} \quad \mathcal{H}_1 : \mu_x > \mu_y\ .$$

   The test statistic is

   $$T = \frac{\sqrt{n_x + n_y - 2}}{\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \frac{\overline{X} - \overline{Y}}{\sqrt{n_x S_x^2 + n_y S_y^2}}\ ,$$

   for which high values will be rejected.

   $$\textit{Reject } \mathcal{H}_0 \iff T > l\ .$$

   The bound l *is such that a variable following the Student distribution with parameter* $9 + 13 - 2 = 20$ *is larger with probability* 0.05*, hence* $l = 1.7247$*. Here the test statistic takes a value of* 1.3055*, therefore the equality of expectations cannot be rejected: the decrease on average that has been observed is not significant at threshold* 5%*. The p-value is the probability that a variable following the Student distribution* $\mathcal{T}(20)$ *be larger than* 1.3055*. In the table,* 1.3055 *lies between the quantiles of order* 0.8 *and* 0.9*, close to that of order* 0.9*. Hence the p-value lies between* 0.1 *and* 0.2*. The numerical value is* 0.1033*.*

3. In another study on 110 sunflower oil users, a mean of -0.82 has been observed, with standard deviation 0.29, while 130 olive oil users, had a mean of -0.93, with standard deviation 0.31. Find the p-value of the test deciding whether the improvement is significant. At threshold 0.05, what is your conclusion?

   This is a test of expectation comparison on large samples. The test statistic is:

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}} ,$$

   that follows the $\mathcal{N}(0,1)$ distribution under the hypothesis $\mathcal{H}_0$. The calculated value is 2.8366. The p-value is the probability for a variable with $\mathcal{N}(0,1)$ distribution to be larger than 2.8366, that is 0.0023. At any threshold lower than 0.23% (and in particular of course at thresholds 5% and 1%), the null hypothesis $\mathcal{H}_0$ is rejected, and it can be concluded that olive oil significantly decreases the risk of thrombosis.

47. The activity of the seric PDE enzyme is studied, depending on different factors. The results are expressed in international unit per liter of serum. For two groups of women, pregnant or not, the following results were obtained:

non pregnant	1.5	1.6	1.4	2.9	2.2	1.8	2.7	1.9
pregnant	4.2	5.5	4.6	5.4	3.9	5.4	2.7	3.9

non pregnant	2.2	2.8	2.1	1.8	3.7	1.8	2.1
pregnant	4.1	4.1	4.6	3.9	3.5		

(Indications: $\sum x_i = 32.5$, $\sum x_i^2 = 75.83$, $\sum y_i = 55.8$, $\sum y_i^2 = 247.32$).

1. Describe the modeling hypotheses.

2. Test the equality of variances at threshold 5%.

3. Can it be claimed that the seric PDE enzyme activity is significantly different between pregnant and non pregnant women?

4. Can it be claimed that the seric PDE enzyme activity is significantly higher among pregnant women?

48. The IQ's of 9 children in a district of a large city have empirical mean 107 and standard deviation 10. The IQ's of 12 children in another district have empirical mean 112 and standard deviation 9.

1. Describe the modeling hypotheses.

2. Test the equality of variances at threshold 5%.

3. Can it be said the children in the second district have significantly higher IQ's than those in the first district? Give an interval containing the p-value for the test.

49. The maximal tensions of gastrocnemian muscles (expressed in g) for the frog vary, according whether the nerves have been removed or not. During an experiment made on 9 frogs, the following measures have been obtained:

With nerves	75	96	32	41	50	39	59	45	30
Without nerves	53	67	32	29	35	27	37	30	21

1. Describe the modeling hypotheses.
2. Test the equality of variances at threshold 5%.
3. At threshold 5%, can it be said that the mean maximal tension is different among the two groups? Give an interval containing the p-value of that test.

50. During a study comparing different sampling methods for forest soil, the K_2O concentration has been measured, on the one hand for 20 samples of soil individually extracted, and on the other hand for 10 mixed samples, each obtained from 25 differents soils. For the individual samples, the following values have been found:

$$\sum x_i = 259.2 \quad \text{and} \quad \sum x_i^2 = 3662.08 \ ,$$

and for the mixed samples:

$$\sum y_i = 109.2 \quad \text{and} \quad \sum y_i^2 = 1200.8 \ .$$

It can be expected that the two sampling methods give very different variances. Justify intuitively why, and prove it using the Fisher test.

51. To determine the average weights of ears of two kinds of ears of wheat, 9 ears of each kind are weighed. The empirical mean and variance for the two samples are:

$$\overline{x} = 170.7 \ ; \quad \overline{y} = 168.5 \ ; \quad s_x^2 = 432.90 \ ; \quad s_y^2 = 182.70 \ .$$

1. Describe the modeling hypotheses.
2. Test the equality of variances at threshold 5%.
3. Give an interval containing the p-value for the test deciding whether the two kinds are significantly different. What is your conclusion?

52. In a farming cooperative, the effect of a fertilizer on wheat production is to be tested. For this, 2000 plots of land of the same size are chosen. Half of them are treated with the fertilizer, the other half are not. The crops in tons for the untreated plots give $\sum x_i = 61.6$, $\sum x_i^2 = 292.18$ and for the fertilized plots $\sum y_i = 66.8$, $\sum y_i^2 = 343.48$. Test the hypothesis “the fertilizer is not efficient”, against “the fertilizer is efficient” at thresholds 0.01 and 0.05.

53. In a city A, 36 people out of 300 smoke at least two packs a day. In city B, 8 out of 100 smoke at least two packs a day. The following hypotheses must be tested: $\mathcal{H}_0$: "there is no difference between the two cities" against $\mathcal{H}_1$: "more people smoke more than two packs a day in city A than in B".

1. Let p_A (respectively p_B) denote the proportion of people smoking more than two packs a day in A (respectively in B). What variables would you propose to model the problem? Give their expectations and variances as a function of p_A and p_B.
2. What test would you propose for $\mathcal{H}_0$ against $\mathcal{H}_1$?
3. Give the p-value for that test. What is your conclusion?

54. Let p_A be the probability of curing a certain disease by treatment A. A group of 50 patients had that treatment and 28 were cured. Another treatment B cures that same disease with probability p_B. Out of 60 patients having had that new treatment, 38 were cured.

1. What test would you propose to decide whether treatment B is better than treatment A?
2. Give the p-value for that test. What is your conclusion?

3.4 The chi-squared adjustment test

Let r be the number of classes. For $i = 1, \dots, r$, denote by n_i the *observed* number of class i, and np_i its *theoretical* number.

- The chi-squared test statistic is:

$$T = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} \; .$$

- Under the null hypothesis where the theoretical model is true, T follows the chi-squared distribution with parameter $d = r-1-k$:
  - ⋆ r is the number of classes,
  - ⋆ k is the number of parameters that have been estimated from the data to establish the theoretical model.
- The test applies to a large sample ($n \geqslant 50$). The theoretical numbers in each class must be large enough ($np_i \geqslant 8$). If necessary, classes can be grouped to fulfill the second condition.

Example. White-flowered sweet peas are crossed with red-flowered sweet peas. The colors of 600 plants from the second generation are distributed as follows:

Phenotype	Red	Pink	White
Number	141	325	134

With the 600 plants, 150 bunches of 4 are made, out of which the number of white-flowered plants is counted. The observed numbers were the following.

White flowers	0	1	2	3	4
Numbers	53	68	23	4	2

1. Give the theoretical proportions of the Mendelian distribution for the three colors. Find the chi-squared test statistic. Give an interval containing the p-value. What is your conclusion?

   Denote by R the allele inducing the red color and by B the allele inducing the white color. It is supposed that the phenotypes "red flowers", "pink flowers", and "white flowers" correspond respectively to genotypes RR, RB, and BB. If genotype RR is crossed with BB, all offsprings at the first generation are RB hybrids. At the second generation, crossing two hybrids should produce one fourth genotypes RR, half genotypes RB, one fourth genotypes BB; therefore, one fourth red flowered plants, half pink flowered and one fourth white flowered should be observed. The theoretical numbers would be 150, 300, 150.

   The chi-squared test statistic takes a value of:

   $$T = \frac{(141-150)^2}{150} + \frac{(325-300)^2}{300} + \frac{(134-150)^2}{150} = 4.33\ .$$

   This value should be compared to the quantiles of the chi-squared distribution with parameter $3-1=2$. The p-value is the probability for a variable following the $\mathcal{X}^2(2)$ distribution to be larger than 4.33. *According to the table, the p-value lies between* 0.1 *and* 0.5. *The exact value is* 0.1147. *The hypothesis that the empirical distribution agrees with the theoretical one is accepted.*

2. What theoretical model would you propose for the number of white flowered plants in a bunch of 4? Make the appropriate grouping of classes. Find the chi-squared test statistic. Give an interval containing the p-value. What is your conclusion?

   If the bunches are formed at random, the distribution of the number of white flowered plants in a bunch of 4 is the binomial distribution with parameters 4 *(total number of plants) and* 1/4 *(theoretical proportion of white flowered plants). For $i = 0, \ldots, 4$, the theoretical number for the number of bunches with i white flowered plants is:*

   $$np_i = 150 \binom{4}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{4-k}\ .$$

White flowers	0	1	2	3	4
Observed number	53	68	23	4	2
Theoretical number	47.46	63.28	31.64	7.03	0.59

In order to reach a theoretical distribution with at least 8 *in each class, the last three classes may be grouped.*

White flowers	0	1	$2,3,4$
Observed number	53	68	29
Theoretical number	47.46	63.28	39.26

The chi-squared test statistic takes a value of 3.6786*. The p-value is the probability for a variable following the* $\mathcal{X}^2(2)$ *distribution, to be larger than* 3.6786*. According to the table, the p-value is between* 0.1 *and* 0.2*. The exact value is* 0.1589*. The hypothesis that the empirical distribution agrees with the theoretical one is accepted.*

3. Let $\widehat{p}$ be the observed proportion of white flowered plants. For 4-plant bunches, test the adjustment of the observed distribution with the binomial distribution $\mathcal{B}(4, \widehat{p})$: calculate the test statistic and give an interval containing the p-value.

The total number of white flowered plants is 134*, their proportion is* $\widehat{p} = \frac{134}{600} \simeq 0.2233$*. The theoretical numbers are now calculated using the* $\mathcal{B}(4, \widehat{p})$ *distribution.*

White flowers	0	1	$2,3,4$
Observed number	53	68	29
Theoretical number	54.59	62.78	32.64

The chi-squared test statistic takes a value of 0.8855*. Since one parameter has been estimated to establish the theoretical distribution, the parameter of the chi-squared distribution is* $3 - 1 - 1 = 1$*. According to the table, the p-value is between* 0.3 *and* 0.4*, the exact value is* 0.3467*. The hypothesis that the empirical distribution agrees with the theoretical one is accepted.*

55. Here is the frequency table of blood types in France:

Groupe Factor	O	A	B	AB
Rhesus +	0.370	0.381	0.062	0.028
Rhesus –	0.070	0.072	0.012	0.005

The transfusion center of Pau has observed the following distribution out of 5000 blood donors.

Group Factor	O	A	B	AB
Rhesus +	2291	1631	282	79
Rhesus –	325	332	48	12

The following questions need to be answered statistically. For each question, write down the table of observed and theoretical distributions, calculate the test statistic, give an interval containing the p-value, and specify your conclusion.

1. Is the distribution of the 8 group-rhesus types in Pau different from the overall French distribution?

2. Is the distribution of the two rhesuses in Pau different from the overall French distribution?

3. Among type O people, is the distribution of the two rhesuses in Pau different from the overall French distribution?

4. Among people with positive rhesus, is the distribution of the four groups in Pau different from the overall French distribution?

5. Among people with negative rhesus, is the distribution of the four groups in Pau different from the overall French distribution?

56. A group of 162 students have been asked to evaluate the time they spend cooking per month:

Hours	[0 ; 5[	[5 ; 10[	[10 ; 15[	$\geqslant 15$
Students	63	49	19	31

Previous studies in the overall population established the following distribution:

Hours	[0 ; 5[	[5 ; 10[	[10 ; 15[	$\geqslant 15$
Proportion	40%	35%	15%	10%

Test the adjustment of the observed distribution among students to that of the general population. Give an interval containing the p-value. What is your conclusion?

57. The sleeping time of twelve year-old children is studied. On a sample of size $n = 50$ the sleeping times (expressed in hours) have been recorded. The following values are given: $\sum x_i = 424$ and $\sum x_i^2 = 3828$, with the following distribution in classes.

Class	$\leqslant 8$	]8 ; 9]	]9 ; 10]	> 10
Number	19	12	9	10

1. It is generally considered that the sleeping time of a 12 year old child follows a $\mathcal{N}(9, 3)$ distribution. Apply the adjustment test of the observed distribution with the theoretical one. Give the value taken by the test statistic, an interval for the p-value and your conclusion.

2. Find the empirical mean $\overline{x}$ and variance s^2. Answer the previous question again, replacing the $\mathcal{N}(9, 3)$ distribution by $\mathcal{N}(\overline{x}, s^2)$.

58. A biometry study on the length of cuckoo eggs has produced the fol-lowing results. The following values are given: $n = 152$, $\sum x_i = 6200$, $\sum x_i^2 = 255200$, with the following distribution by classes:

class	< 32	$[32; 34[$	$[34; 36[$	$[36; 38[$	$[38; 40[$	$[40; 42[$	$[42; 44[$	$[44; 46[$	$[46; 48[$	$\geqslant 48$
number	2	7	6	18	25	40	23	20	6	5

1. Previous studies had shown that the length of a cuckoo egg can be modeled by a normal distribution with expectation 40 and standard deviation 4. Apply the adjustment test of the observed distribution with the theoretical one. Give the value taken by the test statistic, an interval for the p-value and your conclusion.

2. Find the empirical mean $\overline{x}$ and variance s^2. Answer the previous question again, replacing the $\mathcal{N}(40, 4^2)$ distribution by $\mathcal{N}(\overline{x}, s^2)$.

3.5 The chi-squared independence test

This is a particular case of the chi-squared adjustment test, that allows to test the mutual independence of two discrete variables.

- The *contingency table* presents the *joint numbers.* At line i, column j, the table gives n_{ij}, i.e. the number of individuals in class i for the first variable and class j for the second. If the number of modalities for the two variables are r and s, the table has r lines and s columns.

- The *marginal numbers* are the sums by line or column in the contingency table; $n_{i\bullet} = \sum_j n_{ij}$ is the total number of individuals in class i for the first variable; $n_{\bullet j} = \sum_i n_{ij}$ is the total number of individuals in class j for the second variable. The total number is $n = \sum_i n_{i\bullet} = \sum_j n_{\bullet j}$.

- The test statistic is:
$$T = n\left(-1 + \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i\bullet}n_{\bullet j}}\right) .$$

- Under the null hypothesis where both variables are independent, T follows the chi-squared distribution with parameter $d = (r-1)(s-1)$.

Example. The blood transfusion center in Pau has observed the following type distribution on 5000 blood donors.

Group Factor	O	A	B	AB
Rhesus +	2291	1631	282	79
Rhesus −	325	332	48	12

1. Complete the contingency table with the marginal numbers.

   The table gives the joint numbers. The marginal numbers are obtained by summing over lines and columns.

