Probability and Statistics
FS 2017
Session Exam
22.08.2017
Time Limit: 180 Minutes

Name: ____________________________

Student ID: ____________________________

This exam contains 16 pages (including this cover page) and 10 questions. A Formulae sheet is provided with the exam.

Grade Table (for grading use only, please leave empty)

Question	Points	Score
1	10	
2	10	
3	10	
4	10	
5	10	
6	10	
7	10	
8	10	
9	10	
10	10	
Total:	100	

Informations. Read this carefully.

- Please justify all your statements carefully. Explain the steps of your reasoning. Otherwise no points will be given.
- You are expected to write full sentences when giving your answer.
- DO NOT WRITE with red or green pens. DO NOT WRITE with a pencil.
- Your answers should be *readable*.
- Write your name on all the sheets you intend to hand in before the end of the exam.

GOOD LUCK

1. (10 points) **Counting Problems**

   (a) (3 points) A small voting district has K female voters and L male voters. A random sample of N voters, where $N \leq K + L$, is drawn uniformly at random from the population. Let $n \leq \min\{K, N\}$. *What is the probability that exactly n of the N voters will be female? Do you recognize a known distribution? Write down its name.*

   (b) (7 points) At a wedding, a group of $n \geqslant 3$ people (including the married couple, Alice and Bob) wants to take a picture. They all stand in a line, the order of the people in the line taken uniformly at random among the permutations of n elements. *What is the probability that exactly k guests stand between Alice and Bob, for $k \in \{0, 1, ..., n-2\}$?*

Solution

(a) There are $\binom{L+K}{N}$ ways to pick the N voters among the $L+K$ district voters. There are $\binom{K}{n}\binom{L}{N-n}$ ways to choose the n female and $N-n$ male voters. Let X be the random variable that models the number of female voters when choosing N voters at random. The probability to pick n female voters when choosing N voters at random is then

$$P(X = n) = \frac{\binom{K}{n}\binom{L}{N-n}}{\binom{L+K}{N}}.$$

This is a hypergeometric distribution.

(b) First, order the guests and assign them a number between 3 and n (1 is given to Alice and 2 to Bob). We are looking for the number of permutations π such that $|\pi(1) - \pi(2)| = k + 1$. Let $\mathcal{P}_k$ be defined as the set of such permutations.

Such a permutation is constructed as follows (without loss of generality, assume that $\pi(1) < \pi(2)$, any such permutation in $\mathcal{P}_k$ corresponds to another permutation π' in $\mathcal{P}_k$ with $\pi'(1) = \pi(2)$ and $\pi'(2) = \pi(1)$):

- Choose a value for $\pi(1)$ in $\{1, 2, ..., n-k-1\}$.
- Choose k of the guests (integer in $\{3, ..., n\}$). There are $k!$ ways to assign them values between $\pi(1) + 1$ and $\pi(1) + k = \pi(2) - 1$.
- Then there are $(n-2-k)!$ ways to assign the remaining guests to the remaining places in the line.

Then there are $2(n-k-1)\binom{n-2}{k}k!(n-2-k)!$ permutations in $\mathcal{P}_k$. Since the guest permutation is chosen at random and the total number of permutations of n elements is $n!$, the probability that exactly k guests are between Alice and Bob is given by

$$\frac{2(n-k-1)\binom{n-2}{k}k!(n-2-k)!}{n!} = \frac{2(n-k-1)}{n(n-1)}.$$

2. (10 points) **Conditional Probabilities**

A group of people is considered in order to assess the reliability of a disease test procedure. In this group of people we know that the probability that an individual taken at random has the disease is $\frac{1}{3}$. The presence of the disease is then tested by the detection of some markers in the blood. After examination of the results, it is found that the probability of a true positive (markers are detected by the procedure in a sick patient) is $\frac{4}{5}$. The probability of a false positive (markers are found in the blood sample of a healthy patient) is $\frac{1}{5}$.

