CERGE EI

Instructor: **Paolo Zacchia / Yaroslav Korobka**[1] Updated: November 30, 2025

Course: **Statistics**

1 Basic probability theory

Exercise 1.1. For each of the following experiments, describe the sample space.

(a) Toss a coin four times.

(b) Count the number of insect-damaged leaves on a plant.

(c) Measure the lifetime (in hours) of a particular brand of light bulb.

(d) Record the weights of 10-day-old rats.

(e) Observe the proportion of defectives in a shipment of electronic components.

Solution. The sample space S is the set of all possible outcomes.

(a) For each coin toss in the series we receive either Head (H) or Tail (T). The example of the realization of the series of four tosses: THHH, meaning 1st toss gives T and the remaining tosses give H. The sample space S is the set of all such combinations. One may want to calculate the cardinality of this set. It is $2^4 = 16$: on each position (toss) we have either T or H (2 options) and we observe four positions.

(b) The number of leaves should be a nonnegative integer. Thus, $S = \{0, 1, 2, 3, \ldots\}$.

(c) The lifetime may be less than an hour or some integer number of hours plus the remaining fraction of an hour. Moreover, the lifetime cannot be negative. Thus, $S = \{0, 1, 2, 3, \ldots\}$ in case the lifetime is measured in hours, and $S = \{t : t \geq 0\}$ in case the lifetime is measured in infinitesimal units of time.

(d) We need to choose measurement units. Consider grams, then we may say $S = (0, +\infty)$. We may also define some upper bound, for example 1 kilogram. Then $S = (0, 1000]$.

(e) One can define n to be the number of items in the shipment. Since we need a proportion, the sample space is $S = \{\frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \ldots, \frac{n}{n}\}$.

Exercise 1.2. Verify the following identities:

(a) $A \setminus B = A \setminus (A \cap B) = A \cap B^c$,

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

(c) $B \setminus A = B \cap A^c$,

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

[1]The author thanks Arsenii Scherbov for providing materials for the first half of the course.

Solution. When working with set expressions one may use Venn diagrams to understand how to proceed. However, the illustration does not constitute a formal proof and thus is not sufficient as an answer. To proof that the expression is true one must explicitly show both directions: if $x \in$ LHS (left-hand side) then $x \in$ RHS (right-hand side) and if $x \in$ RHS then $x \in$ LHS.

(a) $A \setminus B \iff x \in A$ and $x \notin B \iff x \in A$ and $x \notin A \cap B \iff x \in A \setminus (A \cap B)$. At the same time, $x \in A$ and $x \notin B \iff x \in A$ and $x \in B^c \iff x \in A \cap B^c$.

(b) By definition $A \cup A^c = S$. Then using the Distributive Law $(B \cap A) \cup (B \cap A^c) = B \cap (A \cup A^c) = B \cap S = B$. One may also show the equality by the technique used in point (a).

(c) $B \setminus A \iff x \in B$ and $x \notin A \iff x \in B$ and $x \in A^c \iff x \in B \cap A^c$.

(d) $A \cup B = A \cup [(B \cap A) \cup (B \cap A^c)] = A \cup (B \cap A) \cup A \cup (B \cap A^c) = A \cup [A \cup (B \cap A^c)] = A \cup (B \cap A^c)$.

Exercise 1.3. For events A and B, find formulas for the probabilities of the following events in terms of the quantities $\mathbb{P}(A)$, $\mathbb{P}(B)$, and $\mathbb{P}(A \cap B)$:

(a) either A or B or both,

(b) either A or B but not both,

(c) at least one of A or B,

(d) at most one of A or B.

Solution.

(a) "either A or B or both" means $A \cup B$, so we have

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

This is because $A \cup B = A \cup (B \cup A^c)$ (Exercise 1.2. point (d)), from it $\mathbb{P}(A \cup B) = \mathbb{P}(A \cup (B \cup A^c)) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

In the last expression, the second equality uses the fact that A and $B \cap A^c$ are disjoint; the last equality comes from rearranging the following $\mathbb{P}(B) = \mathbb{P}((B \cap A) \cup (B \cap A^c)) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c)$ (combines Exercise 1.2. point (b) and the fact that two sets are disjoint).

It is easier to draw the picture and note the fact that we count the area of the intersection two times if we do not subtract $\mathbb{P}(A \cap B)$.

(b) "either A or B but not both" is $(A \cap B^c) \cup (B \cap A^c)$, so we have

$$\begin{aligned}\mathbb{P}((A \cap B^c) \cup (B \cap A^c)) &= \mathbb{P}(A \cap B^c) + \mathbb{P}(B \cap A^c) \\ &= [\mathbb{P}(A) - \mathbb{P}(A \cap B)] + [\mathbb{P}(B) - \mathbb{P}(B \cap A)] \\ &= \mathbb{P}(A) + \mathbb{P}(B) - 2\mathbb{P}(A \cap B).\end{aligned}$$

The easy way is to draw a picture and note that this probability is equal to $\mathbb{P}(A \cup B) - \mathbb{P}(A \cap B)$.

(c) "at least one of A or B" is $A \cup B$, so we have

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

(d) "at most one of A or B" is $(A \cap B)^c$, so we have

$$\mathbb{P}((A \cap B)^c) = 1 - \mathbb{P}(A \cap B).$$

Note that this event also includes the possibility of not A and not B.

Exercise 1.4. Consider two different setups:

(a) A fair dice is cast until a 6 appears. What is the probability that it must be cast more than five times?

(b) Prove that if $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$, then:

- if A and B are mutually exclusive, they cannot be independent,
- if A and B are independent, they cannot be mutually exclusive.

Solution.

(a) It must be cast more than five times if we did not observe the appearance of 6 on first five casts. We do not observe 6 on each cast with probability $5/6$. Since casts of a die are independent (adequate additional interpretation of the setup which you should explicitly mention in your solution) the probability to not receive 6 in first five casts is $(5/6)^5 \approx 0.4$.

(b)
- Let A and B be mutually exclusive. Suppose, for the sake of contradiction, that A and B are independent, then $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. However, from the initial property $A \cap B = \emptyset$ and $\mathbb{P}(A \cap B) = 0$ (mutually exclusive). Since we are given events A and B with positive probabilities, we come to a contradiction $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) = 0$ with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. Thus, mutually exclusive events cannot be independent.
- Independence of A and B together with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$ gives $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) > 0$. It means that $A \cap B \neq \emptyset$. Therefore, A and B are not mutually exclusive by definition. Thus, independent events cannot be mutually exclusive.

  One may also proceed by contradiction: let A and B be independent. Suppose, for the sake of contradiction, that A and B are mutually exclusive, then $\mathbb{P}(A \cap B) = 0$. However, from independence and $\mathbb{P}(A) > 0$, $\mathbb{P}(B) > 0$ we have $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) > 0$. This is a desired contradiction.

Exercise 1.5. Two coins, one with $\mathbb{P}(\text{head}) = u$ and one with $\mathbb{P}(\text{head}) = w$, are to be tossed together independently. Let

$$p_0 = \mathbb{P}(0 \text{ heads occur}), \quad p_1 = \mathbb{P}(1 \text{ heads occur}), \quad p_2 = \mathbb{P}(2 \text{ heads occur}).$$

Can u and w be chosen such that $p_0 = p_1 = p_2$? Prove your answer.

Solution. Start with probabilities: $p_0 = (1-u)(1-w)$, that is, first is T and second is T; $p_1 = (1-u)w + u(1-w)$, that is, first is T and second is H or first is H and second is T; and $p_2 = uw$, that is, both are H. Equating these probabilities, we receive

$$p_0 = p_2 \Rightarrow u + w = 1, \quad p_1 = p_2 \Rightarrow uw = \frac{w+u}{3}.$$

These together imply

$$u(1-u) = \frac{1}{3}.$$

This equation has no solution in the real numbers (more specifically, the right-hand side must be $\leq \frac{1}{4}$ for the solution in real numbers to exist). Thus, we cannot choose legitimate u and w to satisfy the conditions.

Exercise 1.6. Consider telegraph signals "dot" and "dash" sent in the proportion 3:4, where erratic transmissions cause a dot to become a dash with probability $1/4$ and a dash to become a dot with probability $1/3$.

(a) If a dash is received, what is the probability that a dash has been sent? If a dot is received, what is the probability that a dot has been sent?

(b) Assuming independence between signals, if the message dot-dot was received, what is the probability distribution of the four possible messages that could have been sent?