Group *Factor*	*O*	*A*	*B*	*AB*	*Total*
Rhesus +	2291	1631	282	79	4283
Rhesus −	325	332	48	12	717
Total	2616	1963	330	91	5000

2. Find the value of the chi-squared independence test statistic.

$$T = 5000\left(-1 + \frac{2291^2}{2616 \times 4283} + \cdots + \frac{12^2}{717 \times 91}\right) = 18.5104\ .$$

3. At threshold 1% what is your conclusion?

   Under the independence hypothesis, the test statistic follows the the chi-squared distribution with parameter $(4-1)(2-1) = 3$*. The quantile of order* 0.99 *for this distribution is* 11.3449*. Since* 18.5104 *is larger, the conclusion is that there is a dependence between the blood group and rhesus. The exact p-value is* 0.000345.

59. Two treatments A and B are available for a certain disease. The observed results for the evolution of the disease with the two treatments on two groups of patients are as follows:

Effect Treatment	Cured	Improved	Stationary
A	280	210	110
B	220	90	90

1. Complete the contingency table.

2. Find the value of the independence test statistic.

3. Give an interval containing the p-value for the chi-squared test of independence. Would you say that treatments A and B yield significantly different results?

60. During 10 years, 240 persons have been followed. Among them:
- 110 had sunflower oil on a regular basis
- 25 had olive oil and cardio-vascular troubles
- 78 had sunflower oil and did not have cardio-vascular troubles.

1. Write the contingency table matching these observations.

2. Find the value of the independence test statistic.

3. Give an interval containing the p-value for the chi-squared test of independence. Would you say that cardio-vascular troubles are independent of the type of oil?

61. The observations of a couple (X, Y) of physiological variables for 100 individuals in a population has led, after the choice of two classes for X and three for Y, to the following contingency table.

Y X	1	2	3	Total
1	4	11	7	22
2	16	39	23	78
Total	20	50	30	100

1. Find the value of the independence test statistic.

2. Give an interval containing the p-value for the test. What is your conclusion?

62. After the treatment for a certain disease, 40 young patients out of 70, and 50 old patients out of 100 have shown improvements.

1. Write the contingency table matching these observations.

2. Find the value of the independence test statistic.

3. Give an interval containing the p-value for the test. Would you say that the effect of the treatment depends on the age of the patient?

63. The following contingency table concerns 592 women grouped accord-ing to their eye and hair color.

Hair Eyes	Black	Brown	Red	Blond
Brown	68	119	26	7
Hazel	15	54	14	10
Green	5	29	14	16
Blue	20	84	17	94

1. Complete this contingency table.

2. Find the value of the independence test statistic.

3. Give an interval containing the p-value for the test. Would you say that the hair color is independent of the eye color?

Linear regression

4.1 Regression line and prediction

The data are n couples of real numbers. The first coordinate is a variable considered as *deterministic* and *explaining*. The second one is considered as *random* and *explained*. The following quantities are calculated:

- mean of explaining variable: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- mean of explained variable: $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- variance of explaining variable: $s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$
- variance of explained variable: $s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2$
- *covariance* of the two variables: $c_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}$
- *correlation coefficient*: $r_{xy} = \frac{c_{xy}}{\sqrt{s_x^2\, s_y^2}}$.
- *slope* of the regression line: $\widehat{a} = \frac{c_{xy}}{s_x^2}$
- *intercept with y-axis*: $\widehat{b} = \overline{y} - \widehat{a}\,\overline{x}$
- *estimated variance*: $\widehat{\sigma}^2 = \frac{n}{n-2} s_y^2 (1 - r_{xy}^2)$
- *prediction* of an ordinate for a given abscissa x_*: $y_* = \widehat{a}\, x_* + \widehat{b}$.

Example. In order to measure the dependence between age and thrombosis risk, 12 patients have been considered. Their age in years (variable X) and the logarithm of D-dimer concentration (variable Y) are known. The following quantities are given:

$$\sum x_i = 596 \ ; \ \sum x_i^2 = 32435 \ ; \ \sum y_i = -5.2 \ ; \ \sum y_i^2 = 4.3 \ ; \ \sum x_i y_i = -188.58 \ .$$

1. Find the correlation coefficient between X and Y.

$$\overline{x} = 49.667 \ ; \quad \overline{y} = -0.43333 \ ; \quad s_x^2 = 236.139 \ ; \quad s_y^2 = 0.17056 \ ;$$
$$c_{xy} = 5.8072 \ ; \quad r_{xy} = 0.91506 \ .$$

Note that r_{xy} is close to 1, indicating a strong correlation.

2. Find the equation for the regression line for Y onto X.

$$\widehat{a} = 0.02459 \; ; \quad \widehat{b} = -1.6548 \; .$$

The equation of the regression line is $y = 0.02459\,x - 1.6548$*. It is increasing* $(a > 0)$ *because the correlation is positive: the D-dimer concentration logarithm tends to increase with the age.*

3. Find the estimated variance of the regression.

$$\widehat{\sigma}^2 = 0.0333 \; .$$

4. What value for Y would you predict for a 60 year old person?

The prediction for $x_* = 60$ *is* $y_* = 0.02459 \times 60 - 1.6548 = -0.1792$*.*

64. The air pollution has been recorded in 41 American cities by the vari-able Y, measuring the volume of SO_2 in micrograms per m^3 of air, as well as the average annual temperature X in degrees Fahrenheit. The following numerical results are given:

$$\sum x_i = 2286 \; ; \; \sum y_i = 1232 \; ; \; \sum x_i^2 = 129549 \; ; \; \sum y_i^2 = 59050 \; ; \; \sum x_i y_i = 74598 \; .$$

1. Find the correlation coefficient of X and Y.

2. Give the equation of the regression line for Y onto X.

3. What value for Y do you predict for a city where the average annual temperature is 60°F?

65. In a study of the duration of the vegetation period in mountains, weather stations have been installed at different altitudes. The average temperature (variable Y in degrees Celsius) and the altitude (variable X in meters) for each station are given in the table below.

altitude	1040	1230	1500	1600	1740	1950	2200	2530	2800	3100
temperature	7.4	6	4.5	3.8	2.9	1.9	1	-1.2	-1.5	-4.5

The following values are given:

$$\sum x_i = 19690; \; \sum y_i = 20.3; \; \sum x_i^2 = 42925500; \; \sum y_i^2 = 162.41; \; \sum x_i y_i = 17671 \, .$$

1. Find the correlation coefficient.

2. Find the estimates for the parameters a, b and σ^2 of the regression of Y onto X.

3. What average temperature do you predict at 1100 m?

66. The gain in weight of a young sheep in one year (variable Y in kilo-grams) is thought to depend on the initial weight (variable X also in kilograms). For 10 sheep, the following results have been obtained:

$$\sum x_i = 406 \ ; \ \sum y_i = 423 \ ; \ \sum x_i^2 = 16570 \ ; \ \sum y_i^2 = 18057 \ ; \ \sum x_i y_i = 17280 \ .$$

1. Find the correlation coefficient.
2. Estimate the parameters a, b and σ^2 for the regression of Y onto X.
3. According to this model, how much weight should a sheep with an initial weight of 50 kg gain? Same question for a 30 kg sheep.

67. The volume Y of exhaled air is a standard measurement for lung condition. In order to identify a population with weak lung condition, a model for the normal lung condition is needed. In order to do this, the volume Y in liters and the size X in centimeters for 12 boys between 10 and 15 years old were measured.

The following values were obtained:

$$\sum x_i = 1872 \ ; \ \sum y_i = 32.3 \ ; \ \sum x_i^2 = 294320 \ ; \ \sum y_i^2 = 93.11 \ ; \ \sum x_i y_i = 5156.20 \ .$$

1. Find the correlation coefficient.
2. Find the estimates for the coefficients of the regression line of Y onto X, and its variance.
3. What volume of air should a boy 1.60 m tall exhale?

68. One wishes to predict the height H of a tree as a function of the diameter D of its trunk. To make a linear regression, the logarithms of the variables are used: $Y = \ln(H)$ and $X = \ln(D)$. Here are the results for 5 trees:

X	-1.61	-1.20	-0.97	-0.51	-0.42
Y	2.22	2.27	2.38	2.60	2.65

The following results are given:

$$\sum x_i = -4.71 \ ; \ \sum y_i = 12.12 \ ; \ \sum x_i^2 = 5.4095 \ ;$$

$$\sum y_i^2 = 29.5282 \ ; \ \sum x_i y_i = -11.0458 \ .$$

1. Find the correlation coefficient of X and Y.
2. Give the equation of the regression line of Y onto X.
3. Find the predicted height for a tree with a trunk diameter 0.7.

4.2 Confidence and prediction intervals

The intervals given in what follows have level $1-\alpha$, and t_α is the quantile of order $1-\alpha/2$ for the Student distribution $\mathcal{T}(n-2)$.

- Confidence interval for the slope a:

$$\left[\widehat{a} \pm t_\alpha \sqrt{\frac{\widehat{\sigma}^2}{n s_x^2}}\right] .$$

- Confidence interval for $ax_* + b$:

$$\left[\widehat{a}x_* + \widehat{b} \pm t_\alpha \sqrt{\frac{\widehat{\sigma}^2(s_x^2 + (x_* - \overline{x})^2)}{n s_x^2}}\right] .$$

- *Prediction* interval for $Y_* = ax_* + b + E$:

$$\left[\widehat{a}x_* + \widehat{b} \pm t_\alpha \sqrt{\frac{\widehat{\sigma}^2((n+1)s_x^2 + (x_* - \overline{x})^2)}{n s_x^2}}\right] .$$

Example. In order to measure the dependence between age and thrombosis risk, 12 patients have been observed. Their age in years (variable X) and the logarithm of D-dimer concentration (variable Y) are known. The following quantities are given:

$$\sum x_i = 596 \; ; \; \sum x_i^2 = 32435 \; ; \; \sum y_i = -5.2 \; ; \; \sum y_i^2 = 4.3 \; ; \; \sum x_i y_i = -188.58 \; .$$

1. Give a 99% confidence interval for the slope of the regression line.

   The quantile of order 0.995 *for the Student distribution with parameter* $12-2 = 10$ *is* 3.169. *The confidence interval is* $[0.0137\,;\,0.0355]$.

2. Give a 99% confidence interval for the intercept of the regression line.

   A confidence interval for b is obtained by letting $x_* = 0$ *in the formula giving the confidence interval for* $ax_* + b$. *The desired interval is* $[-2.2195\,;\,-1.0900]$.

3. Give a 99% confidence interval for the average value of Y among 60 year old people.

   We want a confidence interval for $ax_* + b$, *with* $x_* = 60$. *The interval is* $[-0.380\,;0.022\,]$.

4. Give a prediction interval at level 0.99 for the value of Y of one 60 year old person in particular.

   We want a prediction interval for $Y_* = ax_* + b + E$, *with* $x_* = 60$. *The interval* $[-0.791\,;0.433\,]$. *Take care not to confuse:*

- estimation *of the* mean *value of D-dimer concentration among* all 60 *year old persons*
- prediction *of the value of the D-dimer concentration for* one 60 *year old person in particular.*

In the second case, the interval is necessarily wider than in the first.

69. The air pollution has been recorded in 41 American cities through the variable Y, measuring the volume of SO_2 in micro-grams per m^3 of air, as well as the average annual temperature X in degrees Fahrenheit. The following numerical results are given:

$$\sum x_i = 2286, \sum y_i = 1232, \sum x_i^2 = 129549, \sum y_i^2 = 59050, \sum x_i y_i = 74598 .$$

1. Give a 95% confidence interval for the slope and intercept of the regression line.
2. Give a 95% confidence interval for the mean value of Y in the cities where the average temperature is 60°F.
3. Give a prediction interval at level 0.95 for the value of Y in a given city where the average temperature is 60°F.

70. In a study of the duration of the vegetation period in mountains, weather stations have been installed at different altitudes. The average temperature (variable Y in degrees Celsius) and the altitude (variable X in meters) for each station are given in the table below.

altitude	1040	1230	1500	1600	1740	1950	2200	2530	2800	3100
temperature	7.4	6	4.5	3.8	2.9	1.9	1	-1.2	-1.5	-4.5

The following results are given:

$$\sum x_i = 19690; \sum y_i = 20.3; \sum x_i^2 = 42925500; \sum y_i^2 = 162.41; \sum x_i y_i = 17671 .$$

1. Give a 95% confidence interval for the slope and the intercept of the regression line.
2. Give a 95% confidence interval for the mean temperature at 1100 m.
3. Give a prediction interval at level 0.95 for the temperature of a given place at 1100 m.

71. The gain in weight of a young sheep in one year (variable Y in kilo-grams) is thought to depend on the initial weight (variable X also in kilograms). For 10 sheep, the following results are given:

$$\sum x_i = 406 ; \sum y_i = 423 ; \sum x_i^2 = 16570 ; \sum y_i^2 = 18057 ; \sum x_i y_i = 17280 .$$

1. Give a 99% confidence interval for the two coefficients of the regression line.
2. Give a 99% confidence interval for the mean weight gain of sheep initially weighing 30 kg.
3. Give a prediction interval at level 0.99 for the weight gain of a given sheep initially weighing 30 kg.

72. The volume Y of exhaled air is a standard measurement for lung condition. In order to identify a population with weak lung condition, a model for the normal lung condition is needed. In order to do this, the volume Y in liters and the size X in centimeters for 12 boys between 10 and 15 years old were measured.

The following numerical values were obtained:

$$\sum x_i = 1872 \ ; \ \sum y_i = 32.3 \ ; \ \sum x_i^2 = 294320 \ ; \ \sum y_i^2 = 93.11 \ ; \ \sum x_i y_i = 5156.20 \ .$$

1. Give a 99% confidence interval for the slope and intercept of the regression line.
2. Give a 99% confidence interval for the mean volume of air exhaled by boys measuring 1.60 m.
3. Give a prediction interval at level 0.99 for the volume of air exhaled by one given boy measuring 1.60 m.

73. One wishes to predict the height H of a tree as a function of the diameter D of its trunk. To make a linear regression, the logarithms of the variables are used: $Y = \ln(H)$ and $X = \ln(D)$. Here are the results for 5 trees:

X	-1.61	-1.20	-0.97	-0.51	-0.42
Y	2.22	2.27	2.38	2.60	2.65

The following results are given:

$$\sum x_i = -4.71, \ \sum y_i = 12.12, \ \sum x_i^2 = 5.4095,$$

$$\sum y_i^2 = 29.5282, \ \sum x_i y_i = -11.0458.$$

1. Give a 95% confidence interval for the two coefficients of the regression line.
2. Give a 95% confidence interval for the mean height of trees with diameter 0.7.
3. Give a prediction interval at level 0.95 for the height of one given tree with diameter 0.7.

4.3 Tests on a regression

Under the hypothesis $\mathcal{H}_0$, the model is $Y = ax + b + E$, where E follows the normal distribution $\mathcal{N}(0, \sigma^2)$. The parameters a, b and σ^2 are unknown. They are estimated by $\widehat{a}$, $\widehat{b}$ and $\widehat{\sigma}^2$. To test particular values, the following results are used. They give the distribution of the test statistic under $\mathcal{H}_0$.