(a) (5 points) *Compute the probability of the patient being sick given that the markers were found in the blood sample, i.e. the test result is positive.*

(b) (5 points) *Compute the probability of the patient being sick given that no markers were found in the blood sample, i.e. the test result is negative.*

Solution

Let us first define the following events:

- D :={The patient has the disease.}
- P :={The result of the test is positive.}

By the problem formulation, we know that $\mathbb{P}(D) = \frac{1}{3}$, $\mathbb{P}(P|D) = \frac{4}{5}$ and $\mathbb{P}(P|D^c) = \frac{1}{5}$.

(a) We want to compute the probabilty $\mathbb{P}(D|P)$. For this, we use Bayes theorem (or the definition of the conditional probability and the law of total probabilities). We get

$$\begin{aligned}\mathbb{P}(D|P) &= \frac{\mathbb{P}(P|D)\,\mathbb{P}(D)}{\mathbb{P}(P|D)\,\mathbb{P}(D) + \mathbb{P}(P|D^c)\,\mathbb{P}(D^c)} \\ &= \frac{\frac{4}{5}\cdot\frac{1}{3}}{\frac{4}{5}\cdot\frac{1}{3}+\frac{1}{5}\cdot\frac{2}{3}} \\ &= \frac{4}{15}\cdot\frac{15}{6} \\ &= \frac{2}{3}.\end{aligned}$$

(b) We want to compute the probabilty $\mathbb{P}[D|P^c]$. By the same arguments as above, we obtain,

$$\begin{aligned}\mathbb{P}(D|P^c) &= \frac{\mathbb{P}(P^c|D)\,\mathbb{P}(D)}{\mathbb{P}(P^c|D)\,\mathbb{P}(D) + \mathbb{P}(P^c|D^c)\,\mathbb{P}(D^c)} \\ &= \frac{\frac{1}{5}\cdot\frac{1}{3}}{\frac{1}{5}\cdot\frac{1}{3}+\frac{4}{5}\cdot\frac{2}{3}} \\ &= \frac{1}{15}\cdot\frac{15}{9} \\ &= \frac{1}{9.}\end{aligned}$$

3. (10 points) **Doubling Strategy**

A fair coin is used in a simple game, i.e. the probability that it falls on head when tossed is $\frac{1}{2}$. The mechanism of the game is the following: the player bets k dollars and the coin is tossed. If head shows, he gets his k dollars back and wins k additional dollars. If tail shows, he loses his k dollars.

The player now follows the following strategy: he first bets 1 dollar. Then, he applies the following algorithm:

- if he wins, he stops playing,
- if he loses, he plays again and doubles his bet (2 dollars for the second game, 4 dollars for the third, etc...).

As long as he loses, he keeps on playing, betting 2^k dollars in the $(k+1)$-th game. We assume that the throws of the coin are independent. Let X be the amount of dollars the player bets in the last game.

(a) (3 points) *Write down the probability distribution of X, i.e. give $\mathbb{P}\left(X = 2^k\right)$ for $k \in \{0, 1, 2, ...\}$.*

(b) (3 points) *Show that the expectation of X does not exist.*

(c) (3 points) *Compute the overall amount of money earned at the end of the game.*

(d) (1 point) *Is the strategy applied in this exercise a good strategy? Why / Why not?*

Solution

(a) Let $k \in \{0, 1, 2, ...\}$. Then $X = 2^k$ if and only if the player loses the first k games and wins the $(k+1)$-th game. Using the independence of the throws of the coin, we get

$$\mathbb{P}\left(X = 2^k\right) = \left(\frac{1}{2}\right)^k \cdot \frac{1}{2} = \frac{1}{2^{k+1}}.$$

(b) We have

$$\mathbb{E}\left[X\right] = \sum_{k=0}^{\infty} P\left(X = 2^k\right) 2^k = \sum_{k=0}^{\infty} \frac{1}{2^{k+1}} 2^k = \lim_{n \to \infty} \sum_{k=0}^{n} \frac{1}{2} = \infty.$$

Hence the expectation of X does not exist.

(c) If the player wins at the n-th round, he has lost in the $(n-1)$ previous games

$$\sum_{k=0}^{n-1} 2^k = \frac{1-2^n}{1-2} = 2^n - 1.$$

Therefore after the last game he has won 1 dollar.