Solution. Let "DA" mean "dash", "DO" mean "dot", "R" mean "received", "S" mean "sent". From dot/dash 3:4 proportion we can calculate unconditional probabilities to observe dots and dashes:

$$\mathbb{P}(\text{dot sent}) := \mathbb{P}(\text{DOS}) = \frac{3}{3+4} = \frac{3}{7}, \quad \mathbb{P}(\text{dash sent}) := \mathbb{P}(\text{DAS}) = \frac{4}{3+4} = \frac{4}{7}.$$

From probabilities of mistakes we deduce other important objects, such as

$$\mathbb{P}(\text{DOR}|\text{DOS}) = 1 - \frac{1}{4} = \frac{3}{4}, \quad \mathbb{P}(\text{DAR}|\text{DAS}) = 1 - \frac{1}{3} = \frac{2}{3},$$

and

$$\mathbb{P}(\text{DOR}|\text{DAS}) = \frac{1}{3}, \quad \mathbb{P}(\text{DAR}|\text{DOS}) = \frac{1}{4}.$$

(a) Using the Bayes' rule, we calculate the following:

$$\begin{aligned}\mathbb{P}(\text{DAS}|\text{DAR}) &= \frac{\mathbb{P}(\text{DAR}|\text{DAS})\cdot\mathbb{P}(\text{DAS})}{\mathbb{P}(\text{DAR}|\text{DAS})\cdot\mathbb{P}(\text{DAS})+\mathbb{P}(\text{DAR}|\text{DOS})\cdot\mathbb{P}(\text{DOS})}\\ &= \frac{(\frac{2}{3})(\frac{4}{7})}{(\frac{2}{3})(\frac{4}{7})+(\frac{1}{4})(\frac{3}{7})} = \frac{32}{41},\end{aligned}$$

which is precisely the probability that a dash has been sent if a dash is received. Next,

$$\begin{aligned}\mathbb{P}(\text{DOS}|\text{DOR}) &= \frac{\mathbb{P}(\text{DOR}|\text{DOS})\cdot\mathbb{P}(\text{DOS})}{\mathbb{P}(\text{DOR}|\text{DOS})\cdot\mathbb{P}(\text{DOS})+\mathbb{P}(\text{DOR}|\text{DAS})\cdot\mathbb{P}(\text{DAS})}\\ &= \frac{(\frac{3}{4})(\frac{3}{7})}{(\frac{3}{4})(\frac{3}{7})+(\frac{1}{3})(\frac{4}{7})} = \frac{27}{43},\end{aligned}$$

which is the probability that a a dot has been sent if a dot is received.

(b) We need yet another probability with DOR condition,

$$\begin{aligned}\mathbb{P}(\text{DAS}|\text{DOR}) &= \frac{\mathbb{P}(\text{DOR}|\text{DAS})\cdot\mathbb{P}(\text{DAS})}{\mathbb{P}(\text{DOR}|\text{DAS})\cdot\mathbb{P}(\text{DAS})+\mathbb{P}(\text{DOR}|\text{DOS})\cdot\mathbb{P}(\text{DOS})}\\ &= \frac{(\frac{1}{3})(\frac{4}{7})}{(\frac{1}{3})(\frac{4}{7})+(\frac{3}{4})(\frac{3}{7})} = \frac{16}{43}.\end{aligned}$$

Independence greatly simplifies calculations. We have DOR two times, so the possible signals sent with corresponding probabilities are,

$$\begin{aligned}\mathbb{P}(\text{DOS}|\text{DOR})\cdot\mathbb{P}(\text{DOS}|\text{DOR}) &= \frac{27}{43}\cdot\frac{27}{43},\\ \mathbb{P}(\text{DOS}|\text{DOR})\cdot\mathbb{P}(\text{DAS}|\text{DOR}) &= \frac{27}{43}\cdot\frac{16}{43},\\ \mathbb{P}(\text{DAS}|\text{DOR})\cdot\mathbb{P}(\text{DOS}|\text{DOR}) &= \frac{16}{43}\cdot\frac{27}{43},\\ \mathbb{P}(\text{DAS}|\text{DOR})\cdot\mathbb{P}(\text{DAS}|\text{DOR}) &= \frac{16}{43}\cdot\frac{16}{43},\end{aligned}$$

which is the final distribution we are looking for.

Exercise 1.7. A students takes a test that consists of 20 multiple-choice questions, each with 4 possible answers. It is necessary to answer 10 questions correctly to pass the test. Find the probability that the student passes the test, given that she is guessing.

Solution. "At least 10 questions correct" means that we need to account for the situations when there are 10 correct guesses or up to all 20 guesses are correct. Thus, we will have a sum of probabilities.

What is the probability to correctly guess k times? For each question the probability of a correct answer is $\frac{1}{4}$, there are 20 questions and the order of correct answers does not matter (all questions are awarded equally). Each test results with k correct guesses has a probability of

$$\left(\frac{1}{4}\right)^k\left(\frac{3}{4}\right)^{20-k}.$$

There are $\binom{n}{k} = C_n^k = \frac{n!}{k!(n-k)!}$ ways to choose the order of correct answers (here $n = 20$). Thus, the probability to correctly guess exactly k answers is

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

To account for different values of $k = 10, \ldots, 20$, we add up these probabilities and receive the final answer

$$\sum_{k=10}^{n=20}\binom{n}{k}\left(\frac{1}{4}\right)^k\left(\frac{3}{4}\right)^{n-k} = 0.014.$$

2 Random variables

Exercise 2.1. Consider a random variable X with the following probability density function and $c < \infty$ being some constant:

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

(a) Determine the value of the parameter c that makes $f_X(x)$ a pdf. Derive the cumulative density function (cdf) of X.

(b) Derive mean, variance, skewness, and kurtosis of X.

(c) Consider another random variable $Y = X^3$, and derive its pdf.

(d) Derive mean, variance, skewness, and kurtosis of Y.

(e) For $g(x) = x^3$ compare $\mathbb{E}[g(X)]$ and $g(\mathbb{E}[X])$. What about $z(x) = x^{\frac{1}{3}}$?

(f) Determine the value of c that makes $f(x) = c \cdot e^{-|x|}$, $-\infty \leq x \leq \infty$ a pdf.

Solution.

(a) By the definition of the pdf,

$$\begin{aligned} \int_0^1 x(1-x)dx &= \int_0^1 (x - x^2)dx = \int_0^1 xdx - \int_0^1 x^2 dx \\ &= \frac{x^2}{2}\Big|_0^1 - \frac{x^3}{3}\Big|_0^1 = \frac{1}{2} - 0 - \frac{1}{3} + 0 = \frac{1}{6}. \end{aligned}$$

Thus, we need $c = 6$ for the pdf to integrate to 1 over the entire support.

By the definition of the cdf,

$$\begin{aligned} F_X(x) &= \int_0^x 6x(1-x)dx \\ &= 6\left(\frac{x^2}{2}\Big|_0^x - \frac{x^3}{3}\Big|_0^x\right) = 6\left(\frac{3x^2 - 2x^3}{6}\right) = 3x^2 - 2x^3, \end{aligned}$$

so that

$$F_X(x) = \begin{cases} 1, & \text{if } x \geq 1, \\ 3x^2 - 2x^3, & \text{if } x \in (0,1), \\ 0, & \text{if } x \leq 0. \end{cases}$$

(b) Again by the definition, we derive

- mean: $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx = \int_0^1 6x^2(1-x)dx = 6\left(\frac{x^3}{3}\Big|_0^1 - \frac{x^4}{4}\Big|_0^1\right) = 6\left(\frac{1}{3} - \frac{1}{4}\right) = 0.5$,
- variance[2]: $\text{var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 0.3 - (0.5)^2 = 0.05$, because the second moment is
$$\mathbb{E}[X^2] = \int_0^1 6x^3(1-x)dx = 6\left(\frac{x^4}{4}\Big|_0^1 - \frac{x^5}{5}\Big|_0^1\right) = 6\left(\frac{1}{4} - \frac{1}{5}\right) = 0.3,$$
- skewness, κ_3, which is a rescaled version of the third *centered* moment:

$$\begin{aligned} \mathbb{E}[(X - \mathbb{E}[X])^3] &= \int_0^1 (x - \mathbb{E}[x])^3 f_X(x)dx = \int_0^1 (x - 0.5)^3 6x(1-x)dx \\ &= 6\int_0^1 (x^3 - 1.5x^2 + 0.75x - 0.125)x(1-x)dx \\ &= 6\left(-\frac{x^6}{6} + \frac{x^5}{2} - \frac{9x^4}{16} + \frac{7x^3}{24} - \frac{x^2}{16}\right)\Big|_0^1 = 0, \end{aligned}$$

[2]Note that you can also proceed by the definition $\text{var}[X] = \int_0^1 (x - \mathbb{E}[x])^2 f_X(x)dx$.

so that the skewness is

$$\kappa_3 = \frac{\mathbb{E}[(X - \mathbb{E}[X])^3]}{(\mathbb{E}[(X - \mathbb{E}[X])^2])^{\frac{3}{2}}} = 0.$$

- kurtosis, κ_4, which is a rescaled version of the *forth* moment,

$$\begin{aligned}\mathbb{E}[(X - \mathbb{E}[X])^4] &= \int_0^1 (x - \mathbb{E}[x])^4 f_X(x)dx = \int_0^1 (x - 0.5)^4 6x(1-x)dx \\ &= 6\left(-\frac{x^7}{7} + \frac{x^6}{2} - \frac{7x^5}{10} + \frac{x^4}{2} - \frac{3x^3}{16} + \frac{x^2}{32}\right)\Bigg|_0^1 \approx 0.005,\end{aligned}$$

so that the kurtosis is

$$\kappa_4 = \frac{\mathbb{E}[(X - \mathbb{E}[X])^4]}{(\mathbb{E}[(X - \mathbb{E}[X])^2])^2} \approx 2.14.$$

Note that the skewness is equal to zero, the fact that follows from the symmetry of the distribution. If the skewness is positive, the distribution is positively skewed or skewed right, meaning that the right tail of the distribution is longer than the left. If the skewness is negative, the distribution is negatively skewed or skewed left, meaning that the left tail is longer.