- $\sqrt{\dfrac{ns_x^2}{\widehat{\sigma}^2}}\Big(\widehat{a} - a\Big)$ follows the $\mathcal{T}(n-2)$.
- $\sqrt{\dfrac{ns_x^2}{\widehat{\sigma}^2(s_x^2 + (x_* - \overline{x})^2)}}\Big(\widehat{a}x_* + \widehat{b} - ax_* - b\Big)$ follows the $\mathcal{T}(n-2)$.
- $\sqrt{\dfrac{ns_x^2}{\widehat{\sigma}^2((n+1)s_x^2 + (x_* - \overline{x})^2)}}\Big(Y_* - \widehat{a}x_* - \widehat{b}\Big)$ follows the $\mathcal{T}(n-2)$.
- $(n-2)\dfrac{\widehat{\sigma}^2}{\sigma^2}$ follows the $\mathcal{X}^2(n-2)$.

The *pertinence* or validity test for the regression consists of testing $\mathcal{H}_0 : a = 0$ against $\mathcal{H}_1 : a \neq 0$, by using the first result. The regression is declared valid by rejecting $\mathcal{H}_0$.

Example. In order to measure the dependence between age and thrombosis risk, 12 patients have been observed. Their age in years (variable X) and the logarithm of D-dimer concentration (variable Y) are known. The following quantities are given:

$$\sum x_i = 596 \ ; \ \sum x_i^2 = 32435 \ ; \ \sum y_i = -5.2 \ ; \ \sum y_i^2 = 4.3 \ ; \ \sum x_i y_i = -188.58 \ .$$

1. Test the pertinence of the regression at threshold 1%.

   This is a two-tailed test of $\mathcal{H}_0 : a = 0$ *against* $\mathcal{H}_1 : a \neq 0$. *The test statistic is:*

   $$T = \sqrt{\frac{ns_x^2}{\widehat{\sigma}^2}}\,\widehat{a} \ .$$

   Under hypothesis $\mathcal{H}_0$, T *follows the Student distribution with parameter* 10. *The decision rule is:*

   $$\textit{Reject } \mathcal{H}_0 \implies T \notin [-t_\alpha \, ; +t_\alpha] \ ,$$

   where t_α *is the quantile of order* $1-\alpha/2$ *for the Student distribution* $\mathcal{T}(10)$, *that is* 3.169. *Here, the value taken by* T *is* 7.177. *The null hypothesis* $\mathcal{H}_0$ *is rejected, the pertinence of the regression is accepted.*

2. Previous studies had shown a linear dependence betwen age and logarithm of D-dimer concentration of the form $Y = 0.02x - 2$. At threshold 1%, test whether the values of a and b previously found can still be accepted.

We first test $\mathcal{H}_0 : a = 0.02$ against $\mathcal{H}_1 : a \neq 0.02$. The test statistic is:

$$T = \sqrt{\frac{ns_x^2}{\widehat{\sigma}^2}}\,(\widehat{a} - 0.02)\ .$$

It takes a value of 1.341 *which is in the interval* $[-3.169\,;\,+3.169]$*. Thus we accept $\mathcal{H}_0$ (declare that the estimated value of a is not significantly far from* 0.02*).*

We now test $\mathcal{H}_0 : b = -2$ against $\mathcal{H}_1 : b \neq -2$. The test statistic is:

$$T = \sqrt{\frac{ns_x^2}{\widehat{\sigma}^2(s_x^2 + (0 - \overline{x})^2)}}(\widehat{b} - (-2))\ .$$

It takes a value of 1.935 *which is in the interval* $[-3.169\,;\,+3.169]$*. Thus we accept $\mathcal{H}_0$ (declare that the estimated value of b is not significantly far from* -2*).*

Finally, both tests accept the previously admitted values for a and b.

3. A 60 year old patient has a value of Y equal to 0.14: should he worry?

   We are testing here a value of $Y_* = ax_* + b + E$, with $x_* = 60$. The test statistic is:

$$T = \sqrt{\frac{ns_x^2}{\widehat{\sigma}^2((n+1)s_x^2 + (x_* - \overline{x})^2)}}(Y_* - \widehat{a}x_* - \widehat{b})\ .$$

   It takes a value of 5.028*. This value is higher than the quantile of order* 0.0005 *for the $\mathcal{T}(10)$ distribution, therefore it is unusually high (by reference to the available data). The patient should see a doctor.*

4. At threshold 1% test the hypothesis $\mathcal{H}_0 : \sigma^2 = 0.03$ against $\mathcal{H}_1 : \sigma^2 > 0.03$.

   The test statistic is:

$$10\frac{\widehat{\sigma}^2}{0.03}\ .$$

   Under hypothesis $\mathcal{H}_0$, T follows the chi-squared distribution with parameter 10*. At threshold* 1%*, the test rejects those values higher than the quantile of order* 0.99 *for the $\mathcal{X}^2(10)$ distribution, that is* 23.21*. Here, T takes a value of* 11.09*, therefore $\mathcal{H}_0$ is accepted.*

74. The air pollution has been recorded in 41 American cities through the variable Y, measuring the volume of SO_2 in micro-grams per m^3 of air, as well as the average annual temperature X in degrees Fahrenheit. The following numerical results are given:

$$\sum x_i = 2286,\ \sum y_i = 1232,\ \sum x_i^2 = 129549,\ \sum y_i^2 = 59050,\ \sum x_i y_i = 74598\ .$$

1. Test the pertinence of the regression at threshold 5%.

2. Test $\mathcal{H}_0 : a = 3$ against $\mathcal{H}_1 : a < 3$, at threshold 5%.
3. You were asked to fix an upper bound for the pollution of a city the average temperature of which is 60°F. You want this upper bound to be overpassed only in 5% of the cases. What upper bound will you choose?

75. During a study on the duration of the vegetation period in mountains, weather stations have been installed at different altitudes. The average temperature (variable Y in degrees Celsius) and the altitude (variable X in meters) for each station are given in the table below.

altitude	1040	1230	1500	1600	1740	1950	2200	2530	2800	3100
temperature	7.4	6	4.5	3.8	2.9	1.9	1	-1.2	-1.5	-4.5

The following results are given:

$$\sum x_i = 19690;\ \sum y_i = 20.3;\ \sum x_i^2 = 42925500;\ \sum y_i^2 = 162.41;\ \sum x_i y_i = 17671\,.$$

1. Test the pertinence of the regression at threshold 1%.
2. In one given place at 1100 meters, a mean temperature of 3.2 degrees has been observed. At threshold 1%, would you say that it is unusually cold up there?

76. The gain in weight of a young sheep in one year (variable Y in kilo-grams) is thought to depend on the initial weight (variable X also in kilograms). For 10 sheep, the following results are given:

$$\sum x_i = 406\ ;\ \sum y_i = 423\ ;\ \sum x_i^2 = 16570\ ;\ \sum y_i^2 = 18057\ ;\ \sum x_i y_i = 17280\ .$$

1. Test the pertinence of the regression at threshold 1%.
2. An old saying states that the weight of a sheep should double in one year. At threshold 1%, can you confirm?
3. A sheep with initial weight 30 kg, has gained only 20 kg after one year. At threshold 1%, should the shepherd worry?

77. The volume Y of exhaled air is a standard measurement for lung condition. In order to identify a population with weak lung condition, a model for the normal lung condition is needed. In order to do this, the volume Y in liters and the size X in centimeters for 12 boys between 10 and 15 years old were measured.

The following numerical results were obtained:

$$\sum x_i = 1872\ ;\ \sum y_i = 32.3\ ;\ \sum x_i^2 = 294320\ ;\ \sum y_i^2 = 93.11\ ;\ \sum x_i y_i = 5156.20\ .$$

1. Test the pertinence of the regression at threshold 1%.

2. A 1.60 m boy exhales 2.1 litres: should he worry?

78. One wishes to predict the height H of a tree as a function of the diameter D of its trunk. To make a linear regression, the logarithms of the variables are used: $Y = \ln(H)$ and $X = \ln(D)$. Here are the results for 5 trees:

X	-1.61	-1.20	-0.97	-0.51	-0.42
Y	2.22	2.27	2.38	2.60	2.65

The following results are given:

$$\sum x_i = -4.71,\ \sum y_i = 12.12,\ \sum x_i^2 = 5.4095,$$

$$\sum y_i^2 = 29.5282,\ \sum x_i y_i = -11.0458.$$

1. Test the pertinence of the regression at threshold 5%.
2. A tree with trunk diameter 0.7 m was 20 m in height. Was it unusually tall?

79. Suppose a family contains two children of different ages, and we are interested in the gender of these children. Let F denote that a child is female and M that the child is male and let a pair such as FM denote that the older child is female and the younger is male. There are four points in the set S of possible observations:

S = {FF, FM, MF, MM}.

Let A denote the subset of possibilities containing no males; B, the subset containing two males; and C, the subset containing at least one male. List the elements of $A, B, C, A \cap B,\ A \cup B, A \cap C, A \cup C, B \cap C, B \cup C,$ *and* $C \cap \bar{B}$

80. Draw Venn diagrams to verify DeMorgan's laws. That is, for any two sets A and B, $\overline{(A \cup B)} = \bar{A} \cap \bar{B}$ and $\overline{(A \cap B)} = \bar{A} \cup \bar{B}$.

81. Use the identities A = A ∩ S and S = B ∪ $\bar{B}$ and a distributive law to prove that

a. A = (A ∩ B) ∪ (A ∩ $\bar{B}$).

b. If B ⊂ A then A = B ∪ (A ∩ $\bar{B}$).

c. Further, show that (A ∩ B) and (A ∩ $\bar{B}$) are mutually exclusive and therefore that A is the union of two mutually exclusive sets, (A ∩ B) and (A ∩ $\bar{B}$).

d. Also show that B and (A ∩ $\bar{B}$) are mutually exclusive and if B ⊂ A, A is the union of two mutually exclusive sets, B and (A ∩ $\bar{B}$).

82. A group of five applicants for a pair of identical jobs consists of three men and two women. The employer is to select two of the five applicants for the jobs. Let S denote the set of all possible outcomes for the employer's selection. Let A denote the subset of outcomes corresponding to the selection of two men and B the subset corresponding to the selection of at least one woman. List the outcomes in $A, B, A \cup \bar{B}, A \cap B, and\ A \cap \bar{B}$. (Denote the different man and women by M_1, M_2, M_3 and W_1, W_2, respectively.)

83. A manufacturer has five seemingly identical computer terminals available for shipping. Unknown to her, two of the five are defective. A particular order calls for two of the terminals and is filled by randomly selecting two of the five that are available.

a. List the sample space for this experiment.

b. Let *A* denote the event that the order is filled with two nondefective terminals. List the sample points in *A*.

c. Construct a Venn diagram for the experiment that illustrates event *A*.

d. Assign probabilities to the simple events in such a way that the information about the experiment is used and the axioms in Definition 2.6 are met.

e. Find the probability of event *A*.

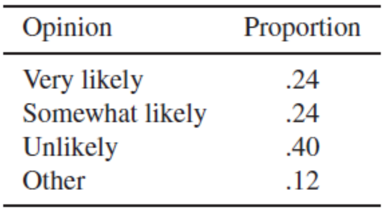
84. Every person's blood type is A, B, AB, or O. In addition, each individual either has the Rhesus (Rh) factor (+) or does not (−). A medical technician records a person's blood type and Rh factor. List the sample space for this experiment.

85. A sample space consists of five simple events, E1, E2, E3, E4, and E5.

a. If $P(E1) = P(E2) = 0.15, P(E3) = 0.4, \mathit{and}\ P(E4) = 2P(E5),$ find the probabilities of E4 and E5.

b. If $P(E1) = 3P(E2) = 0.3$, find the probabilities of the remaining simple events if you know that the remaining simple events are equally probable.

86. Americans can be quite suspicious, especially when it comes to government conspiracies. On the question of whether the U.S. Air Force has withheld proof of the existence of intelligent life on other planets, the proportions of Americans with varying opinions are given in the table.



Opinion	Proportion
Very likely	.24
Somewhat likely	.24
Unlikely	.40
Other	.12

Suppose that one American is selected and his or her opinion is recorded.

a. What are the simple events for this experiment?

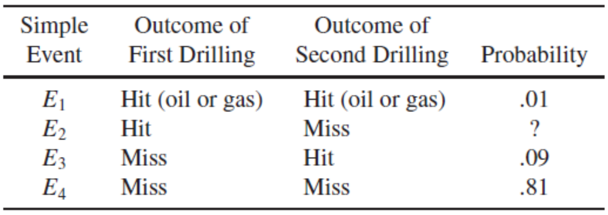
b. Are the simple events that you gave in part (a) all equally likely? If not, what are the probabilities that should be assigned to each?

c. What is the probability that the person selected finds it at least somewhat likely that the Air Force is withholding information about intelligent life on other planets?

87. An oil prospecting firm hits oil or gas on 10% of its drillings. If the firm drills two wells, the four possible simple events and three of their associated probabilities are as given in the accompanying table. Find the probability that the company will hit oil or gas

a. on the first drilling and miss on the second.

b. on at least one of the two drillings.



Simple Event	Outcome of First Drilling	Outcome of Second Drilling	Probability
E_1	Hit (oil or gas)	Hit (oil or gas)	.01
E_2	Hit	Miss	?
E_3	Miss	Hit	.09
E_4	Miss	Miss	.81

88. Hydraulic landing assemblies coming from an aircraft rework facility are each inspected for defects. Historical records indicate that 8% have defects in shafts only, 6% have defects in bushings only, and2%have defects in both shafts and bushings. One of the hydraulic assemblies is selected randomly. What is the probability that the assembly has

a. a bushing defect?

b. a shaft or bushing defect?

c. exactly one of the two types of defects?

d. neither type of defect?

89. A business office orders paper supplies from one of three vendors, V1, V2, or V3. Orders are to be placed on two successive days, one order per day. Thus, (V2, V3) might denote that vendor V2 gets the order on the first day and vendor V3 gets the order on the second day.

a. List the sample points in this experiment of ordering paper on two successive days.

b. Assume the vendors are selected at random each day and assign a probability to each sample point.

c. Let A denote the event that the same vendor gets both orders and B the event that V2 gets at least one order. Find $P(A), P(B), P(A \cup B),$ and $P(A \cap B)$ by summing the probabilities of the sample points in these events.

90. Consider the problem of selecting two applicants for a job out of a group of five and imagine that the applicants vary in competence, 1 being the best, 2 second best, and so on, for 3, 4, and 5. These ratings are of course unknown to the employer. Define two events A and B as:

A: The employer selects the best and one of the two poorest applicants (applicants 1 and 4 or 1 and 5).

B: The employer selects at least one of the two best.

Find the probabilities of these events.

91. A balanced coin is tossed three times. Calculate the probability that exactly two of the three tosses result in heads.

92. The odds are two to one that, when A and B play tennis, A wins. Suppose that A and B play two matches. What is the probability that A wins at least one match?

93. A single car is randomly selected from among all of those registered at a local tag agency. What do you think of the following claim? "All cars are either Volkswagens or they are not. Therefore, the probability is 1/2 that the car selected is a Volkswagen."