(d) The strategy applied in this exercise is a good strategy only if we can borrow an unlimited amount of money and if we are also allowed to bet an unlimited amount of money, since then we always have an overall win of 1 dollar.

4. (10 points) **Pareto and the 80-20-"Rule"**

Let $\gamma > 1$ and $m > 0$. Let X be a random variable with density function f given by

$$f(x) = \begin{cases} \frac{\gamma m^\gamma}{x^{\gamma+1}}, & \text{if } x \geq m, \\ 0 & \text{else.} \end{cases}$$

The random variable X is said to have a Pareto distribution with parameters γ and m.

(a) (4 points) Let $x \geq a \geq m$. *Compute* $\mathbb{P}(X > a)$, $\mathbb{P}(X > x|X > a)$ *and the density function of* X *given* $\{X > a\}$.

Hint: Recall that the density function of X *given* $\{X > a\}$ *is obtained by differentiating the distribution function of* X *given* $\{X > a\}$, *i.e.*

$$f_{X|\{X>a\}}(x) = \frac{d\mathbb{P}(X \leq x|X > a)}{dx}.$$

(b) (1 point) Let $a \geq m$. *Compute* $\mathbb{E}[X|X > a]$.

Assume now and for the rest of the exercise, that $\gamma = \frac{\ln(5)}{\ln(4)}$ and $m = 1$.

(c) (2 points) *Compute the* 0.8-*quantile* $q_{0.8}$ *of this Pareto distribution.*

Hint: Recall that in general the α-*quantile* q_α *is defined as a real number that satisifes* $P(X \leq q_\alpha) \geq \alpha$ *and* $P(X \geq q_\alpha) \geq 1 - \alpha$.

(d) (3 points) *Compute* $\mathbb{E}[X]$ *and* $\mathbb{E}\left[X \mathbb{1}_{\{X > q_{0.8}\}}\right]$. *Compare the two values by computing the ratio* $\frac{\mathbb{E}\left[X \mathbb{1}_{\{X > q_{0.8}\}}\right]}{\mathbb{E}[X]}$ *and give an interpretation.*

Solution

(a) We compute

$$\begin{aligned} \mathbb{P}(X > a) &= \int_a^\infty \frac{\gamma m^\gamma}{x^{\gamma+1}} dx \\ &= \left[-\frac{m^\gamma}{x^\gamma} \right]_a^\infty \\ &= \left(\frac{m}{a} \right)^\gamma \end{aligned}$$

and

$$\begin{aligned} \mathbb{P}(X > x|X > a) &= \frac{\mathbb{P}(X > x)}{\mathbb{P}(X > a)} \\ &= \left(\frac{a}{x} \right)^\gamma. \end{aligned}$$

Then $\mathbb{P}(X \leq x | X > a) = 1 - \mathbb{P}[X > x | X > a] = 1 - \left(\frac{a}{x}\right)^{\gamma}$ and the density function of X given $\{X > a\}$ at $x > a$ is obtained by differentiation:

$$f_{\{X|X>a\}}(x) = \frac{\gamma a^{\gamma}}{x^{\gamma+1}}.$$

(b) We have

$$\begin{aligned}\mathbb{E}[X|X > a] &= \int_a^{\infty} \frac{\gamma a^{\gamma}}{x^{\gamma+1}} x dx \\ &= \left[-\frac{\gamma a^{\gamma}}{(\gamma-1)x^{\gamma-1}}\right]_a^{\infty} \\ &= \frac{\gamma}{\gamma-1} a.\end{aligned}$$

(c) By definition and since the distribution is absolutely continuous, $q_{0.8}$ satisfies

$$P(X \geq q_{0.8}) = 1 - 0.8 = 0.2.$$

On the other hand, using (a), we have

$$P(X \geq q_{0.8}) = \left(\frac{m}{q_{0.8}}\right)^{\gamma} = \left(\frac{1}{q_{0.8}}\right)^{\frac{\ln(5)}{\ln(4)}}.$$