The kurtosis is a measure of the combined sizes of the two tails. If the kurtosis is greater than 3, then the distribution has heavier tails than a normal distribution. If the kurtosis is less than 3, then the distribution has lighter tails than a normal distribution. Note that kurtosis is always greater or equal to 0.

(c) Let us proceed by the definition of a cdf,

$$\begin{aligned}F_Y(y) &= \mathbb{P}(Y \le y) = \mathbb{P}(X^3 \le y) = \mathbb{P}(X \le y^{1/3}) \\ &= \int_0^{y^{1/3}} 6x(1-x)dx = (3x^2 - 2x^3)\Big|_0^{y^{1/3}} = 3y^{2/3} - 2y.\end{aligned}$$

Note that the support does not change, since $g(0) = 0$ and $g(1) = 1$. The full answer would be:

$$F_Y(y) = \begin{cases} 1, & \text{if } y \ge 1, \\ 3y^{2/3} - 2y, & \text{if } y \in (0,1), \\ 0, & \text{if } y \le 0. \end{cases}$$

The first equality is the definition; the second equality uses our transformation $g(x) = x^3$; the third inequality applies the inverse $g^{-1}(x) = x^{1/3}$ which is increasing on the support of X $(0,1)$ (and thus, preserves the order); the fourth inequality is again a definition.

Using the definition, we can also calculate the pdf,

$$f_Y(y) = \frac{dF_Y(y)}{dy} = 2y^{-1/3} - 2,$$

for $y \in (0,1)$ and 0 otherwise.

(d) Again by the definition,

- mean: $\mathbb{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y)dy = 0.2$,
- variance: $\text{var}[Y] = \int_{-\infty}^{\infty} (y - \mathbb{E}[y])^2 f_Y(y)dy = \frac{13}{300} \approx 0.04$,
- skewness: $\kappa_3[Y] = \frac{\mathrm{E}[(X - \mathrm{E}[X])^3]}{\mathrm{E}[(X - \mathrm{E}[X])^2]^{\frac{3}{2}}} = \frac{\frac{63}{5500}}{(\frac{13}{300})^{\frac{3}{2}}} \approx 1.27$,
- kurtosis: $\kappa_4[Y] = \frac{\mathrm{E}[(X - \mathrm{E}[X])^4]}{\mathrm{E}[(X - \mathrm{E}[X])^2]^2} = \frac{\frac{713}{96250}}{(\frac{13}{300})^2} \approx 3.95$.

(e) Here, the following result is useful:

Theorem 2.2 (Jensen's inequality)**.** *For any random variable X, if $g(x)$ is a convex function, then*

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)].$$

If $g(x)$ is a concave function, the inequality is reversed.

Proof. We focus on the convex case. Let $a + bx$ be the tangent line to $g(x)$ at $x = \mathbb{E}[X]$. Since $g(x)$ is convex, $g(x) \geq a + bx$. Evaluating at $x = X$ and taking expectations, we find

$$\mathbb{E}[g(X)] \geq a + b\mathbb{E}[X] = g(\mathbb{E}[X])$$

as claimed. □

Now, we have

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx = \int_0^1 x^3 6x(1-x) dx = 0.2,$$
$$g(\mathbb{E}[X]) = g(0.5) = 0.5^3 = 0.125.$$

so $\mathbb{E}[g(X)] > g(\mathbb{E}[X])$ in line with Theorem 2.2.
For $z(x) = x^{1/3}$ we have

$$\mathbb{E}[z(X)] = \int_{-\infty}^{\infty} z(x) f_X(x) dx = \int_0^1 x^{1/3} 6x(1-x) dx \approx 0.77,$$
$$z(\mathbb{E}[X]) = z(0.5) = 0.5^{\frac{1}{3}} \approx 0.79,$$

so $\mathbb{E}[z(X)] < z(\mathbb{E}[X])$. It is, again, in line with Theorem 2.2 because $z(x)$ is concave.

(f) $\int_{-\infty}^{\infty} e^{-|x|} dx = \int_{-\infty}^{0} e^{x} dx + \int_0^{\infty} e^{-x} dx = 1 + 1 = 2$, so we need $c = \frac{1}{2}$ for pdf to integrate to 1 over the entire support. Non-negativity is obvious.

Exercise 2.3. In each of the following find the pdf of Y, and show that the pdf integrates to 1.

(a) $Y = X^3$ and $f_X(x) = 42x^5(1-x)$, $0 < x < 1$,

(b) $Y = 4X + 3$ and $f_X(x) = 7e^{-7x}$, $0 < x < \infty$,

(c) $Y = X^2$ and $f_X(x) = 30x^2(1-x)^2$, $0 < x < 1$.

Solution. Here, two following results are useful (we state them separately even though Theorem 2.5 is used in the proof of Theorem 2.4):

Theorem 2.4 (Monotonic transformation)**.** *If X has a pdf $f_X(x)$, $f_X(x)$ is continuous on its support $\mathcal{X}$, $g(x)$ is strictly monotone, and $g^{-1}(y)$ is continuously differentiable on $\mathcal{Y}$, then for $y \in \mathcal{Y}$*

$$f_Y(y) = f_X\left(g^{-1}(y)\right) J(y), \quad J(y) := \left|\frac{d}{dy} g^{-1}(y)\right|.$$

Theorem 2.5. *Let X and $Y = g(X)$ be two random variables, $\mathcal{X}$ and $\mathcal{Y}$ be their supports, and let $F_X(x)$ be the cdf of X.*

(a) If g is increasing in $\mathcal{X}$, we have $F_Y(y) = F_X(g^{-1}(y))$ for all $y \in \mathcal{Y}$.

(b) If g is decreasing in $\mathcal{X}$ and X is a continuous random variable, we have $F_Y(y) = 1 - F_X(g^{-1}(y))$ for all $y \in \mathcal{Y}$.

(a) We have that $y = g(x) = x^3$, $g'(x) = 3x^2$ so $g(x)$ is monotone and increasing. If $0 < x < 1$ then $[g(0) = 0] < [g(x) = y] < [g(1) = 1]$, thus $0 < y < 1$ is our new support. $g^{-1}(y) = y^{\frac{1}{3}}$, $\frac{d}{dy} g^{-1}(y) = \frac{1}{3} y^{-\frac{2}{3}}$.

Applying Theorem 2.4, we have

$$f_Y(y) = f_X\left(g^{-1}(y)\right) \left|\frac{d}{dy} g^{-1}(y)\right| = 42 y^{\frac{5}{3}} \left(1 - y^{\frac{1}{3}}\right) \left(\frac{1}{3} y^{-\frac{2}{3}}\right) = 14y - 14y^{\frac{4}{3}}$$

for $0 < y < 1$ and 0 otherwise. It is a valid pdf, since

$$\int_0^1 14y - 14y^{\frac{4}{3}} dy = 7y^2 - 6y^{\frac{7}{3}} |_0^1 = 1 - 0 = 1.$$

(b) An alternative way how to compute transformations is by using the cdf of the random variable. First, we derive the cdf as

$$F_X(x) = \int_0^x 7e^{-7x}dx = -e^{-7x}\Big|_0^x = -e^{-7x} + e^0 = 1 - e^{-7x},$$

and then $y = g(x) = 4x + 3$, $g'(x) = 4 > 0$ so $g(x)$ is monotone and increasing, and $g^{-1}(y) = \frac{y-3}{4}$. If $0 < x < \infty$ then $[g(0) = 3] < [g(x) = y] < [g(\infty) = \infty]$, thus $3 < y < \infty$ is our new support. Now apply Theorem 2.5 to get

$$F_Y(y) = F_X(g^{-1}(y)) = 1 - e^{-\frac{7}{4}(y-3)}$$

for $3 < y < \infty$ and 0 otherwise. From here, we can differentiate to come to the pdf,

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{7}{4}e^{-\frac{7}{4}(y-3)}$$

for $3 < y < \infty$ and 0 otherwise. It is a valid pdf, since

$$\int_3^\infty \frac{7}{4}e^{-\frac{7}{4}(y-3)}dy = -e^{-\frac{7}{4}(y-3)}|_3^\infty = 0 - (-1) = 1.$$

(c) Yet another alternative, is to proceed "by definition": note that with $g(x) = x^2$ and $0 < x < 1$ we have $[g(0) = 0] < [g(x) = y] < [g(1) = 1]$, thus $0 < y < 1$ is our new support. By definition, $F_Y(y) = \mathbb{P}(Y < y) = \mathbb{P}(X < \sqrt{y}) = F_X(\sqrt{y})$. In this step we find a region in X such that $Y < y$ holds. Differentiate both sides with respect to y (come to pdfs) to receive

$$f_Y(y) = \frac{1}{2\sqrt{y}}f_X(\sqrt{y}) = \frac{1}{2\sqrt{y}}30(\sqrt{y})^2(1-\sqrt{y})^2 = 15y^{\frac{1}{2}}(1-y^{\frac{1}{2}})^2$$

for $0 < y < 1$ and 0 otherwise. It is a valid pdf, since

$$\int_0^1 15y^{\frac{1}{2}}(1-y^{\frac{1}{2}})^2dy = \int_0^1 15y^{\frac{1}{2}} - 30y + 15y^{\frac{3}{2}}dy = 15\frac{2}{3} - 30\frac{1}{2} + 15\frac{2}{5} = 1.$$

3 Parametric distributions

Exercise 3.1. Consider the following:

(a) Derive the moment-generating function (mgf) of $\mathcal{B}(n, p)$ distribution.