94. Two additional jurors are needed to complete a jury for a criminal trial. There are six prospective jurors, two women and four men. Two jurors are randomly selected from the six available.

a. Define the experiment and describe one sample point. Assume that you need describe only the two jurors chosen and not the order in which they were selected.

b. List the sample space associated with this experiment.

c. What is the probability that both of the jurors selected are women?

95. The Bureau of the Census reports that the median family income for all families in the United States during the year 2003 was $43,318. That is, half of all American families had incomes exceeding this amount, and half had incomes equal to or below this amount. Suppose that four families are surveyed and that each one reveals whether its income exceeded $43,318 in 2003.

a. List the points in the sample space.

b. Identify the simple events in each of the following events:

A: At least two had incomes exceeding $43,318.

B: Exactly two had incomes exceeding $43,318.

C: Exactly one had income less than or equal to $43,318.

c. Make use of the given interpretation for the median to assign probabilities to the simple events and find P(A), P(B), and P(C).

96. A boxcar contains six complex electronic systems. Two of the six are to be randomly selected for thorough testing and then classified as defective or not defective.

a. If two of the six systems are actually defective, find the probability that at least one of the two systems tested will be defective. Find the probability that both are defective.

b. If four of the six systems are actually defective, find the probabilities indicated in part (a).

97. The names of 3 employees are to be randomly drawn, without replacement, from a bowl containing the names of 30 employees of a small company. The person whose name is drawn first receives $100, and the individuals whose names are drawn second and third receive $50 and $25, respectively. How many sample points are associated with this experiment?

98. Suppose that an assembly operation in a manufacturing plant involves four steps, which can be performed in any sequence. If the manufacturer wishes to compare the assembly time for each of the sequences, how many different sequences will be involved in the experiment?

99. A labor dispute has arisen concerning the distribution of 20 laborers to four different construction jobs. The first job (considered to be very undesirable) required 6 laborers; the second, third, and fourth utilized 4, 5, and 5 laborers, respectively. The dispute arose over an alleged random distribution of the laborers to the jobs that placed all 4 members of a particular ethnic group on job 1. In considering whether the assignment represented injustice, a mediation panel desired the probability of the observed event. Determine the number of sample points in the sample space S for this experiment. That is, determine the number of ways the 20 laborers can be divided into groups of the appropriate sizes to fill all of the jobs. Find the probability of the observed event if it is assumed that the laborers are randomly assigned to jobs.

100. An airline has six flights from New York to California and seven flights from California to Hawaii per day. If the flights are to be made on separate days, how many different flight arrangements can the airline offer from New York to Hawaii?

101. A businesswoman in Philadelphia is preparing an itinerary for a visit to six major cities. The distance traveled, and hence the cost of the trip, will depend on the order in which she plans her route.

a. How many different itineraries (and trip costs) are possible?

b. If the businesswoman randomly selects one of the possible itineraries and Denver and San Francisco are two of the cities that she plans to visit, what is the probability that she will visit Denver before San Francisco?

102. An experiment consists of tossing a pair of dice.

a. Use the combinatorial theorems to determine the number of sample points in the sample space S.

b. Find the probability that the sum of the numbers appearing on the dice is equal to 7.

103. How many different seven-digit telephone numbers can be formed if the first digit cannot be zero?

104. A fleet of nine taxis is to be dispatched to three airports in such a way that three go to airport A, five go to airport B, and one goes to airport C. In how many distinct ways can this be accomplished?

105. Students attending the University of Florida can select from 130 major areas of study. A student's major is identified in the registrar's records with a two-or three-letter code (for example, statistics majors are identified by STA, math majors by MS). Some students opt for a double major and complete the requirements for both of the major areas before graduation. The registrar was asked to consider assigning these double majors a distinct two- or three-letter code so that they could be identified through the student records' system.

a. What is the maximum number of possible double majors available to University of Florida students?

b. If any two- or three-letter code is available to identify majors or double majors, how many major codes are available?

c. How many major codes are required to identify students who have either a single major or a double major?

d. Are there enough major codes available to identify all single and double majors at the University of Florida?

106. A local fraternity is conducting a raffle where 50 tickets are to be sold—one per customer. There are three prizes to be awarded. If the four organizers of the raffle each buy one ticket, what is the probability that the four organizers win

a. all of the prizes?

b. exactly two of the prizes?

c. exactly one of the prizes?

d. none of the prizes?

107. Five firms, F1, F2, . . . , F5, each offer bids on three separate contracts, C1,C2, and C3. Any one firm will be awarded at most one contract. The contracts are quite different, so an assignment of C1 to F1, say, is to be distinguished from an assignment of C2 to F1.

a. How many sample points are there altogether in this experiment involving assignment of contracts to the firms? (No need to list them all.)

b. Under the assumption of equally likely sample points, find the probability that F3 is awarded a contract.

108. A study is to be conducted in a hospital to determine the attitudes of nurses toward various administrative procedures. A sample of 10 nurses is to be selected from a total of the 90 nurses employed by the hospital.

a. How many different samples of 10 nurses can be selected?

b. Twenty of the 90 nurses are male. If 10 nurses are randomly selected from those employed by the hospital, what is the probability that the sample of ten will include exactly 4 male (and 6 female) nurses?

109. Two cards are drawn from a standard 52-card playing deck. What is the probability that the draw will yield an ace and a face card?

110. Five cards are dealt from a standard 52-card deck. What is the probability that we draw

a. 1 ace, 1 two, 1 three, 1 four, and 1 five (this is one way to get a "straight")?

b. any straight?

111. Suppose that we ask *n* randomly selected people whether they share your birthday.

a. Give an expression for the probability that no one shares your birthday (ignore leap years).

b. How many people do we need to select so that the probability is at least .5 that at least one shares your birthday?

112.The eight-member Human Relations Advisory Board of Gainesville, Florida, considered the complaint of a woman who claimed discrimination, based on sex, on the part of a local company. The board, composed of fivewomen and three men, voted 5–3 in favor of the plaintiff, the five women voting in favor of the plaintiff, the three men against. The attorney representing the company appealed the board's decision by claiming sex bias on the part of the board members. If there was no sex bias among the board members, it might be reasonable to conjecture that any group of five board members would be as likely to vote for the complainant as any other group of five. If this were the case, what is the probability that the vote would split along sex lines (five women for, three men against)?

113. Consider the following events in the toss of a single die: A: Observe an odd number, B: Observe an even number, C: Observe a 1 or 2.

a. Are A and B independent events?
b. Are A and C independent events?

114. Three brands of coffee, X, Y , and Z, are to be ranked according to taste by a judge. Define the following events:

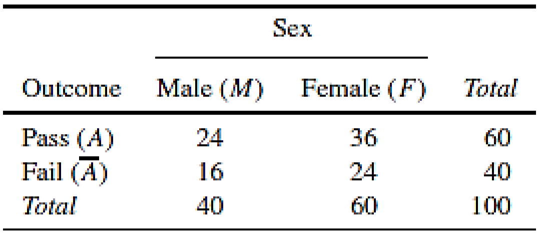
A: Brand X is preferred to Y .
B: Brand X is ranked best.
C: Brand X is ranked second best.
D: Brand X is ranked third best.

If the judge actually has no taste preference and randomly assigns ranks to the brands, is event A independent of events B, C, and D?

115. If two events, A and B, are such that $P(A) = .5$, $P(B) = .3$, and $P(A \cap B) = .1$, find the following:

a. P(A|B)
b. P(B|A)
c. P(A|A ∪ B)
d. P(A|A ∩ B)
e. P(A ∩ B|A ∪ B)

116. Gregor Mendel was a monk who, in 1865, suggested a theory of inheritance based on the science of genetics. He identified heterozygous individuals for flower color that had two alleles (one r = recessive white color allele and one R = dominant red color allele). When these individuals were mated, 3/4 of the offspring were observed to have red flowers, and 1/4 had white flowers. The following table summarizes this mating; each parent gives one of its alleles to form the gene of the offspring.



	Sex		
Outcome	Male (M)	Female (F)	*Total*
Pass (A)	24	36	60
Fail ($\bar{A}$)	16	24	40
Total	40	60	100

We assume that each parent is equally likely to give either of the two alleles and that, if either one or two of the alleles in a pair is dominant (R), the offspring will have red flowers. What is the probability that an offspring has

a. at least one dominant allele?

b. at least one recessive allele?

c. one recessive allele, given that the offspring has red flowers?

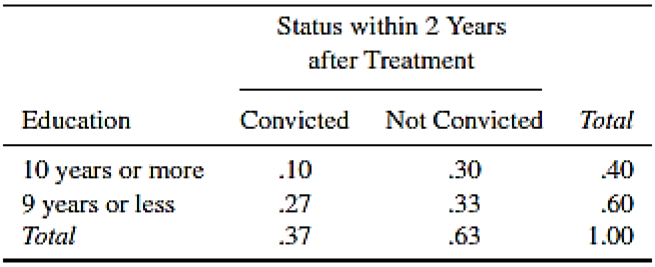
117.Cards are dealt, one at a time, from a standard 52-card deck.

a. If the first 2 cards are both spades, what is the probability that the next 3 cards are also spades?

b. If the first 3 cards are all spades, what is the probability that the next 2 cards are also spades?

c. If the first 4 cards are all spades, what is the probability that the next card is also a spade?

118.A study of the posttreatment behavior of a large number of drug abusers suggests that the likelihood of conviction within a two-year period after treatment may depend upon the offenders education. The proportions of the total number of cases falling in four education–conviction categories are shown in the following table:



	Status within 2 Years after Treatment		
Education	Convicted	Not Convicted	*Total*
10 years or more	.10	.30	.40
9 years or less	.27	.33	.60
Total	.37	.63	1.00

Suppose that a single offender is selected from the treatment program. Define the events: A: The offender has 10 or more years of education. B: The offender is convicted within two years after completion of treatment. Find the following:

a. P(A).

b. P(B).

c. P(A ∩ B).

d. P(A ∪ B).

e. $P(\bar{A})$.

f. P($\overline{A\ \cup\ B}$).

g. P($\overline{A\ \cap\ B}$).

h. P(A|B).

i. P(B|A).

119. Suppose that A and B are mutually exclusive events, with P(A) > 0 and P(B) < 1. Are A and B independent? Prove your answer.

120. If A and B are mutually exclusive events and P(B) > 0, show that

$$P(A|A \cup B) = \frac{P(A)}{P(A) + P(B)}.$$

121. If A and B are independent events, show that A and $\bar{B}$ are also independent. Are $\bar{A}$ and $\bar{B}$ independent?

122. Suppose that A and B are two events such that P(A) + P(B) > 1.

a. What is the smallest possible value for P(A ∩ B)?

b. What is the largest possible value for P(A ∩ B)?

123. Suppose that A and B are two events such that P(A) + P(B) < 1.

a. What is the smallest possible value for P(A ∩ B)?

b. What is the largest possible value for P(A ∩ B)?

124. Can A and B be mutually exclusive if P(A) = .4 and P(B) = .7? If P(A) = .4 and P(B) = .3? Why?

125. Two events A and B are such that P(A) = .2, P(B) = .3, and P(A ∪ B) = .4. Find the following:

a. P(A ∩ B)

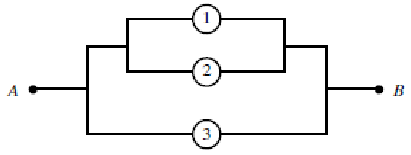
b. P($\bar{A}$ ∪ $\bar{B}$)

c. P($\bar{A}$ ∩ $\bar{B}$)

d. P($\bar{A}$ |B)

126. In a game, a participant is given three attempts to hit a ball. On each try, she either scores a hit, H, or a miss, M. The game requires that the player must alternate which hand she uses in successive attempts. That is, if she makes her first attempt with her right hand, she must use her left hand for the second attempt and her right hand for the third. Her chance of scoring a hit with her right hand is .7 and with her left hand is .4. Assume that the results of successive attempts are independent and that she wins the game if she scores at least two hits in a row. If she makes her first attempt with her right hand, what is the probability that she wins the game?

127. Consider the following portion of an electric circuit with three relays. Current will flow from point a to point b if there is at least one closed path when the relays are activated. The relays may malfunction and not close when activated. Suppose that the relays act independently of one another and close properly when activated, with a probability of .9.



a. What is the probability that current will flow when the relays are activated?

b. Given that current flowed when the relays were activated, what is the probability that relay 1 functioned?

128. Articles coming through an inspection line are visually inspected by two successive inspectors. When a defective article comes through the inspection line, the probability that it gets by the first inspector is 0.1. The second inspector will "miss" five out of ten of the defective items that get past the first inspector. What is the probability that a defective item gets by both inspectors?

129. It is known that a patient with a disease will respond to treatment with probability equal to 0.9. If three patients with the disease are treated and respond independently, find the probability that at least one will respond.

130. A monkey is to demonstrate that she recognizes colors by tossing one red, one black, and one white ball into boxes of the same respective colors, one ball to a box. If the monkey has not learned the colors and merely tosses one ball into each box at random, find the probabilities of the following results:

a. There are no color matches.

b. There is exactly one color match.

131. An advertising agency notices that approximately 1 in 50 potential buyers of a product sees a given magazine ad, and 1 in 5 sees a corresponding ad on television. One in 100 sees both. One in 3 actually purchases the product after seeing the ad, 1 in 10 without seeing it. What is the probability that a randomly selected potential customer will purchase the product?

132. A state auto-inspection station has two inspection teams. Team 1 is lenient and passes all automobiles of a recent vintage; team 2 rejects all autos on a first inspection because their “headlights are not properly adjusted.” Four unsuspecting drivers take their autos to the station for inspection on four different days and randomly select one of the two teams.

a. If all four cars are new and in excellent condition, what is the probability that three of the four will be rejected?

b. What is the probability that all four will pass?

133. A football team has a probability of .75 of winning when playing any of the other four teams in its conference. If the games are independent, what is the probability the team wins all its conference games?

134. Suppose that two balanced dice are tossed repeatedly and the sum of the two uppermost faces is determined on each toss. What is the probability that we obtain

a. a sum of 3 before we obtain a sum of 7?

b. a sum of 4 before we obtain a sum of 7?

135. A new secretary has been given n computer passwords, only one of which will permit access to a computer file. Because the secretary has no idea which password is correct, he chooses one of the passwords at random and tries it. If the password is incorrect, he discards it and randomly selects another password from among those remaining, proceeding in this manner until he finds the correct password.

a. What is the probability that he obtains the correct password on the first try?

b. What is the probability that he obtains the correct password on the second try? The third try?

c. A security system has been set up so that if three incorrect passwords are tried before the correct one, the computer file is locked and access to it denied. If n = 7, what is the probability that the secretary will gain access to the file?

136. An electronic fuse is produced by five production lines in a manufacturing operation. The fuses are costly, are quite reliable, and are shipped to suppliers in 100-unit lots. Because testing is destructive, most buyers of the fuses test only a small number of fuses before deciding to accept or reject lots of incoming fuses. All five production lines produce fuses at the same rate and normally produce only 2% defective fuses, which are dispersed randomly in the output. Unfortunately, production line 1 suffered mechanical difficulty and produced 5% defectives during the month of March. This situation became known to the manufacturer after the fuses had been shipped. A customer received a lot produced in March and tested three fuses. One failed.

a. What is the probability that the lot was produced on line 1?

b. What is the probability that the lot came from one of the four other lines?