Thus we can calculate

$$q_{0.8} = 5^{\frac{\ln(4)}{\ln(5)}} = \exp\left[\frac{\ln(4)}{\ln(5)} \ln(5)\right] = 4.$$

(d) We have

$$\begin{aligned}\mathbb{E}[X] &= \int_1^{\infty} \frac{\gamma}{x^{\gamma+1}} x dx \\ &= \frac{\gamma}{\gamma-1} \\ &= \frac{\ln(5)}{\ln(5)-\ln(4)}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[X \mathbb{1}_{\{X > q_{0.8}\}}\right] &= \int_{q_{0.8}}^{\infty} \frac{\gamma}{x^{\gamma+1}} x dx \\ &= \frac{\gamma}{\gamma-1} q_{0.8}^{1-\gamma} \\ &= \frac{\ln(5)}{\ln(5)-\ln(4)} 4^{\frac{\ln(4)-\ln(5)}{\ln(4)}} \\ &= \frac{\ln(5)}{\ln(5)-\ln(4)} \frac{4}{5}.\end{aligned}$$

So we have $\frac{\mathbb{E}\left[X \mathbb{1}_{\{X > q_{0.8}\}}\right]}{\mathbb{E}[X]} = \frac{4}{5}$. We conclude that the highest 20% percent of the distribution make up for 80% of the distribution's mean.

5. (10 points) **Continuous Joint Distribution**

Let (X, Y) be a random vector with joint probability density function

$$f(x,y) = \begin{cases} cx\mathrm{e}^{-xy} & \text{if } 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 2, \\ 0 & \text{else,} \end{cases}$$

for some constant $c > 0$.

(a) (3 points) *Compute the value of c.*

(b) (3 points) *Compute the marginal density of X.*

(c) (4 points) *Compute the expectation of the product XY.*

Solution

(a-b) The marginal density function of X is

$$f_X(x) = \int_{\mathbb{R}} f(x,y)\,dy = \begin{cases} c(1-\mathrm{e}^{-2x}) & \text{for } 0 \leq x \leq 1, \\ 0 & \text{else}\,. \end{cases}$$

The constant c making f_X a probability density function should satisfy

$$\int_{\mathbb{R}} f_X(x)\,dx = 1.$$

We calculate

$$\begin{aligned} \int_0^1 c(1-\mathrm{e}^{-2x})dx &= c\left([x]_0^1 - \left[-\frac{1}{2}\mathrm{e}^{-2x}\right]_0^1\right) \\ &= c\left(1 + \frac{1}{2}\mathrm{e}^{-2} - \frac{1}{2}\right) \\ &= \frac{c}{2}\left(1+\mathrm{e}^{-2}\right). \end{aligned}$$

Thus $\int_{\mathbb{R}} f_X(x)\,dx = 1$ if and only if $c = \frac{2}{1+\mathrm{e}^{-2}}$. We conclude that the marginal density of X is given by

$$f_X(x) = \begin{cases} \frac{2(1-\mathrm{e}^{-2x})}{1+\mathrm{e}^{-2}} & \text{for } 0 \leq x \leq 1, \\ 0 & \text{else}\,. \end{cases}$$

(c) We calculate

$$\begin{aligned}
\mathbb{E}[XY] &= \int_{\mathbb{R}} \int_{\mathbb{R}} xy f_{X,Y}(x,y)\, dydx \\
&= \frac{2}{1+\mathrm{e}^{-2}} \int_0^1 \int_0^2 x^2 y \mathrm{e}^{-xy}\, dydx \\
&= \frac{2}{1+\mathrm{e}^{-2}} \int_0^1 x^2 \left(\left[-\frac{y}{x} \mathrm{e}^{-xy} \right]_0^2 + \int_0^2 \frac{1}{x} \mathrm{e}^{-xy} dy \right) dx \\
&= \frac{2}{1+\mathrm{e}^{-2}} \int_0^1 \left(-2x\mathrm{e}^{-2x} + 1 - \mathrm{e}^{-2x} \right) dx \\
&= \frac{2}{1+\mathrm{e}^{-2}} \left(\left[x\mathrm{e}^{-2x} \right]_0^1 - \int_0^1 \mathrm{e}^{-2x} dx + 1 - \int_0^1 \mathrm{e}^{-2x} dx \right) \\
&= \frac{2}{1+\mathrm{e}^{-2}} \left(\left[x\mathrm{e}^{-2x} \right]_0^1 + \left[\mathrm{e}^{-2x} \right]_0^1 + 1 \right) \\
&= \frac{2}{1+\mathrm{e}^{-2}} \left(\mathrm{e}^{-2} + \mathrm{e}^{-2} \right) \\
&= \frac{4}{1+\mathrm{e}^{2}}.
\end{aligned}$$