(b) It has been determined that 5% of drivers checked at a road stop show traces of alcohol and 10% of drivers checked do not wear seat belts. Assume that the two infractions are independent from one another. If an officer stops five drivers at random: *(i)* calculate the probability that exactly three of the drivers have committed any one of the two offenses, *(ii)* calculate the probability that at least one of the drivers checked has committed at least one of the two offenses.

(c) Derive the mgf of the Poisson(λ) distribution with parameter λ.

(d) Find mean and variance of the Poisson distribution.

(e) Discuss the connection between Poisson(λ) and $\mathcal{B}(n, p)$.

(f) In the manufacture of glassware, bubbles can occur in the glass which reduces the status of the glassware to that of a low quality. If, on average, one in every 1000 items produced has a bubble, calculate the probability that exactly six items in a batch of three thousand are of a low quality.

Solution.

(a) First, note that the pmf of $\mathcal{B}(n, p)$ is $f(x) = C_n^x p^x (1-p)^{n-x}$, where $x \geq 0$ is an integer, and p is a probability of a success. The mgf is then

$$M_X(t) = \mathbb{E}[e^{tX}] = \sum_{x=0}^n e^{tx}C_n^x p^x(1-p)^{n-x} = \sum_{x=0}^n C_n^x(pe^t)^x(1-p)^{n-x} = [pe^t + (1-p)]^n,$$

where the last equality uses the binomial formula $\sum_{x=0}^n C_n^x u^x v^{n-x} = (u+v)^n$ with $u = pe^t$ and $v = 1-p$. You can check that $\mathbb{E}[X] = np$ and $\text{var}[X] = np(1-p)$ by using this mgf.

(b) Denote X the number of drivers who have committed any one of the two offenses; A meaning a driver shows traces of alcohol, and B meaning a driver does not wear a seat belt. We have

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = 0.05 + 0.1 - 0.05 \cdot 0.1 = 0.145 := p.$$

n is 5 so we are having $\mathcal{B}(n = 5, p = 0.145)$. Then for *(i)*, we have $\mathbb{P}(X = 3) = C_5^3(0.145)^3(1 - 0.145)^{5-3} = 0.0223$.

For *(ii)*, use the rule that $\mathbb{P}(\text{at least one}) = 1 - \mathbb{P}(\text{no one})$. The resulting probability is $1 - C_5^0(1 - 0.145)^5 = 0.543$.

(c) First, note that the pmf is $f(x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$, where $x \geq 0$ is an integer and λ is a positive constant. We have

$$M_X(t) = \mathbb{E}[e^{tX}] = \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} e^{tx} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^t\lambda)^x}{x!} = e^{-\lambda} e^{e^t\lambda} = e^{\lambda(e^t-1)},$$

where the fourth equality uses power series expansion of exponential function $e^x = \sum_{n=0}^{\infty} \dfrac{x^n}{n!} = 1 + x + \dfrac{x^2}{2!} + \dfrac{x^3}{3!} + \ldots$, which is valid $\forall x \in \mathbb{R}$.

(d) To compute any nth moment, differentiate the mgf w.r.t. t n times and evaluate the derivative at $t = 0$,

$$\mathbb{E}[X^n] = \frac{d^n M_X(t)}{dt^n}\Big|_{t=0}.$$

For our needs, we require the first and the second moment. We have

$$\begin{aligned}
\mathbb{E}[X] &= \frac{de^{\lambda(e^t-1)}}{dt}\Big|_{t=0} = e^{\lambda(e^t-1)}\lambda e^t|_{t=0} = e^{\lambda(1-1)}\lambda e^0 = \lambda, \\
\mathbb{E}[X^2] &= \frac{d^2e^{\lambda(e^t-1)}}{dt^2}\Big|_{t=0} = e^{\lambda(e^t-1)}\lambda e^t \lambda e^t + e^{\lambda(e^t-1)}\lambda e^t|_{t=0} = \lambda^2 + \lambda, \\
\text{var}\,[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.
\end{aligned}$$

(e) Binomial probabilities can be approximated by Poisson probabilities (the latter are easier to calculate). Approximation is valid if n is large and np is small. To get the result, consider $X \sim \mathcal{B}(n, p)$ and $Y \sim$ Poisson(λ) with $\lambda = np$. The approximation is then defined as having $P(X = x) \approx P(Y = x)$ for large n and small np. Formally, we show the result if we show that their mgfs converge, thus having $F_X(u) = F_Y(u)$ for all u (see Theorem 2.3.11 in *Casella & Berger*).

Define $p = \dfrac{\lambda}{n}$, let $n \to \infty$, and look at the limit of $M_X(t)$,

$$M_X(t) = [pe^t + (1 - p)]^n = [1 + p(e^t - 1)]^n = \left[1 + \frac{\lambda(e^t - 1)}{n}\right]^n \to e^{\lambda(e^t-1)} = M_Y(t),$$

where the limit uses the fact that $\lim_{n\to\infty}\left(1 + \dfrac{a_n}{n}\right)^n = e^a$ with $a = \lim_{n\to\infty} a_n$.

(f) Denote X the number of items with bubbles (low quality). Clearly, $X \sim \mathcal{B}(n = 3000, p = 0.001)$, so we are in the context where n is large and np is small. Thus, the Poisson approximation should be accurate.

Then $\lambda = np = 3000 \cdot 0.001 = 3$ and $\mathbb{P}(X = 6) \approx \dfrac{\lambda^x e^{-\lambda}}{x!} = \dfrac{3^6 e^{-3}}{6!} \approx 0.0498 \cdot 1.0125 \approx 0.05041$. Using any statistical programming language, we can compute the exact probability, which is 0.05038.

Note: remember the usual interpretation of Poisson as the number of occurrences in a given interval for which the average rate of occurrences is λ (think of an example with a student waiting for a bus at a bus stop).

Exercise 3.2. A random point (X, Y) is distributed uniformly on the square with vertices $(1, 1)$, $(1, -1)$, $(-1, 1)$, and $(-1, -1)$. That is, the joint pdf is $f(x, y) = \frac{1}{4}$ on the square.

(a) Determine the probabilities of the following events: *(i)* $X^2 + Y^2 \leq 1$, *(ii)* $2X - Y \geq 0$, *(iii)* $|X + Y| \leq 2$.

(b) Derive the marginal and the conditional densities of X. What would you conclude regarding the same densities of Y?

(c) A and B agree to meet at a certain place between 1 PM and 2 PM. Suppose they arrive at the meeting place independently and randomly during the hour. Find the distribution of the length of time A waits for B. If B arrives before A, define A's waiting time as 0.

Solution.

(a) The easiest way to proceed in this example is to invoke the geometric interpretation of events and probabilities and compare areas on a square. The total area is equal to $2 \cdot 2 = 4$.

For *(i)*, the condition is true only in a circle of radius one which can be summarized by the inequality $x^2 + y^2 \leq 1$. The area of this circle is equal to $\pi \cdot r^2 = \pi$. The the probability of the event will equal to the ratio of the area of the event (circle) to the total area (square). Thus, $\mathbb{P}(X^2 + Y^2 \leq 1) = \pi/4$.

For *(i)*, the condition may be rearranged as $2x \geq y$. The corresponding area is the one below the line $y = 2x$. Due to symmetry (or using graphical argument) one can conclude that the area is exactly the half of the initial square. Thus, $\mathbb{P}(2X - Y \geq 0) = 2/4 = 1/2$.

(iii) is straightforward, because 2 is an upper bound for $|X + Y|$ when $X \in [-1, 1]$ and $Y \in [-1, 1]$. Thus, $\mathbb{P}(|X + Y| \leq 2) = 1$.

(b) Start with the marginal distribution of X. By definition:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy = \int_{-1}^{1} \frac{1}{4}dy = \frac{1}{4}y\Big|_{-1}^{1} = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

Thus, the marginal density of X is equal to $1/2$ on the support $X \in [-1, 1]$ and we conclude that $X \sim \mathcal{U}[-1, 1]$. Due to symmetry, the same conclusion is valid for the marginal density of Y.

Conditional density of X given Y is by definition

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}$$

when $Y \in [-1, 1]$ and $X \in [-1, 1]$, and 0 in other cases. Thus, $X|Y \sim \mathcal{U}[-1, 1]$ and the conditional distribution is the same as the marginal distribution. The convenience is easy to ruin. For example, one may rotate (45°) the square around $(0, 0)$ point and work out the integration to see the result (concentrate on bounds of integration). Note that here $f_{X,Y}(x, y) = f_X(x)f_Y(y)$, so X and Y are statistically independent.

(c) Let A = time that A arrives and B = time that B arrives. The random variables A and B are independent $\mathcal{U}[1, 2]$. Their properties (except the support) will not change if we preserve the length of the interval and look at $\mathcal{U}[0, 1]$ instead (the problem is the same for 1 PM – 2 PM, 2 PM – 3 PM, 12 PM – 1 PM). So we can consider joint pdf which is uniform on the square $(0, 1) \times (0, 1)$. Define X = amount of time A waits for B. Then if $x < 0$, $F_X(x) = \mathbb{P}(X \leq x) = 0$, and if $x > 1$, $F_X(x) = \mathbb{P}(X \leq x) = 1$.