137. A diagnostic test for a disease is such that it (correctly) detects the disease in 90% of the individuals who actually have the disease. Also, if a person does not have the disease, the test will report that he or she does not have it with probability .9. Only 1% of the population has the disease in question. If a person is chosen at random from the population and the diagnostic test indicates that she has the disease, what is the conditional probability that she does, in fact, have the disease? Are you surprised by the answer? Would you call this diagnostic test reliable?

138. Males and females are observed to react differently to a given set of circumstances. It has been observed that 70% of the females react positively to these circumstances, whereas only 40% of males react positively. A group of 20 people, 15 female and 5 male, was subjected to these circumstances, and the subjects were asked to describe their reactions on a written questionnaire. A response picked at random from the 20 was negative. What is the probability that it was that of a male?

139. A student answers a multiple-choice examination question that offers four possible answers. Suppose the probability that the student knows the answer to the question is .8 and the probability that the student will guess is .2. Assume that if the student guesses, the probability of selecting the correct answer is .25. If the student correctly answers a question, what is the probability that the student really knew the correct answer?

140. Of the travelers arriving at a small airport, 60% fly on major airlines, 30% fly on privately owned planes, and the remainder fly on commercially owned planes not belonging to a major airline. Of those traveling on major airlines, 50% are traveling for business reasons, whereas 60% of those arriving on private planes and 90% of those arriving on other commercially owned planes are traveling for business reasons. Suppose that we randomly select one person arriving at this airport. What is the probability that the person

a. is traveling on business?

b. is traveling for business on a privately owned plane?

c. arrived on a privately owned plane, given that the person is traveling for business reasons?

d. is traveling on business, given that the person is flying on a commercially owned plane?

141. Five identical bowls are labeled 1, 2, 3, 4, and 5. Bowl i contains i white and 5 − i black balls, with i = 1, 2, . . . , 5. A bowl is randomly selected and two balls are randomly selected (without replacement) from the contents of the bowl.

a. What is the probability that both balls selected are white?

b. Given that both balls selected are white, what is the probability that bowl 3 was selected?

142. Define an experiment as tossing two coins and observing the results. Let Y equal the number of heads obtained.

a. Identify the sample points in S, assign a value of Y to each sample point, and identify the sample points associated with each value of the random variable Y .

b. Compute the probabilities for each value of Y.

DISCRETE RANDOM VARIABLES AND THEIR PROBABILITY DISTRUBITION

143. A supervisor in a manufacturing plant has three men and three women working for him. He wants to choose two workers for a special job. Not wishing to show any biases in his selection, he decides to select the two workers at random. Let Y denote the number of women in his selection. Find the probability distribution for Y.

144. When the health department tested private wells in a county for two impurities commonly found in drinking water, it found that 20% of the wells had neither impurity, 40% had impurity A, and 50% had impurity B. (Obviously, some had both impurities.) If a well is randomly chosen from those in the county, find the probability distribution for Y , the number of impurities found in the well.

145. A group of four components is known to contain two defectives. An inspector tests the components one at a time until the two defectives are located. Once she locates the two defectives, she stops testing, but the second defective is tested to ensure accuracy. Let Y denote the number of the test on which the second defective is found. Find the probability distribution for Y.

146. A problem in a test given to small children asks them to match each of three pictures of animals to the word identifying that animal. If a child assigns the three words at random to the three pictures, find the probability distribution for Y , the number of correct matches.

147.Each of three balls are randomly placed into one of three bowls. Find the probability distribution for Y = the number of empty bowls.

148. In order to verify the accuracy of their financial accounts, companies use auditors on a regular basis to verify accounting entries. The company's employees make erroneous entries 5% of the time. Suppose that an auditor randomly checks three entries.

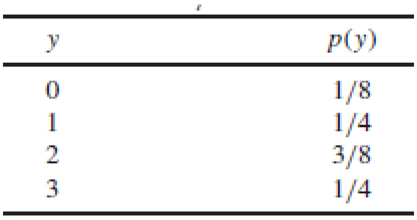
a. Find the probability distribution for Y , the number of errors detected by the auditor.

b. Construct a probability histogram for p(y).

c. Find the probability that the auditor will detect more than one error.

149. Persons entering a blood bank are such that 1 in 3 have type O+ blood and 1 in 15 have type O− blood. Consider three randomly selected donors for the blood bank. Let X denote the number of donors with type O+ blood and Y denote the number with type O− blood. Find the probability distributions for X and Y . Also find the probability distribution for X + Y , the number of donors who have type O blood.

150. The probability distribution for a random variable Y is given in table. Find the mean, variance, and standard deviation of Y .



y	$p(y)$
0	1/8
1	1/4
2	3/8
3	1/4

151. The manager of an industrial plant is planning to buy a new machine of either type A or type B. If t denotes the number of hours of daily operation, the number of daily repairs Y1 required to maintain a machine of type A is a random variable with mean and variance both equal to .10t. The number of daily repairs Y2 for a machine of type B is a random variable with mean and variance both equal to .12t. The daily cost of operating A is CA(t) = 10t + $30Y_1^2$; for B it is CB(t) = 8t + $30Y_2^2$. Assume that the repairs take negligible time and that each night the machines are tuned so that they operate essentially like new machines at the start of the next day. Which machine minimizes the expected daily cost if a workday consists of (a) 10 hours and (b) 20 hours?

152. An insurance company issues a one-year $1000 policy insuring against an occurrence A that historically happens to 2 out of every 100 owners of the policy. Administrative fees are $15 per policy and are not part of the company’s “profit.” How much should the company charge for the policy if it requires that the expected profit per policy be $50? [Hint: If C is the premium for the policy, the company’s “profit” is C−15 if A does not occur and C−15−1000 if A does occur.]

153. Who is the king of late night TV? An Internet survey estimates that, when given a choice between David Letterman and Jay Leno, 52% of the population prefers to watch Jay Leno. Three late night TV watchers are randomly selected and asked which of the two talk show hosts they prefer.

a. Find the probability distribution for Y, the number of viewers in the sample who prefer Leno.

b. Construct a probability histogram for p(y).

c. What is the probability that exactly one of the three viewers prefers Leno?

d. What are the mean and standard deviation for Y?

e. What is the probability that the number of viewers favoring Leno falls within 2 standard deviations of the mean?

154. In a gambling game a person draws a single card from an ordinary 52-card playing deck. A person is paid $15 for drawing a jack or a queen and $5 for drawing a king or an ace. A person who draws any other card pays $4. If a person plays this game, what is the expected gain?

155. Two construction contracts are to be randomly assigned to one or more of three firms: I, II, and III. Any firm may receive both contracts. If each contract will yield a profit of $90,000 for the firm, find the expected profit for firm I. If firms I and II are actually owned by the same individual, what is the owner's expected total profit?

156. A potential customer for an $85,000 fire insurance policy possesses a home in an area that, according to experience, may sustain a total loss in a given year with probability of .001 and a 50% loss with probability .01. Ignoring all other partial losses, what premium should the insurance company charge for a yearly policy in order to break even on all $85,000 policies in this area?

157. Suppose that Y is a discrete random variable with mean μ and variance σ^2 and let W = 2Y.

a. Do you expect the mean of W to be larger than, smaller than, or equal to μ=E(Y)? Why?

b. Express E(W) = E(2Y) in terms of μ = E(Y). Does this result agree with your answer to part (a)?

c. Recalling that the variance is a measure of spread or dispersion, do you expect the variance of W to be larger than, smaller than, or equal to σ^2 = V(Y)? Why?

d. Use the result in part (b) to show that $V(W) = E\{[W - E(W)]^2\} = E[4(Y - \mu)^2] = 4\sigma^2$; that is, W = 2Y has variance four times that of Y .

158. Let Y be a discrete random variable with mean μ and variance σ^2. If a and b are constants, prove that,

a. $E(aY + b) = aE(Y) + b = a\mu + b$.

b. $V(aY + b) = a^2V(Y) = a^2\sigma^2$.

159. Consider the population of voters described in Example 3.6. Suppose that there are N = 5000 voters in the population, 40% of whom favor Jones. Identify the event favors Jones as a success S. It is evident that the probability of S on trial 1 is .40. Consider the event B that S occurs on the second trial. Then B can occur two ways: The first two trials are both successes or the first trial is a failure and the second is a success. Show that P(B) = .4. What is P(B| the first trial is S)? Does this conditional probability differ markedly from P(B)?

160. In 2003, the average combined SAT score (math and verbal) for college-bound students in the United States was 1026. Suppose that approximately 45% of all high school graduates took this test and that 100 high school graduates are randomly selected from among all high school grads in the United States. Which of the following random variables has a distribution that can be approximated by a binomial distribution? Whenever possible, give the values for n and p.

a. The number of students who took the SAT

b. The scores of the 100 students in the sample

c. The number of students in the sample who scored above average on the SAT

d. The amount of time required by each student to complete the SAT

e. The number of female high school grads in the sample

161. A complex electronic system is built with a certain number of backup components in its subsystems. One subsystem has four identical components, each with a probability of .2 of failing in less than 1000 hours. The subsystem will operate if any two of the four components are operating. Assume that the components operate independently. Find the probability that

a. exactly two of the four components last longer than 1000 hours.

b. the subsystem operates longer than 1000 hours.

162. A multiple-choice examination has 15 questions, each with five possible answers, only one of which is correct. Suppose that one of the students who takes the examination answers each of the questions with an independent random guess. What is the probability that he answers at least ten questions correctly?

163. Many utility companies promote energy conservation by offering discount rates to consumers who keep their energy usage below certain established subsidy standards. A recent EPA report notes that 70% of the island residents of Puerto Rico have reduced their electricity usage sufficiently to qualify for discounted rates. If five residential subscribers are randomly selected from San Juan, Puerto Rico, find the probability of each of the following events:

a. All five qualify for the favorable rates.

b. At least four qualify for the favorable rates.

164. A fire-detection device utilizes three temperature-sensitive cells acting independently of each other in such a manner that any one or more may activate the alarm. Each cell possesses a probability of p = .8 of activating the alarm when the temperature reaches 100◦ Celsius or more. Let Y equal the number of cells activating the alarm when the temperature reaches 100◦.

a. Find the probability distribution for Y .

b. Find the probability that the alarm will function when the temperature reaches 100◦.

165. Tay-Sachs disease is a genetic disorder that is usually fatal in young children. If both parents are carriers of the disease, the probability that their offspring will develop the disease is approximately .25. Suppose that a husband and wife are both carriers and that they have three children. If the outcomes of the three pregnancies are mutually independent, what are the probabilities of the following events?

a. All three children develop Tay-Sachs.

b. Only one child develops Tay-Sachs.

c. The third child develops Tay-Sachs, given that the first two did not.

166. In the 18th century, the Chevalier de Mere asked Blaise Pascal to compare the probabilities of two events. Below, you will compute the probability of the two events that, prior to contrary gambling experience, were thought by de Mere to be equally likely.

a. What is the probability of obtaining at least one 6 in four rolls of a fair die?

b. If a pair of fair dice is tossed 24 times, what is the probability of at least one double six?

167. A manufacturer of floor wax has developed two new brands, A and B, which she wishes to subject to homeowners' evaluation to determine which of the two is superior. Both waxes, A and B, are applied to floor surfaces in each of 15 homes. Assume that there is actually no difference in the quality of the brands. What is the probability that ten or more homeowners would state a preference for

a. brand A?

b. either brand A or brand B?

168. Suppose that Y is a binomial random variable with n > 2 trials and success probability p. Use the technique presented in Theorem 3.7 and the fact that

$E\{Y\ (Y-1)(Y-2)\} = E(Y^3) - 3E(Y^{\ 2}) + 2E(Y\ )$ to derive $E(Y^{\ 3})$.

169. Ten motors are packaged for sale in a certain warehouse. The motors sell for \$100 each, but a double-your-money-back guarantee is in effect for any defectives the purchaser may receive. Find the expected net gain for the seller if the probability of any one motor being defective is .08. (Assume that the quality of any one motor is independent of that of the others.)

170. Of the volunteers donating blood in a clinic, 80% have the Rhesus (Rh) factor present in their blood.

a. If five volunteers are randomly selected, what is the probability that at least one does not have the Rh factor?

b. If five volunteers are randomly selected, what is the probability that at most four have the Rh factor?

c. What is the smallest number of volunteers who must be selected if we want to be at least 90% certain that we obtain at least five donors with the Rh factor?

171. Suppose that the probability of engine malfunction during any one-hour period is p = .02. Find the probability that a given engine will survive two hours.

172. If the probability of engine malfunction during any one-hour period is p = .02 and Y denotes the number of one-hour intervals until the first malfunction, find the mean and standard deviation of Y .

173. Suppose that 30% of the applicants for a certain industrial job possess advanced training in computer programming. Applicants are interviewed sequentially and are selected at random from the pool. Find the probability that the first applicant with advanced training in programming is found on the fifth interview.

174.About six months into GeorgeW. Bush's second term as president, a Gallup poll indicated that a near record (low) level of 41% of adults expressed "a great deal" or "quite a lot" of confidence in the U.S. Supreme Court Suppose that you conducted your own telephone survey at that time and randomly called people and asked them to describe their level of confidence in the Supreme Court. Find the probability distribution for Y , the number of calls until the first person is found who does not express "a great deal" or "quite a lot" of confidence in the U.S. Supreme Court.

175. Let Y denote a geometric random variable with probability of success p.

a. Show that for a positive integer a, $P(Y > a) = q^a$.

b. Show that for positive integers a and b, $P(Y > a + b|Y > a) = q^b = P(Y > b)$.

This result implies that, for example, $P(Y > 7|Y > 2) = P(Y > 5)$. Why do you think this property is called the memoryless property of the geometric distribution?

c. In the development of the distribution of the geometric random variable, we assumed that the experiment consisted of conducting identical and independent trials until the first success was observed. In light of these assumptions, why is the result in part (b) "obvious"?

176. A certified public accountant (CPA) has found that nine of ten company audits contain substantial errors. If the CPA audits a series of company accounts, what is the probability that the first account containing substantial errors

a. is the third one to be audited?

b. will occur on or after the third audited account?

177. The probability of a customer arrival at a grocery service counter in any one second is equal to .1. Assume that customers arrive in a random stream and hence that an arrival in any one second is independent of all others. Find the probability that the first arrival

a. will occur during the third one-second interval.

b. will not occur until at least the third one-second interval.

178. How many times would you expect to toss a balanced coin in order to obtain the first head?

179. In responding to a survey question on a sensitive topic (such as "Have you ever tried marijuana?"), many people prefer not to respond in the affirmative. Suppose that 80% of the population have not tried marijuana and all of those individuals will truthfully answer no to your question. The remaining 20% of the population have tried marijuana and 70% of those individuals will lie. Derive the probability distribution of Y , the number of people you would need to question in order to obtain a single affirmative response.

