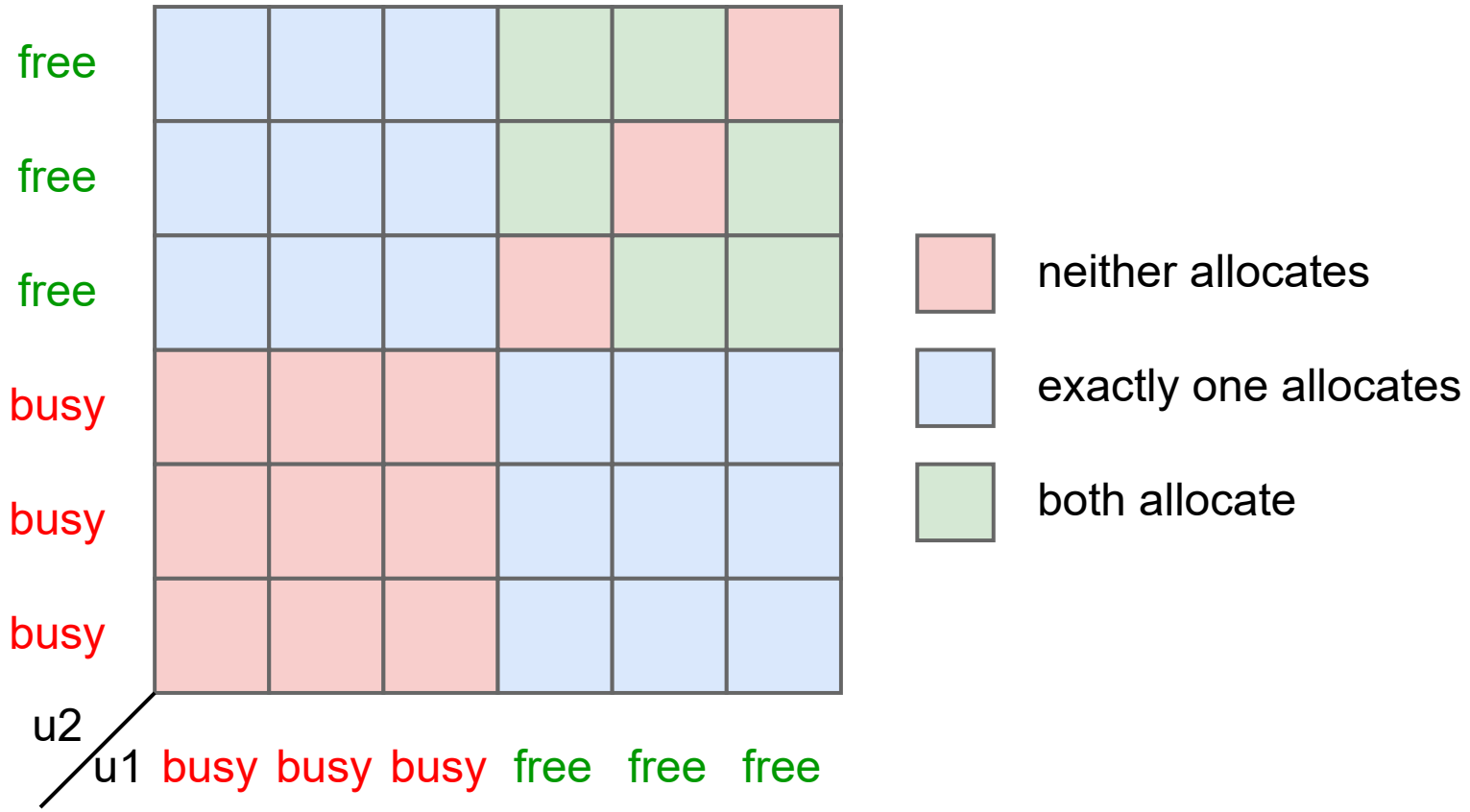
MSP 2024 - Tutorial 1

Exercise 1. A service provider manages a frequency bandwidth divided into 6 channels that can be used for communication. A user that attempts to establish a new connection randomly and uniformly picks the channel; if the channel is free, it is allocated to the user, otherwise the user repeats the attempt. Assume that one attempt takes roughly 1ms.
Assume that currently 3 out of 6 channels are busy and that two users simultaneously attempt to allocate free channels. If both users randomly pick the same free channel, they both restart the procedure.