6. (10 points) **Markov, Chebyshev and Chernoff Inequalities**

   (a) (3 points) Let X be a random variable such that for $t < t_0$ the following holds:
   $$\psi(t) := \mathbb{E}\left[e^{tX}\right] < \infty.$$
   This function is called the moment generating function of X. Let $x \in \mathbb{R}$. Then we define
   $$\kappa(t) := \ln[\psi(t)]$$
   and
   $$S(x) := \sup_{0<t<t_0} \{tx - \kappa(t)\}.$$
   Prove the Chernoff inequality
   $$\mathbb{P}(X \geq x) \leq e^{-S(x)}.$$
   Hint: Apply the generalized Chebyshev inequality with the function $g(y) := e^{ty}$.

   Assume now and for the rest of the exercise, that X is distributed as a standard normal random variable $\mathcal{N}(0,1)$.

   (b) (2 points) *Compute the moment generating function of X. Give the details of your computations.*

   (c) (2 points) Let $x > 0$. *Apply the Chernoff inequality from (a) to $X \sim \mathcal{N}(0,1)$.*

   (d) (2 points) Let $x > 0$. *Apply the Markov inequality to $|X - \mathbb{E}[X]|^2$ to get another upper bound for $\mathbb{P}(X \geq x)$.*

   (e) (1 point) *What can you say about the efficiency of the two inequalities you got in (c) and in (d)?*

Solution

(a) Let $0 < t < t_0$. Since the function g given in the hint is positive and increasing, we have by the generalized Chebyshev inequality that
$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}\left[e^{tX}\right]}{e^{tx}} = e^{-[tx-\kappa(t)]}.$$
Since this holds for all $0 < t < t_0$, we can conclude that
$$\mathbb{P}(X \geq x) \leq \inf_{0<t<t_0} e^{-[tx-\kappa(t)]} = e^{-\sup_{0<t<t_0}\{tx-\kappa(t)\}} = e^{-S(x)}.$$

(b) We calculate
$$\begin{aligned}
\mathbb{E}\left[e^{tX}\right] &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-x^2/2} e^{tx} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{1}{2}\left(x^2-2tx+t^2\right)} e^{\frac{t^2}{2}} dx \\
&= e^{\frac{t^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{1}{2}y^2} dy \\
&= e^{\frac{t^2}{2}},
\end{aligned}$$

where in the third equality the change of variables $y = x - t$ was used.

(c) We have $\kappa(t) = \frac{t^2}{2}$ for $t \in \mathbb{R}_+$. We compute the function S for X:

$$S(x) = \sup_{t \in \mathbb{R}_+ \setminus \{0\}} \{tx - \kappa(t)\} = \sup_{t \in \mathbb{R}_+ \setminus \{0\}} \left\{ tx - \frac{t^2}{2} \right\} = \frac{x^2}{2},$$

since $x > 0$. The Chernoff bound applied to X gives

$$\mathbb{P}(X \geq x) \leq \mathrm{e}^{-\frac{x^2}{2}}.$$

(d) Note that $\mathbb{E}[X] = 0$. We can write

$$\mathbb{P}(X \geq x) \leq \mathbb{P}\left(X^2 \geq x^2\right) = \mathbb{P}\left(|X - \mathbb{E}[X]|^2 \geq x^2\right).$$

Using the Markov inequality and that $\mathrm{Var}(X) = 1$, we find

$$\mathbb{P}(X \geq x) \leq \frac{\mathrm{Var}(X)}{x^2} = \frac{1}{x^2}.$$

(e) The Chernoff bound is much tighter than the Markov bound. However, it requires that the distribution of the random variable has exponential moments, which is a much higher requirement than only second moment.