108. A geological study indicates that an exploratory oil well drilled in a particular region should strike oil with probability .2. Find the probability that the third oil strike comes on the fifth well drilled.

181. A large stockpile of used pumps contains 20% that are in need of repair. A maintenance worker is sent to the stockpile with three repair kits. She selects pumps at random and tests them one at a time. If the pump works, she sets it aside for future use. However, if the pump does not work, she uses one of her repair kits on it. Suppose that it takes 10 minutes to test a pump that is in working condition and 30 minutes to test and repair a pump that does not work. Find the mean and variance of the total time it takes the maintenance worker to use her three repair kits.

182. A geological study indicates that an exploratory oil well should strike oil with probability .2.

a. What is the probability that the first strike comes on the third well drilled?

b. What is the probability that the third strike comes on the seventh well drilled?

c. What assumptions did you make to obtain the answers to parts (a) and (b)?

d. Find the mean and variance of the number of wells that must be drilled if the company wants to set up three producing wells.

183. The employees of a firm that manufactures insulation are being tested for indications of asbestos in their lungs. The firm is requested to send three employees who have positive indications of asbestos on to a medical center for further testing. If each test costs $20, find the expected value and variance of the total cost of conducting the tests necessary to locate the three positives.

184. Ten percent of the engines manufactured on an assembly line are defective. What is the probability that the third nondefective engine will be found

a. on the fifth trial?

b. on or before the fifth trial?

c. Given that the first two engines tested were defective, what is the probability that at least two more engines must be tested before the first nondefective is found?

d. Find the mean and variance of the number of the trial on which a the first nondefective engine is found.

e. Find the mean and variance of the number of the trial on which the third nondefective engine is found.

185. In a sequence of independent identical trials with two possible outcomes on each trial, S and F, and with P(S) = p, what is the probability that exactly y trials will occur before the r th success?

186. An important problem encountered by personnel directors and others faced with the selection of the best in a finite set of elements is exemplified by the following scenario. From a group of 20 Ph.D. engineers, 10 are randomly selected for employment. What is the probability that the 10 selected include all the 5 best engineers in the group of 20?

187. An industrial product is shipped in lots of 20. Testing to determine whether an item is defective is costly, and hence the manufacturer samples his production rather than using a 100% inspection plan. A sampling plan, constructed to minimize the number of defectives shipped to customers, calls for sampling five items from each lot and rejecting the lot if more than one defective is observed. (If the lot is rejected, each item in it is later tested.) If a lot contains four defectives, what is the probability that it will be rejected? What is the expected number of defectives in the sample of size 5? What is the variance of the number of defectives in the sample of size 5?

188. A warehouse contains ten printing machines, four of which are defective. A company selects five of the machines at random, thinking all are in working condition. What is the probability that all five of the machines are nondefective?

189. In southern California, a growing number of individuals pursuing teaching credentials are choosing paid internships over traditional student teaching programs. A group of eight candidates for three local teaching positions consisted of five who had enrolled in paid internships and three who enrolled in traditional student teaching programs. All eight candidates appear to be equally qualified, so three are randomly selected to fill the open positions. Let Y be the number of internship trained candidates who are hired.

a. Does Y have a binomial or hypergeometric distribution? Why?

b. Find the probability that two or more internship trained candidates are hired.

c. What are the mean and standard deviation of Y?

190. Seed are often treated with fungicides to protect them in poor draining, wet environments. A small-scale trial, involving five treated and five untreated seeds, was conducted prior to a large-scale experiment to explore how much fungicide to apply. The seeds were planted in wet soil, and the number of emerging plants were counted. If the solution was not effective and four plants actually sprouted, what is the probability that

a. all four plants emerged from treated seeds?

b. three or fewer emerged from treated seeds?

c. at least one emerged from untreated seeds?

191. A group of six software packages available to solve a linear programming problem has been ranked from 1 to 6 (best to worst). An engineering firm, unaware of the rankings, randomly selected and then purchased two of the packages. Let Y denote the number of packages purchased by the firm that are ranked 3, 4, 5, or 6. Give the probability distribution for Y.

192. Specifications call for a thermistor to test out at between 9000 and 10,000 ohms at 25° Celcius. Ten thermistors are available, and three of these are to be selected for use. Let Y denote the number among the three that do not conform to specifications. Find the probability distributions for Y (in tabular form) under the following conditions:

a. Two thermistors do not conform to specifications among the ten that are available.

b. Four thermistors do not conform to specifications among the ten that are available.

193. A jury of 6 persons was selected from a group of 20 potential jurors, of whom 8 were African American and 12 were white. The jury was supposedly randomly selected, but it contained only 1 African American member. Do you have any reason to doubt the randomness of the selection?

194. Suppose that a radio contains six transistors, two of which are defective. Three transistors are selected at random, removed from the radio, and inspected. Let Y equal the number of defectives observed, where Y = 0, 1, or 2. Find the probability distribution for Y.

195. In an assembly-line production of industrial robots, gearbox assemblies can be installed in one minute each if holes have been properly drilled in the boxes and in ten minutes if the holes must be redrilled. Twenty gearboxes are in stock, 2 with improperly drilled holes. Five gearboxes must be selected from the 20 that are available for installation in the next five robots.

a. Find the probability that all 5 gearboxes will fit properly.

b. Find the mean, variance, and standard deviation of the time it takes to install these 5 gearboxes.

196. Cards are dealt at random and without replacement from a standard 52 card deck. What is the probability that the second king is dealt on the fifth card?

197. Show that the probabilities assigned by the Poisson probability distribution satisfy the requirements that $0 \leq p(y) \leq 1$ for all y and $\sum p(y) = 1$

198. Suppose that a random system of police patrol is devised so that a patrol officer may visit a given beat location Y = 0, 1, 2, 3, . . . times per half-hour period, with each location being visited an average of once per time period. Assume that Y possesses, approximately, a Poisson probability distribution. Calculate the probability that the patrol officer will miss a given location during a half-hour period. What is the probability that it will be visited once? Twice? At least once?

199. A certain type of tree has seedlings randomly dispersed in a large area, with the mean density of seedlings being approximately five per square yard. If a forester randomly locates ten 1-square-yard sampling regions in the area, find the probability that none of the regions will contain seedlings.

200. Industrial accidents occur according to a Poisson process with an average of three accidents per month. During the last two months, ten accidents occurred. Does this number seem highly improbable if the mean number of accidents per month, μ, is still equal to 3? Does it indicate an increase in the mean number of accidents per month?

201. Let Y denote a random variable that has a Poisson distribution with mean λ = 2. Find

a. P(Y = 4).

b. P(Y ≥ 4).

c. P(Y < 4).

d. P(Y ≥ 4|Y ≥ 2).

202. The random variable Y has a Poisson distribution and is such that p(0) = p(1). What is p(2)?

203. Customers arrive at a checkout counter in a department store according to a Poisson distribution at an average of seven per hour. If it takes approximately ten minutes to serve each customer, find the mean and variance of the total service time for customers arriving during a 1-hour period. (Assume that a sufficient number of servers are available so that no customer must wait for service.) Is it likely that the total service time will exceed 2.5 hours?

204. The number of typing errors made by a typist has a Poisson distribution with an average of four errors per page. If more than four errors appear on a given page, the typist must retype the whole page. What is the probability that a randomly selected page does not need to be retyped?

205. The number of knots in a particular type of wood has a Poisson distribution with an average of 1.5 knots in 10 cubic feet of the wood. Find the probability that a 10-cubic-foot block of the wood has at most 1 knot.

206. Assume that the tunnel in Exercise 3.132 is observed during ten two-minute intervals, thus giving ten independent observations $Y1, Y2, \ldots, Y10$, on the Poisson random variable. Find the probability that Y > 3 during at least one of the ten two-minute intervals.

207. A salesperson has found that the probability of a sale on a single contact is approximately .03. If the salesperson contacts 100 prospects, what is the approximate probability of making at least one sale?

208. The probability that a mouse inoculated with a serum will contract a certain disease is .2. Using the Poisson approximation, find the probability that at most 3 of 30 inoculated mice will contract the disease.

209. In the daily production of a certain kind of rope, the number of defects per foot Y is assumed to have a Poisson distribution with mean λ = 2. The profit per foot when the rope is sold is given by X, where X = 50 − 2Y − Y 2. Find the expected profit per foot.

210. A food manufacturer uses an extruder (a machine that produces bite-size cookies and snack food) that yields revenue for the firm at a rate of $200 per hour when in operation. However, the extruder breaks down an average of two times every day it operates. If Y denotes the number of breakdowns per day, the daily revenue generated by the machine is R = 1600−50Y 2. Find the expected daily revenue for the extruder.

211. The number of customers per day at a sales counter, Y , has been observed for a long period of time and found to have mean 20 and standard deviation 2. The probability distribution of Y is not known. What can be said about the probability that, tomorrow, Y will be greater than 16 but less than 24?

212. Let Y be a random variable with mean 11 and variance 9. Using Tchebysheff's theorem, find

a. a lower bound for $P(6 < Y < 16)$.

b. the value of C such that $P(|Y - 11| \geq C) \leq .09$.

213. This exercise demonstrates that, in general, the results provided by Tchebysheff's theorem cannot be improved upon. Let Y be a random variable such that

$$p(-1) = 1/18 \quad ,p(0) = 16/\ 18 \quad ,p(1) = 1/18$$

a. Show that $E(Y) = 0$ and $V(Y) = 1/9$.

b. Use the probability distribution of Y to calculate $P(|Y - \mu| \geq 3\sigma)$. Compare this exact probability with the upper bound provided by Tchebysheff's theorem to see that the bound provided by Tchebysheff's theorem is actually attained when $k = 3$.

214. For a certain type of soil the number of wireworms per cubic foot has a mean of 100. Assuming a Poisson distribution of wireworms, give an interval that will include at least 5/9 of the sample values of wireworm counts obtained from a large number of 1-cubic-foot samples.

215. A balanced coin is tossed three times. Let Y equal the number of heads observed.

a. Use the formula for the binomial probability distribution to calculate the probabilities associated with Y = 0, 1, 2, and 3.

b. Construct a probability distribution similar to the one in Table 3.1.

c. Find the expected value and standard deviation of Y , using the formulas E(Y) = np and V(Y) = npq.

d. Using the probability distribution from part (b), find the fraction of the population measurements lying within 1 standard deviation of the mean. Repeat for 2 standard deviations.

e. How do your results compare with the results of Tchebysheff's theorem and the empirical rule?

216. In May 2005, Tony Blair was elected to an historic third term as the British prime minister. A Gallop U.K. poll (http://gallup.com/poll/content/default.aspx?ci=1710, June 28, 2005) conducted after Blair's election indicated that only 32% of British adults would like to see their son or daughter grows up to become prime minister. If the same proportion of Americans would prefer that their son or daughter grow up to be president and 120 American adults are interviewed,

a. what is the expected number of Americans who would prefer their child grow up to be president?

b. what is the standard deviation of the number Y who would prefer that their child grow up to be president?

c. is it likely that the number of Americans who prefer that their child grow up to be president exceeds 40?

217. For a certain section of a pine forest, the number of diseased trees per acre, Y , has a Poisson distribution with mean λ = 10. The diseased trees are sprayed with an insecticide at a cost of $3 per tree, plus a fixed overhead cost for equipment rental of $50. Letting C denote the total spraying cost for a randomly selected acre, find the expected value and standard deviation for C. Within what interval would you expect C to lie with probability at least .75?

CONTINUOUS RANDOM VARIABLES AND THEIR PROBABILITY DISTRUBITION

218. Let Y be a continuous random variable with probability density function given by

$$f(y) = \begin{cases} 3y^2, & 0 \leq y \leq 1, \\ 0, & \text{elsewhere}. \end{cases}$$

Find F(y). Graph both f (y) and F(y).

219. Given $f\ (y)\ =\ cy^2, 0\ \leq\ y\ \leq\ 2$, and $f\ (y)\ =\ 0$ elsewhere, find the value of c for which $f\ (y)$ is a valid density function.

220. Let Y be a random variable with *p(y)* given in the table below.

a. Give the distribution function, F(y). Be sure to specify the value of F(y) for all y, $-\infty < y < \infty$.

b. Sketch the distribution function given in part (a).

221. Suppose that Y is a random variable that takes on only integer values 1, 2, . . . and has distribution function F(y). Show that the probability function p(y) = P(Y = y) is given by

$$p(y) = \begin{cases} F(1), & y = 1, \\ F(y) - F(y-1), & y = 2, 3, \dots . \end{cases}$$

222. A random variable Y has the following distribution function:

$$F(y) = P(Y \le y) = \begin{cases} 0, & \text{for } y < 2, \\ 1/8, & \text{for } 2 \le y < 2.5, \\ 3/16, & \text{for } 2.5 \le y < 4, \\ 1/2 & \text{for } 4 \le y < 5.5, \\ 5/8, & \text{for } 5.5 \le y < 6, \\ 11/16, & \text{for } 6 \le y < 7, \\ 1, & \text{for } y \ge 7. \end{cases}$$

a. Is Y a continuous or discrete random variable? Why?

b. What values of Y are assigned positive probabilities?

c. Find the probability function for Y.

d. What is the median, $\varphi_{.5}$, of Y ?

223. Suppose that Y possesses the density function

$$f(y) = \begin{cases} cy, & 0 \le y \le 2, \\ 0, & \text{elsewhere.} \end{cases}$$

a. Find the value of c that makes f (y) a probability density function.

b. Find F(y).

c. Graph f (y) and F(y).

d. Use F(y) to find P(1 ≤ Y ≤ 2).

e. Use f (y) and geometry to find P(1 ≤ Y ≤ 2).

224. A supplier of kerosene has a 150-gallon tank that is filled at the beginning of each week. His weekly demand shows a relative frequency behavior that increases steadily up to 100 gallons and then levels off between 100 and 150 gallons. If Y denotes weekly demand in hundreds of gallons, the relative frequency of demand can be modeled by

$$f(y) = \begin{cases} y, & 0 \le y \le 1, \\ 1, & 1 < y \le 1.5, \\ 0, & \text{elsewhere.} \end{cases}$$

a. Find F(y).

b. Find P(0 ≤ Y ≤ .5).

c. Find P(.5 ≤ Y ≤ 1.2).

225. Let the distribution function of a random variable Y be

$$F(y) = \begin{cases} 0, & y \le 0, \\ \dfrac{y}{8}, & 0 < y < 2, \\ \dfrac{y^2}{16}, & 2 \le y < 4, \\ 1, & y \ge 4. \end{cases}$$

a. Find the density function of Y .

b. Find P(1 ≤ Y ≤ 3).

c. Find P(Y ≥ 1.5).

d. Find P(Y ≥ 1|Y ≤ 3).

226. Suppose that Y has density function

$$f(y) = \begin{cases} (3/2)y^2 + y, & 0 \le y \le 1, \\ 0, & \text{elsewhere,} \end{cases}$$

find the mean and variance of Y **.**

227. If Y has distribution function,

$$F(y) = \begin{cases} 0, & y \leq 0, \\ \frac{y}{8}, & 0 < y < 2, \\ \frac{y^2}{16}, & 2 \leq y < 4, \\ 1, & y \geq 4. \end{cases}$$

find the mean and variance of Y .