1. Construct the Markov chain that models such a system.
2. Compute the probability that it takes up to 3ms to find free channels for both users.
3. Compute the probability that it takes exactly 3ms to find free channels for both users.
4. Compute the expected time required to find free channels for both users.

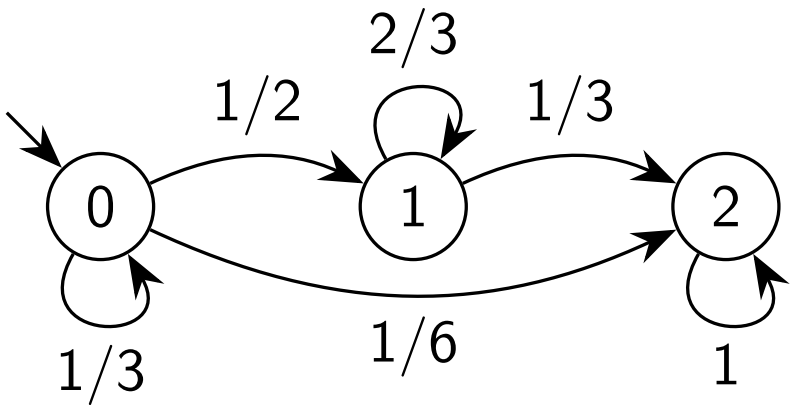
Solution

1. Decision diagram for state 0:



Solution

1. State of the DTMC encodes how many users have the channel allocated.



Solution

2. $P(2 \text{ channels are allocated up to time } 3) = P(X(3) = 2) = \boldsymbol{t}_3(2)$:

$$\boldsymbol{t}_0(0,1,2) = (1,0,0)$$

$$\boldsymbol{t}_1 = \boldsymbol{t}_0 \cdot \boldsymbol{P} = \left(\frac{1}{3}, \frac{1}{2}, \frac{1}{6}\right)$$

$$\boldsymbol{t}_2 = \boldsymbol{t}_1 \cdot \boldsymbol{P} = \left(\frac{1}{9}, \frac{1}{2}, \frac{7}{18}\right)$$

$$\boldsymbol{t}_3 = \boldsymbol{t}_2 \cdot \boldsymbol{P} = \left(\frac{1}{27}, \frac{7}{18}, \frac{31}{54}\right)$$

$\Rightarrow \boldsymbol{t}_3(2) = \frac{31}{54}$

Solution

3

$$P(\text{two channels are allocated exactly at time } 3) = P(X(2) \neq 2, X(3) = 2)$$
$$= P(X(2) = 0, X(3) = 2) + P(X(2) = 1, X(3) = 2)$$
$$= P(X(2) = 0) \cdot P(X(3) = 2 \mid X(2) = 0)$$
$$+ P(X(2) = 1) \cdot P(X(3) = 2 \mid X(2) = 1)$$
$$= \boldsymbol{t}_2(0) \cdot \boldsymbol{P}(0, 2) + \boldsymbol{t}_2(1) \cdot \boldsymbol{P}(1, 2) = \frac{1}{9} \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{1}{3} = \frac{5}{27}$$

Coincidentally, since state 2 is absorbing, $P(X(2) \neq 2, X(3) = 2)$ is also equal to $\boldsymbol{t}_3(2) - \boldsymbol{t}_2(2)$.

Solution

4. Let $T = \{2\}$ be the set of target states and let $\boldsymbol{e}(s)$ denote the expected number of steps (milliseconds) to reach T from s.
$P(0 \to T) = 1$ since the DTMC contains only one (reachable) BSCC and this BSCC contains state 2, thus, $\boldsymbol{e}(0)$ is defined and is obtained by solving the following system:

$$\boldsymbol{e}(0) = 1 + \frac{1}{3}\boldsymbol{e}(0) + \frac{1}{2}\boldsymbol{e}(1) + \frac{1}{6}\boldsymbol{e}(2)$$
$$\boldsymbol{e}(1) = 1 + \frac{2}{3}\boldsymbol{e}(1) + \frac{1}{3}\boldsymbol{e}(2)$$
$$\boldsymbol{e}(2) = 0$$