7. (10 points) **Limit Theorem**

   (a) (3 points) Let X be a random variable with mean μ and finite variance σ^2. *Show that for any distribution of X, the probability that X differs from its mean by more than 3 standard deviations is at most $\frac{1}{9}$.*

   (b) (7 points) Let $(X_i)_{i\in\mathbb{N}}$ be independent, identically distributed random variables with the same distribution as X. For all $n \in \mathbb{N}$, define $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ to be the sample mean of the first n elements of the sequence $(X_i)_{i\in\mathbb{N}}$. *Compute with a limit theorem the approximate minimum value of n such that*

$$P\left(\left|\bar{X}_n - \mu\right| \le \frac{\sigma}{3}\right) \ge 0.9.$$

Solution

(a) Applying the generalized Chebyshev inequality with the function $g(y) := y^2$ for $y > 0$, we get

$$P\left(|X-\mu| \ge 3\sigma\right) \le \frac{\sigma^2}{(3\sigma)^2} = \frac{1}{9}.$$

(b) For all $n \in \mathbb{N}$, let us define $Z_n := \frac{\sqrt{n}}{\sigma}\left(\bar{X}_n - \mu\right)$. By the central limit theorem the distribution function of Z_n converges pointwise to the distribution function Φ of a standard normal random variable, as $n \to \infty$. Therefore, for large n we have

$$P\left(\left|\bar{X}_n - \mu\right| \le \frac{\sigma}{3}\right) = P\left(|Z_n| \le \frac{\sqrt{n}}{3}\right) \approx \Phi\left(\frac{\sqrt{n}}{3}\right) - \Phi\left(-\frac{\sqrt{n}}{3}\right) = 2\Phi\left(\frac{\sqrt{n}}{3}\right) - 1.$$

We want $2\Phi\left(\frac{\sqrt{n}}{3}\right) - 1 \ge 0.9$, i.e. $\Phi\left(\frac{\sqrt{n}}{3}\right) \ge 0.95$. With the table for the standard normal distribution we obtain $\frac{\sqrt{n}}{3} \ge 1.65$ and therefore $n \ge 24.5$. Since n has to be an integer, we get $n_{\min} = 25$.

8. (10 points) **Maximum Likelihood Estimation**

The Pareto distribution introduced in Exercise 4 is used a lot in insurance claim modelling. But to use it for simulation, the parameters need to be estimated. Recall that the density function of a Pareto distribution with parameters γ and m is given by

$$f(x) = \begin{cases} \frac{\gamma m^\gamma}{x^{\gamma+1}} & \text{if } x \geq m, \\ 0 & \text{else.} \end{cases}$$

We observe the reporting of n claims with values $x_1, x_2, \ldots, x_n$.

(a) (4 points) *Define and compute the likelihood function and the log-likelihood function.*

(b) (4 points) *For a fixed m, give the maximum likelihood estimate for the parameter γ.*

(c) (2 points) *For a fixed γ, give the maximum likelihood estimate for the parameter m.*

Solution

(a) The likelihood function is defined as

$$L(\gamma, m | x_1, x_2, \ldots, x_n) := \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \frac{\gamma m^\gamma}{x_i^{\gamma+1}} \mathbb{1}_{\{x_i \geq m\}}.$$

The log-likelihood function is defined as

$$\begin{aligned} l(\gamma, m | x_1, x_2, \ldots, x_n) &:= \ln\left[L(\gamma, m | x_1, x_2, \ldots, x_n)\right] \\ &= \begin{cases} n\ln(\gamma) + n\gamma\ln(m) - (\gamma+1)\sum_{i=1}^n \ln(x_i) & \text{if } \min\limits_{1\leq i\leq n} x_i \geq m, \\ -\infty \text{ otherwise.} \end{cases} \end{aligned}$$