228. For certain ore samples, the proportion Y of impurities per sample is a random variable with density function given by,

$$f(y) = \begin{cases} (3/2)y^2 + y, & 0 \leq y \leq 1, \\ 0, & \text{elsewhere,} \end{cases}$$

The dollar value of each sample is W = 5−0.5Y . Find the mean and variance of W.

229. The temperature Y at which a thermostatically controlled switch turns on has probability density function given by

$$f(y) = \begin{cases} 1/2, & 59 \leq y \leq 61, \\ 0, & \text{elsewhere.} \end{cases}$$

Find E(Y) and V(Y).

230. The pH of water samples from a specific lake is a random variable Y with probability density function given by

$$f(y) = \begin{cases} (3/8)(7 - y)^2, & 5 \leq y \leq 7, \\ 0, & \text{elsewhere.} \end{cases}$$

a. Find E(Y) and V(Y).

b. Find an interval shorter than (5, 7) in which at least three-fourths of the pH measurements must lie.

c. Would you expect to see a pH measurement below 5.5 very often? Why?

231. Daily total solar radiation for a specified location in Florida in October has probability density function given by

$$f(y) = \begin{cases} (3/8)(7-y)^2, & 5 \le y \le 7, \\ 0, & \text{elsewhere.} \end{cases}$$

with measurements in hundreds of calories. Find the expected daily solar radiation for October.

232. If Y is a continuous random variable such that $E[(Y - a)^2] < \infty$ for all a, show that $E[(Y - a)^2]$ is minimized when a = E(Y).

[Hint: $E[(Y - a)^2] = E(\{[Y - E(Y)] + [E(Y) - a]\}^2)$.]

233. Arrivals of customers at a checkout counter follow a Poisson distribution. It is known that, during a given 30-minute period, one customer arrived at the counter. Find the probability that the customer arrived during the last 5 minutes of the 30-minute period.

234. If a parachutist lands at a random point on a line between markers A and B, find the probability that she is closer to A than to B. Find the probability that her distance to A is more than three times her distance to B.

235. A random variable Y has a uniform distribution over the interval (θ1, θ2). Derive the variance of Y .

236. A circle of radius r has area $A = \pi r^{2}$. If a random circle has a radius that is uniformly distributed on the interval (0, 1), what are the mean and variance of the area of the circle?

237. Upon studying low bids for shipping contracts, a microcomputer manufacturing company finds that intrastate contracts have low bids that are uniformly distributed between 20 and 25, in units of thousands of dollars. Find the probability that the low bid on the next intrastate shipping contract

a. is below $22,000.

b. is in excess of $24,000.

238. The failure of a circuit board interrupts work that utilizes a computing system until a new board is delivered. The delivery time, Y , is uniformly distributed on the interval one to five days. The cost of a board failure and interruption includes the fixed cost c0 of a new board and a cost that increases proportionally to Y^2. If C is the cost incurred, $C = c_0 + c_1Y^2$.

a. Find the probability that the delivery time exceeds two days.

b. In terms of c_0 and c_1, find the expected cost associated with a single failed circuit board.

239. A telephone call arrived at a switchboard at random within a one-minute interval. The switch board was fully busy for 15 seconds into this one-minute period. What is the probability that the call arrived when the switchboard was not fully busy?

240. The cycle time for trucks hauling concrete to a highway construction site is uniformly distributed over the interval 50 to 70 minutes. What is the probability that the cycle time exceeds 65 minutes if it is known that the cycle time exceeds 55 minutes?

241. The number of defective circuit boards coming off a soldering machine follows a Poisson distribution. During a specific eight-hour day, one defective circuit board was found.

a. Find the probability that it was produced during the first hour of operation during that day.

b. Find the probability that it was produced during the last hour of operation during that day.

c. Given that no defective circuit boards were produced during the first four hours of operation, find the probability that the defective board was manufactured during the fifth hour.

242. Let Z denote a normal random variable with mean 0 and standard deviation 1.

a. Find $P(Z > 2)$.

b. Find $P(-2 \leq Z \leq 2)$.

c. Find $P(0 \leq Z \leq 1.73)$.

243. The achievement scores for a college entrance examination are normally distributed with mean 75 and standard deviation 10. What fraction of the scores lies between 80 and 90?

244. If Z is a standard normal random variable, find the value z0 such that

a. $P(Z > z_0) = .5$.

b. $P(Z < z_0) = .8643$.

c. $P(-z_0 < Z < z_0) = .90$.

d. $P(-z_0 < Z < z_0) = .99$.

245. What is the median of a normally distributed random variable with mean μ and standard deviation σ?

246. A company that manufactures and bottles apple juice uses a machine that automatically fills 16-ounce bottles. There is some variation, however, in the amounts of liquid dispensed into the bottles that are filled. The amount dispensed has been observed to be approximately normally distributed with mean 16 ounces and standard deviation 1 ounce. determine the proportion of bottles that will have more than 17 ounces dispensed into them.

247. The weekly amount of money spent on maintenance and repairs by a company was observed, over a long period of time, to be approximately normally distributed with mean $400 and standard deviation $20. How much should be budgeted for weekly repairs and maintenance to provide that the probability the budgeted amount will be exceeded in a given week is only .1?

248. A machining operation produces bearings with diameters that are normally distributed with mean 3.0005 inches and standard deviation .0010 inch. Specifications require the bearing diameters to lie in the interval 3.000±.0020 inches. Those outside the interval are considered scrap and must be remachined. What should the mean diameter be in order that the fraction of bearings scrapped be minimized?

249. The grade point averages (GPAs) of a large population of college students are approximately normally distributed with mean 2.4 and standard deviation .8. If students possessing a GPA less than 1.9 are dropped from college, what percentage of the students will be dropped?

250. Wires manufactured for use in a computer system are specified to have resistances between .12 and .14 ohms. The actual measured resistances of the wires produced by company A have a normal probability distribution with mean .13 ohm and standard deviation .005 ohm.

a. What is the probability that a randomly selected wire from company A's production will meet the specifications?

b. If four of these wires are used in each computer system and all are selected from company A, what is the probability that all four in a randomly selected system will meet the specifications?

251. The width of bolts of fabric is normally distributed with mean 950 mm (millimeters) and standard deviation 10 mm.

a. What is the probability that a randomly chosen bolt has a width of between 947 and 958mm?

b. What is the appropriate value for C such that a randomly chosen bolt has a width less than C with probability .8531?

252. A soft-drink machine can be regulated so that it discharges an average of μ ounces per cup. If the ounces of fill are normally distributed with standard deviation 0.3 ounce, give the setting for μ so that 8-ounce cups will overflow only 1% of the time.

253. The SAT and ACT college entrance exams are taken by thousands of students each year. The mathematics portions of each of these exams produce scores that are approximately normally distributed. In recent years, SAT mathematics exam scores have averaged 480 with standard deviation 100. The average and standard deviation for ACT mathematics scores are 18 and 6,respectively.

a. An engineering school sets 550 as the minimum SAT math score for new students. What percentage of students will score below 550 in a typical year?

b. What score should the engineering school set as a comparable standard on the ACT math test?

254. Show that the normal density with parameters μ and σ has inflection points at the values $\mu-\sigma$ and $\mu + \sigma$. (Recall that an inflection point is a point where the curve changes direction from concave up to concave down, or vice versa, and occurs when the second derivative change sign. Such a change in sign may occur when the second derivative equals zero.)

255. ***a***. If $\alpha > 0$, $\Gamma(\alpha)$ is defined by $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\, dy$ show that $\Gamma(1) = 1$.

b. If $\alpha > 1$, integrate by parts to prove that $\Gamma(\alpha) = (\alpha - 1)\, \Gamma(\alpha - 1)$.

256. The operator of a pumping station has observed that demand for water during early afternoon hours has an approximately exponential distribution with mean 100 cfs (cubic feet per second).

a. Find the probability that the demand will exceed 200 cfs during the early afternoon on a randomly selected day.

b. What water-pumping capacity should the station maintain during early afternoons so that the probability that demand will exceed capacity on a randomly selected day is only .01?

257. If Y has an exponential distribution and $P(Y > 2) = .0821$, what is

a. $\beta = E(Y)$?

b. $P(Y \leq 1.7)$?

258. Historical evidence indicates that times between fatal accidents on scheduled American domestic passenger flights have an approximately exponential distribution. Assume that the mean time between accidents is 44 days.

a. If one of the accidents occurred on July 1 of a randomly selected year in the study period, what is the probability that another accident occurred that same month?

b. What is the variance of the times between accidents?

259. A manufacturing plant uses a specific bulk product. The amount of product used in one day can be modeled by an exponential distribution with $\beta = 4$ (measurements in tons). Find the probability that the plant will use more than 4 tons on a given day.

260. The magnitude of earthquakes recorded in a region of North America can be modeled as having an exponential distribution with mean 2.4, as measured on the Richter scale. Suppose that the magnitude of earthquakes striking the region has a gamma distribution with $\alpha = .8$ and $\beta = 2.4$.

a. What is the mean magnitude of earthquakes striking the region?

b. What is the probability that the magnitude an earthquake striking the region will exceed 3.0 on the Richter scale?

c. Compare your answers to Exercise 4.88(a). Which probability is larger? Explain.

d. What is the probability that an earthquake striking the regions will fall between 2.0 and 3.0 on the Richter scale?

261. Explosive devices used in mining operations produce nearly circular craters when detonated. The radii of these craters are exponentially distributed with mean 10 feet. Find the mean and variance of the areas produced by these explosive devices.

262. Four-week summer rainfall totals in a section of the Midwest United States have approximately a gamma distribution with $\alpha = 1.6$ and $\beta = 2.0$.

a. Find the mean and variance of the four-week rainfall totals.

b. What is the probability that the four-week rainfall total exceeds 4 inches?

263. The response times on an online computer terminal have approximately a gamma distribution with mean four seconds and variance eight seconds^2.

a. Use Tchebysheff's theorem to give an interval that contains at least 75% of the response times.

b. What is the actual probability of observing a response time in the interval you obtained in part (a)?

264. The weekly amount of downtime Y (in hours) for an industrial machine has approximately a gamma distribution with $\alpha = 3$ and $\beta = 2$. The loss L (in dollars) to the industrial operation as a result of this downtime is given by $L = 30Y + 2Y\ 2$.

a. Find the expected value and variance of L.

265. Suppose that Y has a gamma distribution with parameters α and β.

a. If a is any positive or negative value such that $\alpha + a > 0$, show that $E(Y^a) = \beta^a(\alpha + a)/(\alpha)$

b. Why did your answer in part (a) require that $\alpha + a > 0$?

c. Show that, with a = 1, the result in part (a) gives $E(Y) = \alpha\beta$.

d. Use the result in part (a) to give an expression for $E(\sqrt{Y})$. What do you need to assume about α?

e. Use the result in part (a) to give an expression for $E(1/Y)$, $E(1/\sqrt{Y})$, and $E(1/Y^2)$. What do you need to assume about α in each case?

266. A gasoline wholesale distributor has bulk storage tanks that hold fixed supplies and are filled every Monday. Of interest to the wholesaler is the proportion of this supply that is sold during the week. Over many weeks of observation, the distributor found that this proportion could be modeled by a beta distribution with $\alpha = 4$ and $\beta = 2$. Find the probability that the wholesaler will sell at least 90% of her stock in a given week.

267. The relative humidity Y , when measured at a location, has a probability density function given following. Find the value of k that makes f (y) a density function.

$$f(y) = \begin{cases} ky^3(1-y)^2, & 0 \le y \le 1, \\ 0, & \text{elsewhere.} \end{cases}$$

268. The percentage of impurities per batch in a chemical product is a random variable Y with density function

$$f(y) = \begin{cases} 12y^2(1-y), & 0 \le y \le 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the mean and variance of the percentage of impurities in a randomly selected batch of the chemical.

269. Verify that if Y has a beta distribution with α = β = 1, then Y has a uniform distribution over (0, 1). That is, the uniform distribution over the interval (0, 1) is a special case of a beta distribution.

270. During an eight-hour shift, the proportion of time Y that a sheet-metal stamping machine is down for maintenance or repairs has a beta distribution with α = 1 and β = 2. That is,

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

a. The cost (in hundreds of dollars) of this downtime, due to lost production and cost of maintenance and repair, is given by $C = 10 + 20Y + 4Y^2$. Find the mean and variance of C.

b. Find an interval for which the probability that C will lie within it is at least .75.

271. Errors in measuring the time of arrival of a wave front from an acoustic source sometimes have an approximate beta distribution. Suppose that these errors, measured in microseconds, have approximately a beta distribution with $\alpha = 1$ and $\beta = 2$.

a. What is the probability that the measurement error in a randomly selected instance is less than .5 μs?

b. Give the mean and standard deviation of the measurement errors.

272. The proportion of time per day that all checkout counters in a supermarket are busy is a random variable Y with a density function given by

a. Find the value of c that makes f (y) a probability density function.

b. Find E(Y). (Use what you have learned about the beta-type distribution.)

c. Calculate the standard deviation of Y .

273. Find μ_k for the uniform random variable with $\theta_1 = 0$ and $\theta_2 = \theta$ for $k=1,2,3$.

274. A machine used to fill cereal boxes dispenses, on the average, μ ounces per box. The manufacturer wants the actual ounces dispensed Y to be within 1 ounce of μ at least 75% of the time. What is the largest value of σ, the standard deviation of Y , that can be tolerated if the manufacturer's objectives are to be met?

275. Find $P(|Y - \mu| \leq 2\sigma)$ for the uniform random variable. Compare with the corresponding probabilistic statements given by Tchebysheff's theorem and the empirical rule.

276. For a sample of 15 students we observe the Time (in minutes, X) usually spent using Facebook per day, and the final grade in the Statistics exam (Y):

Facebook X	Statistics Y
0	25
11	22
17	28
16	30
22	22
17	27
25	26
30	21
27	27
27	28
31	23
35	29
30	30
45	24
60	27

a) Compute mean, median and quartiles for both X and Y
b) Compute the range and the IQR for both X and Y
c) Find the standard deviation and the coefficient of variation for both X and Y. Which variable is more heterogeneous?
d) Plot the box-plot for both X and Y

277. Using the data of Exercise 1, Build a frequency distribution both for X and Y choosing a suitable class grouping, for example (choose one for X and one for Y):

Proposed grouping for X

- [0-10), [10-20), [20-30), [30-40), [40-50), [50-60]
- [0-15), [15,30), [30,60]

Proposed grouping for Y

- [18-22), [22-26), [26-30]
- [18-21), [21-24), [24-27), [27-30]
- [18-20), [20-22), [22-24), [24-26), [26-28), [28-30]

Then, for both X and Y:

a) Plot a histogram using the classes you choose in part a).
b) Indicate the class representing the mode
c) Indicate the class including the median
d) Compute the mean on the basis of the aggregate data (i.e. this table), and compare it to the (exact) mean computed on the original data

278. In the following table is reported the frequency distribution of the age of 20 participants in a clinical trial. Compute the mean and the standard deviation.

	n	p	N	P
class	(freq)	(%)	(cum)	(% cum)
[35-45)	5	25	5	25
[45-55)	2	10	7	35
[55-65)	3	15	10	50
[65-75)	7	35	17	85
[75-85)	3	15	20	100
tot	20			

279. The table on the left below reports the distribution of the final score at the exam of Statistics for the Medicine students (degree course in Italian) of the cohort 2011-2012 (the distribution excludes tests failed and scores rejected; just FYI: 30 out of 45 Score=30 were "30-cum-laude").
The table on the right reports the distribution of the number of attempts = times that each student sat in the exam (only students who passed).

a) What is the average score? How many attempts are necessary on average to pass the exam?
b) A student has 50% of probability of getting a score equal to or higher than?
c) How do you interpret the mode of the distribution of Y?
d) Compute the standard deviation of the Score
e) Assume that the Score is a continuous variable, and represent its distribution with a graph, using the classes 18|-21, 21|-26, 26|-31

Score (X)	
8	12
19	9
20	7
21	6
22	6
23	15
24	12
25	6
26	12
27	27
28	11
29	3
30	45
	171

Attempts (Y)	
1	83
2	46
3	34
4	8
	171

280. The mean and the standard deviation of the score at the exam of Statistics for the past 2 cohorts of students was:
- cohort 2011-2012 (Italian; n=171): mean=25.5, std=3.959
- cohort 2012-2013 (English; n=16): mean=26.4, std=4.11
Compute the overall average and compare the variability.