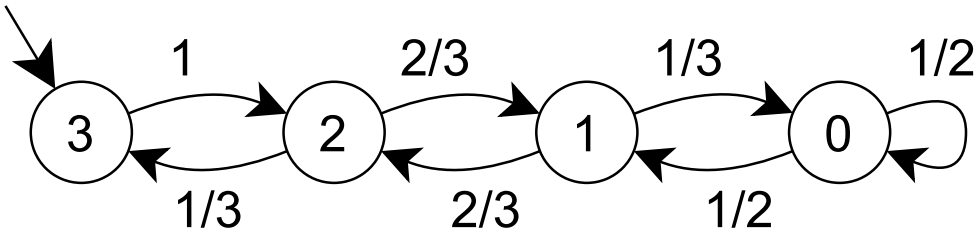
$\Rightarrow \boldsymbol{e}(0) = 3.75$ ms

Exercise 2. Assume a cluster server with three machines. Initially, all machines are on. Every hour, a technician randomly and uniformly chooses a machine to work with: a running machine is shut down for maintenance and an offline machine is turned back on. However, when all machines are off, the probability that the selected machine is restarted is reduced to 50%. The cluster as a whole is considered online if at least one machine is running.

1. Construct the Markov chain that models such a system.
2. Compute the probability that after 4 hours the server is online.
3. Assume that initially all machines are off and the technician is 2 hours late on their first day. Modify the model and recompute the probability from 2.

Solution

1. State of the DTMC encodes the number of running machines.

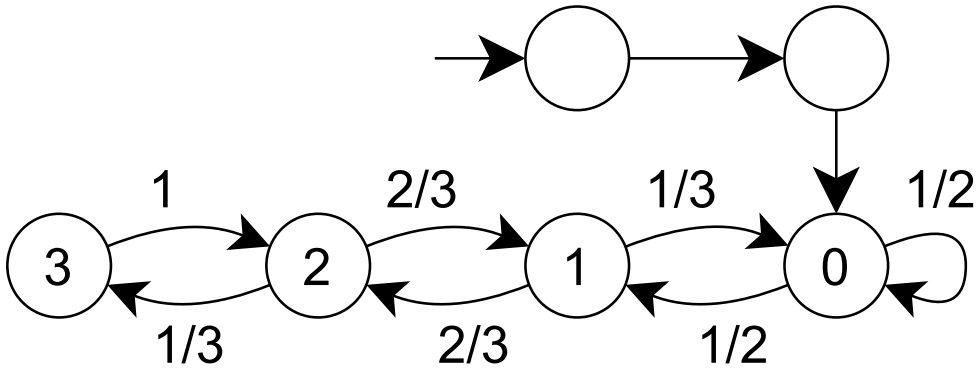


2.

$$\boldsymbol{t}_0(3,2,1,0) = (1,0,0,0)$$
$$\boldsymbol{t}_1 = \boldsymbol{t}_0 \cdot \boldsymbol{P} = (0,1,0,0)$$
$$\boldsymbol{t}_2 = \boldsymbol{t}_1 \cdot \boldsymbol{P} = \left(\frac{1}{3}, 0, \frac{2}{3}, 0\right)$$
$$\boldsymbol{t}_3 = \boldsymbol{t}_2 \cdot \boldsymbol{P} = \left(0, \frac{7}{9}, 0, \frac{2}{9}\right)$$
$$\boldsymbol{t}_4 = \boldsymbol{t}_3 \cdot \boldsymbol{P} = \left(\frac{7}{27}, 0, \frac{17}{27}, \frac{1}{9}\right)$$
$$P(\text{online after 4 hours}) = \boldsymbol{t}_4(1) + \boldsymbol{t}_4(2) + \boldsymbol{t}_4(3) = \frac{8}{9}$$

Solution

3. The Markov chain below models the late arrival



$$\boldsymbol{t}_2(3,2,1,0) = (0,0,0,1)$$

$$\boldsymbol{t}_3 = \boldsymbol{t}_2 \cdot \boldsymbol{P} = \left(0,0,\frac{1}{2},\frac{1}{2}\right)$$