For $x = 0$ (B is first), we have

$$F_X(0) = \mathbb{P}(X = 0) = \mathbb{P}(B \leq A) = \int_0^1 \int_0^a 1dbda = \int_0^1 b\Big|_0^a da = \int_0^1 ada = \frac{a^2}{2}\Big|_0^1 = \frac{1}{2}.$$

If $0 < x < 1$,

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leq x) = 1 - \mathbb{P}(X > x) = 1 - \mathbb{P}(B - A > x) \\ &= 1 - \int_0^{1-x} \int_{a+x}^1 1dbda = 1 - \int_0^{1-x} b\Big|_{a+x}^1 da = 1 - \int_0^{1-x} 1 - a - xda = 1 - \left(a - \frac{a^2}{2} - xa\right)\Big|_0^{1-x} \\ &= 1 - \left(1 - x - \frac{(1-x)^2}{2} - x + x^2\right) = \frac{1}{2} + x - \frac{x^2}{2}. \end{aligned}$$

The answer should be concluded with the full description of the cdf which combines all results obtained.

Exercise 3.3. Consider the following:

(a) Show that $\text{cov}(X, c) = 0$ for any random variable X with finite mean and any constant c.

(b) Let X and Y be independent random variables with means μ_X, μ_Y and variances σ_X^2, σ_Y^2. Find an expression for the correlation of XY and Y in terms of these means and variances.

(c) Let X_1, X_2, and X_3 be uncorrelated random variables, each with mean μ and variance σ^2. Find, in terms of μ and σ^2, $\text{cov}(X_1 + X_2, X_2 + X_3)$ and $\text{cov}(X_1 + X_2, X_1 - X_2)$.

(d) Prove that for any random vector $(X_1, \ldots, X_n)$ the following holds:

$$\text{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{var}(X_i) + 2 \sum_{1 \le i < j \le n} \text{cov}(X_i, X_j).$$

Solution.

(a) $\text{cov}(X, c) = \mathbb{E}[Xc] - \mathbb{E}[X]\mathbb{E}[c] = c\mathbb{E}[X] - c\mathbb{E}[X] = 0$.

(b) We have

$$\rho_{XY,Y} = \frac{\text{cov}(XY, Y)}{\sigma_{XY}\sigma_Y} = \frac{\mathbb{E}[XY^2] - \mu_{XY}\mu_Y}{\sigma_{XY}\sigma_Y} = \frac{\mathbb{E}[X]\mathbb{E}[Y^2] - \mu_X\mu_Y\mu_Y}{\sigma_{XY}\sigma_Y} = \frac{\mu_X(\sigma_Y^2 + \mu_Y^2) - \mu_X\mu_Y^2}{\sigma_{XY}\sigma_Y}$$

where the last line uses independence of X and Y. The remaining quantity is

$$\begin{aligned}\sigma_{XY}^2 &= \mathbb{E}[(XY)^2] - (\mathbb{E}[XY])^2 = \mathbb{E}[X^2]\mathbb{E}[Y^2] - (\mathbb{E}[X])^2(\mathbb{E}[Y])^2 \\ &= (\sigma_X^2 + \mu_X^2)(\sigma_Y^2 + \mu_Y^2) - \mu_X^2\mu_Y^2 = \sigma_X^2\sigma_Y^2 + \sigma_X^2\mu_Y^2 + \sigma_Y^2\mu_X^2.\end{aligned}$$

Thus, $\rho_{XY,Y} = \dfrac{\mu_X(\sigma_Y^2 + \mu_Y^2) - \mu_X\mu_Y^2}{\sigma_Y\sqrt{\sigma_X^2\sigma_Y^2 + \sigma_X^2\mu_Y^2 + \sigma_Y^2\mu_X^2}} = \dfrac{\mu_X\sigma_Y}{\sqrt{\sigma_X^2\sigma_Y^2 + \sigma_X^2\mu_Y^2 + \sigma_Y^2\mu_X^2}}$.

(c) $\text{cov}(X_1 + X_2, X_2 + X_3) = \mathbb{E}[(X_1 + X_2)(X_2 + X_3)] - \mathbb{E}[X_1 + X_2]\mathbb{E}[X_2 + X_3] = (4\mu^2 + \sigma^2) - 4\mu^2 = \sigma^2$,
$\text{cov}(X_1 + X_2, X_1 - X_2) = \mathbb{E}[(X_1 + X_2)(X_1 - X_2)] - \mathbb{E}[X_1 + X_2]\mathbb{E}[X_1 - X_2] = \mathbb{E}[X_1^2 - X_2^2] = 0$.

(d) Denote $\mu_i = \mathbb{E}(X_i)$. Then

$$\begin{aligned}\text{var}\left(\sum_{i=1}^{n} X_i\right) &= \text{var}(X_1 + \ldots + X_n) \\ &= \mathbb{E}\left[((X_1 + \ldots + X_n) - (\mu_1 + \ldots + \mu_n))^2\right] \\ &= \mathbb{E}[((X_1 - \mu_1) + \ldots + (X_n - \mu_n))^2] \\ &= \sum_{i=1}^{n} \mathbb{E}[(X_i - \mu_i)^2] + 2 \sum_{1 \le i < j \le n} \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] \\ &= \sum_{i=1}^{n} \text{var}(X_i) + 2 \sum_{1 \le i < j \le n} \text{cov}(X_i, X_j)\end{aligned}$$

Exercise 3.4. Consider the following:

(a) Let Z be a random variable with pdf $f(z)$. Define z_α to be a number that satisfies $\alpha = \mathbb{P}(Z > z_\alpha) = \int_{z_\alpha}^{\infty} f(z)dz$. Show that if X is a random variable with pdf $\frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right)$ and $x_\alpha = \sigma z_\alpha + \mu$, then $\mathbb{P}(X > x_\alpha) = \alpha$. (Thus if a table of z_α values was available, then values of x_α could be easily computed for any member of the location-scale family).

Theorem 3.5. *Let $f(\cdot)$ be any pdf. Let μ be any real number, and let σ be any positive real number. Then X is a random variable with pdf $\frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right)$ if and only if there exists a random variable Z with pdf $f(z)$ and $X = \sigma Z + \mu$.*

(b) The average number of acres burned by forest and range fires in a large New Mexico county is 4300 acres per year, with a standard deviation of 750 acres. The distribution of the number of acres burned is normal. What is the probability that between 2500 and 4200 acres will be burned in any given year? What number of burnt acres corresponds to the 38^{th} percentile? 87^{th} percentile?

(c) *Smirnov transform*. Let X be a continuous random variable whose distribution function F_X is strictly increasing on the domain of X. Let $U = F_X(X)$ (transformation that applies $F_X(\cdot)$ to X). What is the distribution of U? How one can exploit this fact to simulate independent realizations of X?

Solution.

(a) $\mathbb{P}(X > x_\alpha) = \mathbb{P}(\sigma Z + \mu > \sigma z_\alpha + \mu) = \mathbb{P}(Z > z_\alpha)$. The conclusion about $X = \sigma Z + \mu$ comes from Theorem 3.5.

(b) Let X be the number of acres burned by forest and range fires per year. $X \sim \mathcal{N}(\mu, \sigma)$, $\mu = 4300$, $\sigma = 750$. Then

$$\begin{aligned}\mathbb{P}(2500 < X < 4200) &= \mathbb{P}\left(\frac{2500-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{4200-\mu}{\sigma}\right)\\ &= \mathbb{P}(-2.40 < Z < -0.13) = P(Z < -0.13) - P(Z < -2.40)\\ &= \Phi(-0.13) - \Phi(-2.40) = 1 - \Phi(0.13) - (1 - \Phi(2.40))\\ &= \Phi(2.40) - \Phi(0.13) = 0.9918 - 0.5517 = 0.4401.\end{aligned}$$

Because $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$ and cdf of Z is denoted as $\Phi(Z)$.

Denote 38^{th} percentile by p_{38}, $\mathbb{P}(X < p_{38}) = 0.38$. It holds that $\mathbb{P}\left(\frac{X-\mu}{\sigma} < \frac{p_{38}-\mu}{\sigma}\right) = \mathbb{P}\left(Z < \frac{p_{38}-\mu}{\sigma}\right) = 0.38$. Since the probability is less then 0.5, we are in the negative part of the distribution. So $\mathbb{P}(Z < \frac{p_{38}-\mu}{\sigma}) = \Phi(\frac{p_{38}-\mu}{\sigma}) = 1 - \Phi(-\frac{p_{38}-\mu}{\sigma}) = 0.38$. From here, $\Phi(-\frac{p_{38}-\mu}{\sigma}) = 1 - 0.38 = 0.62$, so $-\frac{p_{38}-\mu}{\sigma} = 0.31$. Finally, $p_{38} = \mu - 0.31\sigma = 4300 - 0.31 \cdot 750 = 4067.5$.

For the 87^{th} percentile, it is evident from table that for Z the value is 1.13. Then for X we have $4300 + 1.13 \cdot 750 = 5147.5$.

(c) $\mathbb{P}(U < u) = \mathbb{P}(F_X(X) < u) = \mathbb{P}(X < F_X^{-1}(u)) = F_X(F_X^{-1}(u)) = u$. Since cdf returns values from $[0,1]$, we conclude that the cdf of the transformation is

$$F_U(u) = \begin{cases} 1, & \text{if } u > 1,\\ u, & \text{if } u \in [0,1],\\ 0, & \text{if } u < 0,\end{cases}$$

so U is uniform random variable on $[0,1]$.