281. There are two urns containing coloured balls. The first urn contains 50 red balls and 50 blue balls. The second urn contains 30 red balls and 70 blue balls. One of the two urns is randomly chosen (both urns have equal probability of being chosen) and then a ball is drawn at random from that urn.

(a) What is the probability to draw a red ball?

(b) If a red ball is drawn, what is the probability that it comes from the first urn?

282. A similar exercise is:

Two hospitals use the same innovative surgical technique for a certain intervention. In the first hospital the probability of successful intervention is 50%. In the second hospital this probability is only 30%. A patient can be admitted to any of the two hospitals with equal probability.

(a) What is the probability that the patient has a successful intervention?

(b) If you know a person who had a successful intervention, what is the probability that he/she was admitted to the first hospital?

283. There are two urns: the first urn contains 3 red balls, 3 blue balls, and 1 black ball. The second urn contains 1 red ball, 2 blue balls, and 1 black ball.

We choose the first urn with probability 2/3 or the second with probability 1/3. Then from the chosen urn we randomly draw one ball (that is, all balls in the urn have equal probabilities).

Let B1 be the event that I choose the first urn, and B2 denote the events that I choose the second, and let A be the event that I draw a red ball.

(a) Compute P(A|B1)

(b) Compute P(A∩B1)

(c) Compute P(A)

284. It is estimated that for the U.S. adult population as a whole, 55 percent are above ideal weight, 20 percent have high blood pressure, and 60 percent either are above ideal weight or have high blood pressure.

(a) What percentage of the population is above ideal weight and have high blood pressure?
(b) Find the conditional probability that a randomly chosen adult of the U.S. population:
  (i) Has high blood pressure given that she or he is above ideal weight
  (ii) Is above ideal weight given that he or she has high blood pressure
(c) Let A be the event that a randomly chosen member of the population is above his or her ideal weight, and let B be the event that this person has high blood pressure. Are A and B independent events?

285.
Consider a given population, a certain disease D and a certain symptom S. We know that 85% of the population members who have the disease D do have the symptom S, while the remaining 15% do not. Suppose that 95% of those who do not have the disease D do not have the symptom S, while 5% do have the symptom S (due to some other cause). Suppose that 10% of the given population have the disease.

(a) What is the probability that a random person chosen from the population has the symptom?
(b) Given that a random person was sampled, and he has the symptom, what is the probability that he or she has disease?
(c) What is the probability that a random person does not have the disease and does not have the symptom?

286. Consider the results of the exams in Statistics for the cohort 2011-2012 reported in Exercise 1.4. Additionally, be aware that in total 228 students tried the exam. Assume that the distribution of scores for those who pass remains the same during the academic year 2013-2014 for n=9 students of the Medicine course in English.

a) What is the probability of passing the exam?
b) If you pass, what is the probability that you get 30 (included 30-cum-laude)?
c) If you pass, what is the probability that you get a score >=27?
d) What is the probability that you pass the exam with score >=27?
e) How many students are expected to pass the exam?
f) What is the probability that all students pass the exam?
g) What is the probability that no student passes the exam?

287. According to a school teacher, when one of his students makes all homeworks assigned, he/she will get the maximum score in 95% of the cases; instead, students who do not make the homeworks properly, have only 45% probability of getting the maximum score. The teacher believes (say, with probability 90%) that his student Tom did not do the homeworks before the last test; however, Tom got the maximum score. What is the probability that Tom cheated at the test? (i.e. Tom did not do the homeworks)

288. A similar exercise is:
There's a leak in an apartment, the owner is pretty sure (say, with probability 90%) that the problem originates in his neighbour's apartment, in this case he can be reimbursed of all the expenses to repair the leak and re-painting. The plumber states that the probability of having the leak if the problem is in the next apartment is only 45%, while there is a 95% probability of having that leak when the problem is in the same apartment. What is the probability that the owner will be reimbursed?

289. Consider a coin and assign to Head and Tail respectively the values 0 and 1. Suppose that P(1)=3/4, and P(0)=1/4. This coin is tossed 3 times independently. Let us define the random variable X = the average result of the 3 tosses.
Write down the distribution of X, that is, the list of all possible values it can take along with their probabilities. Then compute the expectation

** More advanced: compute the variance.*

290. Consider Exercise 3.1. Given the same hypotheses, let us define now the random variable Y = number of 1's obtained in three tosses of the coin. Which is the distribution of Y? and the expected value?

291. Usually (90% of the cases) patients undergoing an intervention in day-hospital actually go home after the intervention, but in 10% of the cases they need to be admitted in hospital. Today there are 3 interventions planned, but only 1 bed available. Compute the probability that today beds will not suffice.

292. Consider the same situation as in Exercise 3.3, but with probability of hospitalization equal to 0.01 and 300 interventions planned; still - incredibly - only 1 bed available. Compute the probability that beds will not suffice in this new situation.

293. Let Z be a standard Normal random variable. Compute the probability of the following intervals: $(-\infty, 2)$; $(-\infty, 2.1)$; $(-\infty,-2.1)$; $(-2.18,+\infty)$; $(0, 2.21)$; $(-2.21, 2.21)$; $(-1, 2.18)$

294. Consider the Score X obtained by the students of Statistics seen in Exercise 1.4 and 2.5, and assume it is a continuous variable with Normal distribution (we know it isn't!) with mean and standard deviation as observed, i.e. μ=25.46 and σ=3.959.

a) If you pass, what is the probability that you get a score >=27? (compare to Exercise 2.5)
b) If you pass, what is the probability that you get a score ≤ 21?

295. The number of bottles of shampoo sold monthly by a certain drug tore is a Normal random variable with mean 212 and standard deviation 40. Find the probability that the next month's shampoo sales will be:

a) Greater than 200
b) Less than 250
c) Greater than 200 but less than 250

296. Consider exactly the same situation as in Exercise 3.4. This time you have 5 beds available. Compute the probability that beds will not suffice in this third situation.

297. Suppose that 46% of the population favours a particular candidate to Major of the town. If a random sample of 200 citizens is chosen, what is the probability that at least 100 of them favour this candidate?

298. The blood cholesterol level of a population of workers has mean 202 and standard deviation 14. A sample of 36 workers is selected. Compute the probability that the sample mean of their blood cholesterol level will lie between 198 and 206. Repeat the exercise for a sample size equal to 64, and explain the change.

299. This exercise helps getting familiar with the quantiles that we use in the inferential procedures that involve using the Normal distribution. Use the table of the standard Normal to find the value(s) z (with precision equal to two decimal places) such that:

a) P(Z > z) = 0.05
b) P(Z < -z) = 0.05
c) P(Z > z) = 0.025
d) P(z1 < Z < z2) = 0.95
e) P(Z > z) = 0.01
f) P(Z > z) = 0.005
g) P(z1 < Z < z2) = 0.99

300. To check compliance with national regulatory, the city board of education needs to estimate the proportion π of women among all secondary school teachers. If there are 518 females in a random sample of 1000 teachers:

- Construct the confidence interval estimate for π at 95%, 90% and 99% confidence level. -
- Before doing this, what do you expect: which confidence interval should be the largest one?
- How do you interpret the 95% CI?
- For example, if the regulatory fixes the target proportion of female teachers to be at least equal to 50%, what does the 95% CI tell about the compliance to this requirement?

301. A similar exercise is:
The most popular drug against menstrual pain indices complete disappearance of pain within 30 minutes from administration in 50% of the women. A pharmaceutical company invested money to produce a drug with larger efficacy. When the drug is tested on 1,000 women, 518 report the target outcome (pain disappeared within 30 min). Did the company achieve the objective? Answer by producing the 95% CI for the proportion π of pain disappearance.

We get (see above) that the interval is between 48.7% and 54.9. So the company is very close to be satisfied, but is not yet sure that the new drug gives more than 50% success.

Notice that this problem can be formulated as an hypothesis test, with H0: π=05 vs. H1: π≠0.5 (we would actually think of a one-sided test, H1: π>0,5; then instead of a too-large conventional alpha=5% we should reduce alpha to 2.5%. This is equivalent to test against the two-sided H1 at the total alpha level of 5%).

Due to the relationship between a 2-sided test at alpha level and the (1-alpha)% confidence interval, we do not need to develop the calculations for the test: the null hypothesis cannot be rejected at 5% level, since the value 0.5 of the null hypothesis belongs to the 95% CI.

302. Continue ex. 4.3-b: as an exercise, proceed anyhow to the test, and:
- compute the p-value
- compute the upper threshold of the rejection region on the original scale (instead of the standardized scale)

303. Medical students undergo a test with score varying between 0 and 100. It is known that the standard deviation of test scores is equal to 11.3, but the mean is unknown. A random sample of 81 students had a sample mean score of 74.6. Compute the 90% confidence interval estimate for the mean μ that represents the average score of all medical students.

304. Traffic authorities claim that the duration of the red light of the town traffic lights is Normally distributed with mean equal to 30 seconds and standard deviation equal to 1.4 seconds. To test this feature a sample of 40 traffic lights's durations was checked, and the observed average duration was 32.2 seconds.
Can we conclude at the 5% level of significance that the authorities are incorrect?
What about using instead a 1% level of significance?

305. A certain disease is known to have a prevalence equal to 1%. However, in a random sample of n=400 members of the population, 11 people were found to have the disease. Test the hypothesis H1 that the true prevalence π in this population is different than 0.01 against the null hypothesis that the prevalence is π= π0=0.01.

306. Using the data of exercise 1.4 on the score at the exam of Statistics for the cohort 2011- 2012 (degree course in Italian) compute a confidence interval for the mean score (at 95% confidence level). Explain in words how we interpret the result. Then test the hypothesis H1 that the mean score is different than 26, with a 2-sided test at 5% significance level.

307. An official ministry office is willing to check whether woman receve lower salaries than men, while being at the same level of experience, expertise, etc. To this purpose they set up a survey among a well-defined population of people with a history of regular employment during the last 10 years, similar experience, similar education etc. These are the results (wages in K euros/year, gross salaries):

Men (group 1): n1=72, mean $\bar{x}_1$=12.2, std = 1.1
Women (group 2): n2=55, mean $\bar{x}_2$ = 10.8, std=0.9

Assume that the variances of the two populations of men and women are equal, and set up a test, writing the null and the alternative hypotheses, computing the p-value, and drawing a conclusion.

308. A public health office detailed the outcome of a prevention project, where 260 elderly people were advised to have a vaccine against flu. A total of 184 agreed to have the vaccine, while the other 76 declined. At the end of the flu season the following outcomes were collected:

	Vaccine	No vaccine
Got flu	10	6
No Flu	174	70

Do the data provide evidence that the people receiving the vaccine had a different chance of contracting the flu from those not receiving the vaccine? Compare these probabilities by an appropriate measure of the effect of "exposure" to vaccine, then test the hypothesis of absence of relation at 5% significance level.

** More advanced (if introduced during the course: repeat the test and compute the p-value using a T-test for two proportions).*

309. A survey investigates on the preferences of Tor Vergata students in terms of computer operating system and type of mobile telephone. The following table represents the results collected on a random sample of size 500.

	Smartphone	Mobile Phone	Tot
MAC OS	102	52	154
WINDOWS	238	40	278
LINUX	52	16	68
Tot	392	108	500

Are the two variables independent at the level of significance α=0.05? What about α=0.01? and how do you interpret your answers?

310. Take the mean and the standard deviation of the score at the exam of Statistics for the 2 cohorts of students seen in exercise 1.5 (and before, in exercise 1.4 and 4.8):

- cohort 2011-2012 (Italian; n=171): mean=25.5, std=3.959
- cohort 2012-2013 (English; n=16): mean=26.4, std=4.11

Do we have evidence to conclude that there is a difference between the two cohorts? Test the hypothesis both computing the p-value and using the rejection region at 5% significance level.

311. Represent and analyze the (linear) relationship between time spent using Facebook and the grades of Statistics exam seen in exercise 1.1, in particular assessing if the time spent on Facebook affects the score in Statistics (discuss the results).

312. The data in the following table show the annual salary (in K euro) X and a measure Y of the productivity of a sample of employees of a company. Some computations are done already in extra columns.

Salary X	Productivity Y	Xi-m(X)	Yi-m(Y)	(Xi-m(X))*(Yi-m(Y))
10.0	1.6	-10.0	-1.3	12.8
15.0	2.0	-5.0	-0.9	4.4
20.0	3.5	0.0	0.6	0.0
21.0	3.0	1.0	0.1	0.1
24.0	3.2	4.0	0.3	1.3
30.0	4.0	10.0	1.1	11.2

Represent the data in a suitable form to show a possible association, then measure it (by the linear correlation coefficient) and find an equation to represent it; furthermore, assess its significance by computing the p-value. Finally, compute the expected productivity of an employee who earns 25 thousand euros.

313. Continues exercise 5.6. The executive director of the company does not like the conclusion that if you pay the employees better, they are more productive. He/she has the following data regarding productivity and salaries in relation to other factors, observed in a larger group of employees: could you suggest him/her some additional analysis to support a position against a salary increase, and find other solutions to increase the productivity?

Impact of Salary on Productivity based on a regression line	b=0.12	p-value<0.0001
Salary according to gender	Salary M: average=22 Salary F: average=19	p-value=0.01
Productivity according to gender	Productivity M: average=2.7 Productivity F: average=2.9	p-value=0.31
Salary according to experience (measured in years)	Salary Exp<5yrs: average=18 Salary Exp≥5 yrs: average=24	p-value=0.002
Productivity according to experience (measured in years)	Productivity Exp<5yrs: average=2.4 Productivity Exp≥5 yrs: average=3.2	p-value=0.012
Salary according to training (measured in hours/year)	Salary <20h: average=19.1 Salary ≥20h: average=21.3	p-value=0.010
Productivity according to training (measured in hours/year)	Productivity <20h: average=2.5 Productivity ≥20h: average=3.0	p-value=0.030