(b) For any given $m \leq \min_{1\leq i\leq n} x_i$, let us differentiate l with respect to γ:

$$\begin{aligned} \frac{\partial l}{\partial \gamma}(\gamma, m | x_1, x_2, \ldots, x_n) &= \frac{n}{\gamma} + n\ln(m) - \sum_{i=1}^n \ln(x_i), \\ \frac{\partial^2 l}{\partial \gamma^2}(\gamma, m | x_1, x_2, \ldots, x_n) &= -\frac{n}{\gamma^2} < 0. \end{aligned}$$

Then the equation $\frac{\partial l}{\partial \gamma}(\gamma, m | x_1, x_2, \ldots, x_n) = 0$ gives a global minimizer of the log-likelihood in γ. This gives the estimate

$$\hat{\gamma} = \frac{1}{\frac{1}{n}\sum_{i=1}^n \ln(\frac{x_i}{m})}.$$

(c) The log-likelihood function is increasing in the parameter m up to $\min_{1\leq i\leq n} x_i$. The maximum likelihood estimate for m is then

$$\hat{m} = \min_{1\leq i\leq n} x_i.$$

9. (10 points) **Posterior Distribution**

Let $X_1, \ldots, X_n$ be independent, identically distributed random variables sampled from the Geometric distribution with parameter $0 < \theta < 1$, which is unknown. Assume that we have a prior distribution for θ, which is the Beta distribution with parameters $\alpha > 0$ and $\beta > 0$. Recall that the geometric distribution is such that

$$P(X = x|\theta = \vartheta) = \begin{cases} \vartheta(1-\vartheta)^{x-1} & \text{if } x \in \mathbb{N}\backslash\{0\}, \\ 0 & \text{else,} \end{cases}$$

and that the Beta distribution with parameters α and β has the density function

$$w(\vartheta|\alpha,\beta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\vartheta^{\alpha-1}(1-\vartheta)^{\beta-1} & \text{if } 0 \leq \vartheta \leq 1, \\ 0 & \text{else.} \end{cases}$$

In an experiment we observe $X_1 = x_1, \ldots, X_n = x_n$ for some $x_1, \ldots, x_n \in \mathbb{N}\backslash\{0\}$. *Show that the posterior distribution of θ given that $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ is the Beta distribution with parameters $n + \alpha$ and $\beta + \sum_{i=1}^n x_i - n$.*

Solution

First we calculate the joint probability density function of $X_1, \ldots, X_n, \theta$ with respect to $\delta_{\mathbb{N}}^n \otimes \Lambda_{|\mathbb{R}_+}$:

$$\begin{aligned} f_{X_1,\ldots,X_n,\theta}(x_1,\ldots,x_n,\vartheta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\vartheta^{\alpha-1}(1-\vartheta)^{\beta-1}\vartheta^n(1-\vartheta)^{y-n} \cdot \mathbf{1}_{\vartheta\in(0,1)} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\vartheta^{n+\alpha-1}(1-\vartheta)^{\beta+y-n-1} \cdot \mathbf{1}_{\vartheta\in(0,1)}, \end{aligned}$$

where $y = x_1 + \cdots + x_n$. The marginal law of $(X_1, \ldots, X_n)$ is obtained by integrating the joint distribution with respect to ϑ:

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\vartheta^{n+\alpha-1}(1-\vartheta)^{\beta+y-n-1}d\vartheta \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(n+\alpha)\Gamma(\beta+y-n)}{\Gamma(\alpha+\beta+y)} \end{aligned}$$

So, conditionally given $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$, θ has the probability density function

$$\begin{aligned} f_{\theta|(X_1,\ldots,X_n)=(x_1,\ldots,x_n)}(\vartheta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\vartheta^{\alpha+y-1}(1-\vartheta)^{\beta+n-y-1}}{f_{X_1,\cdots,X_n}(x_1,\ldots,x_n)} \cdot \mathbf{1}_{\vartheta\in(0,1)} \\ &= \frac{\Gamma(\alpha+\beta+y)}{\Gamma(n+\alpha)\Gamma(\beta+y-n)}\vartheta^{\alpha+n-1}(1-\vartheta)^{\beta+y-n-1} \cdot \mathbf{1}_{\vartheta\in(0,1)}. \end{aligned}$$

We see that the posterior distribution of θ given that $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ is the Beta distribution with parameters $\alpha + n$ and $\beta + y - n$.