The result states that we can take any cdf (well-behaved, meaning invertible and for this we require $F_X(x)$ to be strictly monotone, because it implies that the function is bijective and this is enough for it to be invertible), apply it as a transformation on the random variable with the chosen cdf and receive a uniform random variable on $[0,1]$. But we can go in the opposite direction because of the invertibility of the transformation. So we can simulate independent random variables having distribution function F_X by simulating U, a uniform random variable on $[0,1]$, and then taking $X = F_X^{-1}(U)$. The advantage is that uniform is very easy to simulate. Note that for any distribution it is highly likely that there are more efficient (in terms of computations) ways to do the simulation, but the results is nice and it may help to appreciate other similar algorithms. One of the important drawbacks of this algorithm is that some cdfs are invertible but the inverse has no closed form solution (for example, normal cdf).

Exercise 3.6. A point is generated at random in the plane according to the following polar scheme: a radius R is chosen, where the distribution of R^2 is χ^2 with 2 degrees of freedom; independently, an angle θ is chosen, where $\theta \sim \mathcal{U}(0, 2\pi)$. Find the joint distribution of $X = R\cos\theta$ and $Y = R\sin\theta$.

Solution. Since R and θ are independent, the joint pdf of $T = R^2$ and θ is

$$f_{T,\theta}(t,\theta) = f_T(t) f_\theta(\theta) = \frac{1}{4\pi} e^{-t/2}$$

with $0 < t < \infty$ and $0 < \theta < 2\pi$.

The transformation is $x = \sqrt{t}\cos\theta$, $y = \sqrt{t}\sin\theta$ which is one-to-one (note that we take a 2×1 vector and transform it into a 2×1 vector). From here $t = x^2 + y^2$ (first component of $g^{-1}(x,y)$) and $\theta = \arctan(y/x)$ (second component). We also need the Jacobian (actually, we need only the absolute value of the determinant of this matrix) of the inverse transformation:

$$J = \begin{pmatrix} \dfrac{\partial g_1^{-1}(x,y)}{\partial x} & \dfrac{\partial g_1^{-1}(x,y)}{\partial y} \\ \dfrac{\partial g_2^{-1}(x,y)}{\partial x} & \dfrac{\partial g_2^{-1}(x,y)}{\partial y} \end{pmatrix} = \begin{pmatrix} 2x & 2y \\ \dfrac{-y}{x^2+y^2} & \dfrac{x}{x^2+y^2} \end{pmatrix}.$$

Derivatives of $g_1^{-1}(x,y)$ are obvious. For $g_2^{-1}(x,y) = \arctan(y/x)$ we have

$$\frac{\partial \arctan(y/x)}{\partial x} = -\frac{y}{x^2}\frac{1}{1+\frac{y^2}{x^2}} = \frac{-y}{x^2+y^2}, \quad \frac{\partial \arctan(y/x)}{\partial y} = \frac{1}{x}\frac{1}{1+\frac{y^2}{x^2}} = \frac{1}{x}\frac{x^2}{x^2+y^2} = \frac{x}{x^2+y^2}.$$

The absolute value of the determinant is $|\det(J)| = 2$. Now we combine the ingredients using Theorem 3.1,

$$f_{X,Y}(x,y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}$$

with $-\infty < x, y < \infty$. From here it should be clear that X and Y are independent standard normals. One may want to work out the marginal density of X or Y to convince yourself,

$$f_X(x) = \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} dy = \frac{1}{2\sqrt{\pi}} e^{-\frac{1}{2}x^2} \int_{-\infty}^{\infty} \frac{1}{2\sqrt{\pi}} e^{-\frac{1}{2}y^2} dy = \frac{1}{2\sqrt{\pi}} e^{-\frac{1}{2}x^2}$$

for $-\infty < x < \infty$. This is because in the last equality we integrate $\mathcal{N}(0,1)$ over the entire support.

Exercise 3.7. Let X, Y, and Z be discrete random variables. Show the following generalizations of the law of iterated expectations:

(a) $\mathbb{E}[Z] = \mathbb{E}[\mathbb{E}[Z|X,Y]]$,

(b) $\mathbb{E}[Z|X] = \mathbb{E}[\mathbb{E}[Z|X,Y]|X]$,

(c) $\mathbb{E}[Z] = \mathbb{E}[\mathbb{E}[\mathbb{E}[Z|X,Y]|X]]$.

We start with a stick of length l. We break it at a point which is chosen randomly and uniformly over its length, and keep the piece that contains the left end of the stick. We then repeat the same process on the piece that we were left with.

(a) What is the expected value of the length of the piece that we are left with after breaking twice?

(b) What is the variance of the length of the piece that we are left with after breaking twice?

Solution. In the following we are going to use two very important results:

Theorem 3.8 (Law of iterated expectations)**.** *If* $\mathbb{E}|Y| < \infty$, *then* $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$.

Theorem 3.9 (Law of total variance)**.** $var\,[Y] = \mathbb{E}[var\,[Y|X]] + var\,[\mathbb{E}[Y|X]]$.

(a) We begin by writing the definition for $\mathbb{E}[Z|X,Y]$

$$\mathbb{E}[Z|X=x, Y=y] = \sum_z z f_{Z|X,Y}(z|x,y).$$

Since $\mathbb{E}[Z|X,Y]$ is a function of the random variables X and Y, and is equal to $\mathbb{E}[Z|X=x,Y=y]$ (number) whenever $X=x$ and $Y=y$, which happens with probability $f_{X,Y}(x,y)$, using the rules for manipulations with expectations, we have

$$\begin{aligned}\mathbb{E}[Z|X,Y] &= \sum_x \sum_y \mathbb{E}[Z|X=x,Y=y] f_{X,Y}(x,y) \\ &= \sum_x \sum_y \sum_z z f_{Z|X,Y}(z|x,y) f_{X,Y}(x,y) \\ &= \sum_x \sum_y \sum_z z f_{X,Y,Z}(x,y,z) \\ &= \mathbb{E}[Z].\end{aligned}$$

(b) Using the same definition of $\mathbb{E}[Z|X,Y]$ as in (a) we additionally condition on the event $X=x$:

$$\begin{aligned}\mathbb{E}[\mathbb{E}[Z|X,Y=y]|X=x] &= \sum_y \mathbb{E}[Z|X=x,Y=y] f_{Y|X}(y|x) \\ &= \sum_y \sum_z z f_{Z|X,Y}(z|x,y) f_{Y|X}(y|x) \\ &= \sum_y \sum_z z f_{Y,Z|X}(y,z|x) \\ &= \mathbb{E}[Z|X=x].\end{aligned}$$

Since this is true for all possible values of x, we have $\mathbb{E}[Z|X] = \mathbb{E}[\mathbb{E}[Z|X,Y]|X]$. One may generalize this result for 3+ conditions and refer to it as "the smaller information set always prevails".

(c) Take expectations of both sides of the results from part (b) to obtain

$$\mathbb{E}[\mathbb{E}[Z|X]] = \mathbb{E}[\mathbb{E}[\mathbb{E}[Z|X,Y]|X]]$$

By the law of iterated expectations, the left-hand side above is $\mathbb{E}[Z]$, which establishes the desired result.

Let Y be the length of the piece after we break for the first time. Let X be the length after we break for the second time.

(a) The law of iterated expectations states that $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$, and we have $\mathbb{E}[X|Y] = \frac{Y}{2}$ and $\mathbb{E}[Y] = \frac{l}{2}$. So then

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}\left[\frac{Y}{2}\right] = \frac{1}{2}\mathbb{E}[Y] = \frac{1}{2}\frac{l}{2} = \frac{l}{4}.$$

(b) Recall that the variance of a uniform random variable distributed over $[a,b]$ is $\frac{(b-a)^2}{12}$. Since Y is uniformly distributed over $[0,l]$, we have $\text{var}[Y] = \frac{l^2}{12}$, $\text{var}[X|Y] = \frac{Y^2}{12}$. We know that $\mathbb{E}[X|Y] = \frac{Y}{2}$, and so

$$\text{var}[\mathbb{E}[X|Y]] = \text{var}\left[\frac{Y}{2}\right] = \frac{1}{4}\text{var}[Y] = \frac{l^2}{48}.$$

Also,

$$\mathbb{E}[\text{var}[X|Y]] = \mathbb{E}\left[\frac{Y^2}{12}\right] = \int_0^l \frac{y^2}{12} f_Y(y)dy = \frac{1}{12}\frac{1}{l}\int_0^l y^2 dy = \frac{l^2}{36}.$$

If we combine the results we have,

$$\text{var}[X] = \mathbb{E}[\text{var}[X|Y]] + \text{var}[\mathbb{E}[X|Y]] = \frac{l^2}{36} + \frac{l^2}{48} = \frac{7l^2}{144}.$$

4 Finite-sample properties

Exercise 4.1. Assume that X_i are i.i.d. with $\mathbb{E}[X_i] = \mu$ and $\text{var}[X_i] = \sigma^2$. Define the sample average as $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ with $i = 1, \ldots, n$.

(a) Find $\mathbb{E}[\bar{X}_n]$.

(b) Find $\text{var}\,[\bar{X}_n]$.