10. (10 points) **Hypothesis Test**

The printers of the student computer room are submitted to a heavy workload. Let $X_1, \ldots, X_9$ model the lifetime (in months) for 9 of the printers. We assume that $X_1, \ldots, X_9$ are independent and normally distributed with unknown mean μ and variance equal to 4 months. Moreover, assume that we observe the following number of months of service for the 9 printers:

$$25,\ 26,\ 22,\ 29,\ 23,\ 20,\ 30,\ 28,\ 31.$$

We want to test the hypothesis

$$H_0: \ \mu = 25 \quad \text{against} \quad H_1: \ \mu \neq 25.$$

Build a statistical test to test the above hypothesis at the level $\alpha = 0.05$. In particular, give the test statistic you use, its distribution under H_0, the rejection region and decide whether we can reject the null hypothesis in this particular situation.

Solution Under H_0, we have $X_i \sim \mathcal{N}(25, 4)$ for all $i \in \{1, \ldots, 9\}$. We use the test statistic

$$T := T(X_1, \ldots, X_9) := \sqrt{9}\frac{\frac{1}{9}\sum_{i=1}^{9}(X_i - 25)}{2} = \frac{1}{6}\sum_{i=1}^{9}(X_i - 25),$$

which is standard normally distributed under H_0. For our sample we get

$$T(25, 26, 22, 29, 23, 20, 30, 28, 31) = \frac{1}{6} \cdot 9 = \frac{3}{2}.$$

We want to design a two-sided test, the rejection region should be symmetric around 25. We want to find c_α such that

$$P\left(|T| \geq c_\alpha\right) \leq \alpha.$$

We have

$$\begin{aligned} P\left(|T| \geq c_\alpha\right) &= P\left(T \leq -c_\alpha\right) + P\left(T \geq c_\alpha\right) \\ &= 2P\left(T \leq -c_\alpha\right) \\ &= 2[1 - P\left(T \leq c_\alpha\right)]. \end{aligned}$$

Therefore $P\left(|T| \geq c_\alpha\right) \leq \alpha$ is equivalent to $P\left(T \leq c_\alpha\right) \geq 0.975$, that is to say (from the table) $c_\alpha = 1.96$.

The rejection region is then $(-\infty, -1.96] \cup [1.96, \infty)$. The value of the test statistic for the given set of values is not in the rejection set, so we do not reject the null hypothesis.

Standard normal (cumulative) distribution function.

$$P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y, \quad \text{for } x \geqslant 0$$

	0	1	2	3	4	5	6	7	8	9
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6408	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.998650	.998694	.998736	.998777	.998817	.998856	.998893	.998930	.998965	.998999
3.1	.999032	.999065	.999096	.999126	.999155	.999184	.999211	.999238	.999264	.999289
3.2	.999313	.999336	.999359	.999381	.999402	.999423	.999443	.999462	.999481	.999499
3.3	.999517	.999534	.999550	.999566	.999581	.999596	.999610	.999624	.999638	.999651
3.4	.999663	.999675	.999687	.999698	.999709	.999720	.999730	.999740	.999749	.999758
3.5	.999767	.999776	.999784	.999792	.999800	.999807	.999815	.999822	.999828	.999835
3.6	.999841	.999847	.999853	.999858	.999864	.999869	.999874	.999879	.999883	.999888
3.7	.999892	.999896	.999900	.999904	.999908	.999912	.999915	.999918	.999922	.999925
3.8	.999928	.999931	.999933	.999936	.999938	.999941	.999943	.999946	.999948	.999950
3.9	.999952	.999954	.999956	.999958	.999959	.999961	.999963	.999964	.999966	.999967