(c) Show that

$$\mathbb{E}\left[\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\right] = 0 \quad \text{and} \quad \text{var}\left[\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\right] = 1.$$

Solution.

(a) Plug in the sample average into the expectation operator and use the property of linearity,

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

(b) Similarly,

$$\text{var}\,[\bar{X}_n] = \text{var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}\,[X_i] = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

Notice, there are no covariance terms due to the fact that data are i.i.d.

(c) First, put the constants out of the expectation and then plug in known expressions from (a) and (b),

$$\mathbb{E}\left[\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\right] = \frac{\sqrt{n}}{\sigma}\mathbb{E}\left[(\bar{X}_n - \mu)\right] = \frac{\sqrt{n}}{\sigma}(\mu - \mu) = 0,$$

and similarly for the variance,

$$\text{var}\left[\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\right] = \frac{n}{\sigma^2}\text{var}\left[(\bar{X}_n) - \mu\right] = \frac{n}{\sigma^2}\text{var}\left[\bar{X}_n\right] = \frac{n}{\sigma^2}\frac{\sigma^2}{n} = 1.$$

Thus, the normalization of $\bar{X}_n$ in the Central Limit Theorem gives random variables that have the same mean and variance as the limiting $\mathcal{N}(0,1)$ distribution.

Exercise 4.2. Assume that $X_i \sim \mathcal{N}(\mu_x, \sigma_x^2)$ with $i = 1, \ldots, n_x$, $Y_i \sim \mathcal{N}(\mu_y, \sigma_y^2)$ with $i = 1, \ldots, n_y$ and $X_i \perp Y_i$.

(a) Find $\mathbb{E}[\bar{X}_n - \bar{Y}_n]$.

(b) Find $\text{var}\,[\bar{X}_n - \bar{Y}_n]$.

(c) Find the distribution of $\bar{X}_n - \bar{Y}_n$.

(d) Find the distribution of $\bar{X}_n + \bar{Y}_n$.

(e) Find the distribution of $S_x = \dfrac{n_x\widehat{\sigma_x^2}}{\sigma_x^2}$ where $\widehat{\sigma_x^2} = \frac{1}{n}\sum_{i=1}^{n_x}(X_i - \bar{X}_n)^2$ is the sample variance.

(f) Find the distribution of $\dfrac{n_x\widehat{\sigma_x^2}}{(n_x-1)\sigma_x^2}\Big/\dfrac{n_y\widehat{\sigma_y^2}}{(n_y-1)\sigma_y^2}$ using the following theorem:

Theorem 4.3 (Ratio of χ^2 as F-distribution)**.** *Let $Z_1 \sim \chi^2_m$, $Z_2 \sim \chi^2_p$ where $m, p \in \mathbb{N}$ and $Z_1 \perp Z_2$. Then*

$$\frac{Z_1}{m}\Big/\frac{Z_2}{p} \sim F(m,p)\,,$$

that is, ratio has an F-distribution with m and p degrees of freedom.

Which kind of hypothesis could we test using the statistic above?

Solution.

(a) From Problem 1 we know that

$$\mathbb{E}[\bar{X}_n - \bar{Y}_n] = \mu_x - \mu_y.$$

(b) By definition and using the results from Problem 1,

$$\begin{aligned}\text{var}\left[\bar{X}_n - \bar{Y}_n\right] &= \text{var}\left[\bar{X}_n\right] + \text{var}\left[\bar{Y}_n\right] - 2\text{cov}\left[\bar{X}_n, \bar{Y}_n\right] \\ &= \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y} - \frac{2}{n_x n_y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\text{cov}\left[X_i, Y_j\right] \\ &= \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\end{aligned}$$

due to the fact of the independence,

$$\text{cov}\left[X_i, Y_j\right] = \mathbb{E}[X_i Y_j] - \mathbb{E}[X_i]\mathbb{E}[Y_j] = \mathbb{E}[X_i]\mathbb{E}[Y_j] - \mathbb{E}[X_i]\mathbb{E}[Y_j] = 0.$$

(c) From the property of the normal random variables, we conclude that the sum of the independent normal random variables is normal. Hence, $\bar{X}_n$ and $\bar{Y}_n$ are both normal. Using the results from (a) and (b) we conclude that

$$\bar{X}_n - \bar{Y}_n \sim \mathcal{N}\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right).$$

(d) By similar arguments,

$$\bar{X}_n + \bar{Y}_n \sim \mathcal{N}\left(\mu_x + \mu_y, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right).$$

Notice that the sign influences only the expectation, since it is a linear operator.

(e) Here, I omit x indices to simplify the notation. First, rewrite the expression for $\widehat{\sigma^2}$ so that it contains the true value μ. It will be useful later when computing S directly. First, add and subtract μ,

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 &= \sum_{i=1}^{n}\left[(X_i - \mu) - (\bar{X}_n - \mu)\right]^2 \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X}_n - \mu)\sum_{i=1}^{n}(X_i - \mu) + \sum_{i=1}^{n}(\bar{X} - \mu)^2 \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X}_n - \mu)\cdot n \cdot (\bar{X}_n - \mu) + n(\bar{X}_n - \mu)^2 \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X}_n - \mu)^2 \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - \sum_{i=1}^{n}(\bar{X}_n - \mu)^2.\end{aligned}$$

Hence, we can write

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{X}_n - \mu)^2. \tag{1}$$

Next, denote $Z_i = \frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0,1)$ to be a standardized random variable. Also, denote its sample average as

$$\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i - \mu}{\sigma} = \frac{\frac{1}{n}\sum_{i=1}^{n}X_i - \frac{n}{n}\mu}{\sigma} = \frac{\bar{X}_n - \mu}{\sigma}.$$

Because $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, it follows that $\bar{Z}_n \sim \mathcal{N}\left(0, \frac{1}{n}\right)$ and $\sqrt{n}\bar{Z}_n \sim \mathcal{N}(0,1)$. Squaring both sides and using the fact that square of the standardized normal variable is chi-squared with 1 degree of freedom, we have

$$n\bar{Z}_n^2 \sim \chi_1^2.$$

Hence, we have

$$S = \frac{n\widehat{\sigma^2}}{\sigma^2} = \sum_{i=1}^{n} Z_i^2 - n\bar{Z}_n^2 = Q_n - Q_1.$$

We know that $Z_i \sim \mathcal{N}(0,1)$ and its square is $Z_i^2 \sim \chi_1^2$. Sum of n terms of Z_i^2 is then $\sum_{i=1}^n Z_i^2 \sim \chi_n^2$. Hence, S is distributed as

$$S = \chi_n^2 - \chi_1^2 \sim \chi_{n-1}^2$$

under the condition that $Q_1 \perp S$.

(f) From (e), we conclude that

$$\frac{n_x\widehat{\sigma^2}_x}{(n_x-1)\sigma_x^2} \Bigg/ \frac{n_y\widehat{\sigma^2}_y}{(n_y-1)\sigma_y^2} \sim F(n_x-1, n_y-1).$$

We can test a hypothesis of the equality of the variances σ_x^2 and σ_y^2. For example, $H_0 : \sigma_x^2 = \sigma_y^2$ against $H_1 : \sigma_x^2 \neq \sigma_y^2$.

Exercise 4.4. Suppose that the random variables $Y_1, \ldots, Y_n$ satisfy

$$Y_i = \beta x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where $x_1, \ldots, x_n$ are fixed constants, and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ with σ^2 unknown.

(a) Show that $\hat{\beta} = \sum Y_i / \sum x_i$ is an unbiased estimator of β.

(b) Show that $\tilde{\beta} = [\sum Y_i / x_i] / n$ is also an unbiased estimator of β.

(c) Calculate the exact variances of estimators from (a) and (b). Which one would you prefer?

Solution.

(a) Plug in the expression of the estimator into the expectation and simplify:

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}\left[\frac{\sum_i Y_i}{\sum_i x_i}\right] = \frac{1}{\sum_i x_i}\sum_i \mathbb{E}[Y_i] = \frac{1}{\sum_i x_i}\sum_i \beta x_i = \beta.$$

Hence, $\hat{\beta}$ is an unbiased estimator of β.

(b) Similarly,

$$\mathbb{E}[\tilde{\beta}] = \mathbb{E}\left[\frac{1}{n}\sum_i \frac{Y_i}{x_i}\right] = \frac{1}{n}\sum_i \frac{\mathbb{E}[Y_i]}{x_i} = \frac{1}{n}\sum_i \frac{\beta x_i}{x_i} = \beta.$$

Hence, $\tilde{\beta}$ is an unbiased estimator of β.

(c) First, we calculate the variance of $\hat{\beta}$,

$$\operatorname{var}[\hat{\beta}] = \operatorname{var}\left[\frac{\sum_i Y_i}{\sum_i x_i}\right] = \frac{1}{(\sum_i x_i)^2}\sum_i \operatorname{var}[Y_i] = \frac{\sum_i \sigma^2}{(\sum_i x_i)^2} = \frac{n\sigma^2}{n^2\bar{x}^2} = \frac{\sigma^2}{n\bar{x}^2}.$$

Similarly,

$$\operatorname{var}[\tilde{\beta}] = \operatorname{var}\left[\frac{1}{n}\sum_i \frac{Y_i}{x_i}\right] = \frac{1}{n^2}\sum_i \frac{\operatorname{var}[Y_i]}{x_i^2} = \frac{\sigma^2}{n^2}\sum_i \frac{1}{x_i^2}.$$

To compare the variances, notice that we could compare

$$\frac{1}{\bar{x}^2} \lessgtr \frac{1}{n}\sum_i \frac{1}{x_i^2}.$$

Because $g(u) = 1/u^2$ is convex, using Jensen's inequality we have

$$\frac{1}{\bar{x}^2} \leq \frac{1}{n}\sum_i \frac{1}{x_i^2}.$$

Hence, $\operatorname{var}[\hat{\beta}] \leq \operatorname{var}[\tilde{\beta}]$ and $\hat{\beta}$ should be preferred.

5 Maximum likelihood estimation

Exercise 5.1. A random variable X is said to have a Pareto distribution with parameter β, denoted as $X \sim \text{Pareto}(\beta)$, if it is continuously distributed with density

$$f_X(x;\beta) = \begin{cases} \beta x^{-\beta-1}, & \text{if } x > 1, \\ 0, & \text{otherwise.} \end{cases}$$

A random sample $x_1, \ldots, x_N$ from the Pareto(β) population is available. Derive the maximum-likelihood estimator of β. Does it maximize the log-likelihood function?

Solution. By definition and property of the likelihood function,

$$\begin{aligned} \hat{\beta}_{ML} &= \arg \max_{\beta} \mathcal{L}_N(\beta | x_1, \ldots, x_N) \\ &= \arg \max_{\beta} \log \mathcal{L}_N(\beta | x_1, \ldots, x_N). \end{aligned}$$

Thus, for our case, log-likelihood function looks like

$$\begin{aligned} \log \mathcal{L}_N(\beta | x_1, \ldots, x_N) &= \sum_{i=1}^{N} \log f_X(x_i;\beta) \\ &= \sum_{i=1}^{N} \log(\beta) - (\beta+1)\log(x_i) \\ &= N\log(\beta) - (\beta+1)\sum_{i=1}^{N} \log(x_i). \end{aligned}$$

To obtain the maximum-likelihood estimator of β, we take the derivative w.r.t. β and set it to zero,

$$\frac{\partial \log \mathcal{L}_N}{\partial \beta} = \frac{N}{\hat{\beta}_{ML}} - \sum_{i=1}^{N} \log(x_i) = 0. \tag{2}$$

From (2) we get

$$\hat{\beta}_{ML} = \frac{N}{\sum_{i=1}^{N} \log(x_i)}.$$

To check that $\hat{\beta}_{ML}$ is indeed the optimum, take the second derivative of the log-likelihood w.r.t. β,

$$\frac{\partial^2 \log \mathcal{L}_N}{\partial \beta^2}\Big|_{\beta=\hat{\beta}_{ML}} = -\frac{N}{\beta^2} < 0.$$

Hence, our estimator maximizes the log-likelihood function.

Exercise 5.2. Suppose that the random variables $Y_1, \ldots, Y_n$ satisfy

$$Y_i = \beta x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where $x_1, \ldots, x_n$ are fixed constants, and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ with σ^2 unknown. Find the MLE of β. Is it unbiased?

Solution. Fix the value of σ^2. Write the likelihood function as

$$\mathcal{L}_N(\beta | Y_1, \ldots, Y_N) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{\beta^2}{2\sigma^2}\sum_{i=1}^{N} x_i^2\right) \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N} Y_i^2 + \frac{\beta}{\sigma^2}\sum_{i=1}^{N} Y_i x_i\right),$$

and taking the logarithm, we obtain the log-likelihood function:

$$\log \mathcal{L}_N(\beta | Y_1, \ldots, Y_N) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{\beta^2}{2\sigma^2}\sum_{i=1}^{N} x_i^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N} Y_i^2 + \frac{\beta}{\sigma^2}\sum_{i=1}^{N} Y_i x_i.$$

Taking the first order derivative w.r.t. β and setting it to zero, we have

$$\frac{\partial \log \mathcal{L}_N}{\partial \beta} = -\frac{\hat{\beta}_{ML}}{\sigma^2}\sum_{i=1}^{N} x_i^2 + \frac{1}{\sigma^2}\sum_{i=1}^{N} Y_i x_i = 0.$$

From it, we obtain the MLE of β,

$$\hat{\beta}_{ML} = \frac{\sum_{i=1}^{N} Y_i x_i}{\sum_{i=1}^{N} x_i^2}.$$

To verify that the log-likelihood is indeed maximized, take the second order derivative w.r.t. β,

$$\frac{\partial^2 \log \mathcal{L}_N}{\partial \beta^2}\Big|_{\beta=\hat{\beta}_{ML}} = -\frac{1}{\sigma^2}\sum_{i=1}^{N} x_i^2 < 0.$$

To check whether the estimator is biased,

$$\mathbb{E}[\hat{\beta}_{ML}] = \mathbb{E}\left[\frac{\sum_{i=1}^{N} Y_i x_i}{\sum_{i=1}^{N} x_i^2}\right] = \frac{\sum_{i=1}^{N} \mathbb{E}[Y_i] x_i}{\sum_{i=1}^{N} x_i^2} = \frac{\sum_{i=1}^{N} \beta x_i x_i}{\sum_{i=1}^{N} x_i^2} = \beta,$$

hence, the ML estimator of β is unbiased.

Exercise 5.3. Let $x_1, \ldots, x_N$ be a random sample from a gamma(α, β) population.

(a) Find the MLE of β, assuming α is known.

(b) If α and β are both unknown, there is no explicit formula for the MLE of α, but the maximum can be found numerically. How can we use the result in part (a) to reduce the problem to the maximization of a univariate function?

Solution.

a) Write the likelihood function as

$$\mathcal{L}_N(\beta|x_1, \ldots, x_N) = \prod_{i=1}^{N} \frac{1}{\Gamma(\alpha)\beta^\alpha} x_i^{\alpha-1} e^{-x_i/\beta} = \frac{1}{\Gamma(\alpha)^N \beta^{N\alpha}} \left[\prod_{i=1}^{N} x_i\right]^{\alpha-1} e^{-\sum_i x_i/\beta},$$

and taking the logarithm,

$$\log \mathcal{L}_N(\beta|x_1, \ldots, x_N) = -\log \Gamma(\alpha)^N - N\alpha \log \beta + (\alpha - 1) \log \left[\prod_{i=1}^{N} x_i\right] - \frac{\sum_i x_i}{\beta}.$$

Taking the first order derivative w.r.t. β and setting to zero, we have

$$\frac{\partial \log \mathcal{L}_N}{\partial \beta} = -\frac{N\alpha}{\hat{\beta}_{ML}} + \frac{\sum_i x_i}{\hat{\beta}_{ML}^2} = 0.$$

From the expression above, we get the ML estimator of β,

$$\hat{\beta}_{ML} = \frac{\sum_{i=1}^{N} x_i}{N\alpha} = \frac{\bar{x}}{\alpha}.$$

To check that this is a maximum, calculate

$$\frac{\partial^2 \log \mathcal{L}_N}{\partial \beta^2}\Big|_{\beta=\hat{\beta}_{ML}} = \frac{N\alpha}{\beta^2} - \frac{2\sum_i x_i}{\beta^3}\Big|_{\beta=\hat{\beta}_{ML}} = \frac{(N\alpha)^3}{(\sum_i x_i)^2} - \frac{2(N\alpha)^3}{(\sum_i x_i)^2} = -\frac{(N\alpha)^3}{(\sum_i x_i)^2} < 0.$$

Because $\hat{\beta}_{ML}$ is the unique point where the derivative is zero and it is a local maximum, it is a global maximum. That is, $\hat{\beta}_{ML}$ is the MLE.

b) Now the likelihood function is

$$\mathcal{L}_N(\alpha, \beta | x_1, \ldots, x_N) = \frac{1}{\Gamma(\alpha)^N \beta^{N\alpha}} \left[\prod_{i=1}^{N} x_i \right]^{\alpha-1} e^{-\sum_i x_i/\beta},$$

the same as in part (a) except α and β are both variables now. There is no closed form for the MLEs $\hat{\alpha}_{ML}$ and $\hat{\beta}_{ML}$. One approach to finding $\hat{\alpha}_{ML}$ and $\hat{\beta}_{ML}$ would be to numerically maximize the function of two arguments. But it is usually best to do as much as possible analytically, first, and perhaps reduce the complexity of the numerical problem. From part (a), for each fixed value of α, the value of β that maximizes $\mathcal{L}(\alpha, \beta | x_1, \ldots, x_N)$ is $\sum_i x_i / N\alpha$. Substituting this into the likelihood function, we are left with one variable α,

$$\begin{aligned} \mathcal{L}(\alpha | x_1, \ldots, x_N) &= \frac{1}{\Gamma(\alpha)^N (\sum_i x_i / N\alpha)^{N\alpha}} \left[\prod_{i=1}^{N} x_i \right]^{\alpha-1} e^{-\sum_i x_i / (\sum_i x_i / N\alpha)} \\ &= \frac{1}{\Gamma(\alpha)^N (\sum_i x_i / N\alpha)^{N\alpha}} \left[\prod_{i=1}^{N} x_i \right]^{\alpha-1} e^{-N\alpha} \rightarrow \max_{\alpha}. \end{aligned}$$

From the problem defined above, we get the ML estimator $\hat{\alpha}_{ML}$ which can be used to compute $\hat{\beta}_{ML}$.