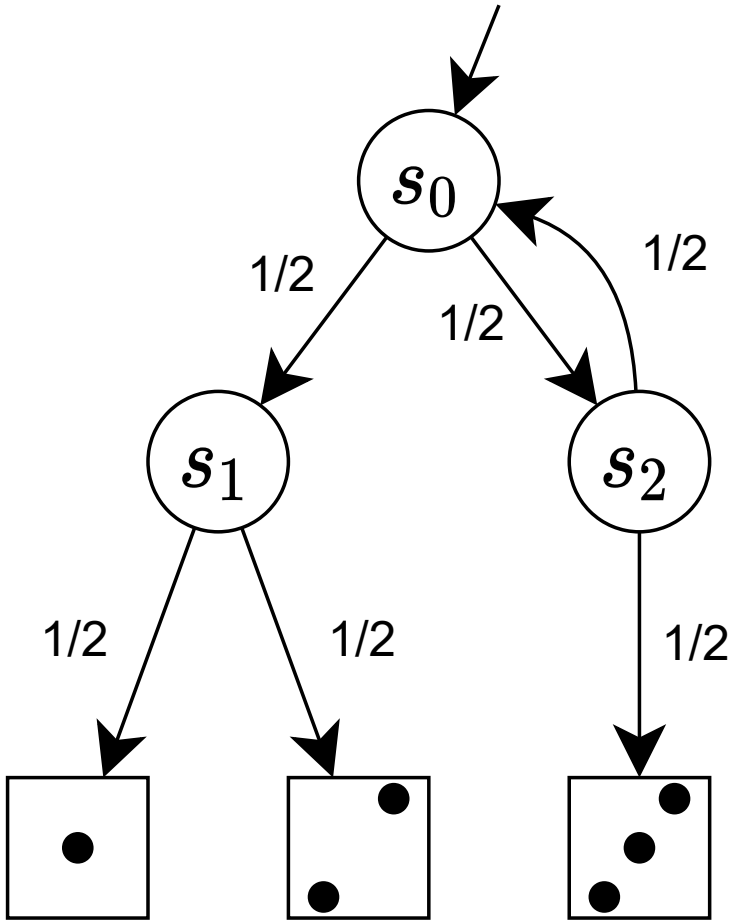
$$\boldsymbol{t}_4 = \boldsymbol{t}_3 \cdot \boldsymbol{P} = \left(0,\frac{1}{3},\frac{1}{4},\frac{5}{12}\right)$$

$$P(\text{online after } 4 \text{ hours}) = \boldsymbol{t}_4(1) + \boldsymbol{t}_4(2) + \boldsymbol{t}_4(3) = \frac{7}{12}$$

Exercise 3. Design a protocol that uses a fair coin to simulate a toss of a 3-sided dice.

1. Verify correctness of your protocol.
2. Analyze efficiency of your protocol.
3. Compute the probability that the protocol does not terminate within 4 steps.

Solution



Solution

1. Correctness: $P(s_0 \to \{\boxdot\}) = P(s_0 \to \{\boxdot\}) = P(s_0 \to \{\boxdot\}) = 1/3$?

   - Let $T = \{\boxdot\}$, $\boldsymbol{x}(s) \coloneqq P(s \to T)$. $S_0 = \{\boxdot, \boxdot\}$

$$\boldsymbol{x}(\boxdot) = 1$$
$$\boldsymbol{x}(\boxdot) = \boldsymbol{x}(\boxdot) = 0$$
$$\boldsymbol{x}(s_0) = \frac{1}{2} \cdot \boldsymbol{x}(s_1) + \frac{1}{2} \cdot \boldsymbol{x}(s_2)$$
$$\boldsymbol{x}(s_1) = \frac{1}{2} \cdot \boldsymbol{x}(\boxdot) + \frac{1}{2} \cdot \boldsymbol{x}(\boxdot)$$
$$\boldsymbol{x}(s_2) = \frac{1}{2} \cdot \boldsymbol{x}(s_0) + \frac{1}{2} \cdot \boldsymbol{x}(\boxdot)$$

   $\Rightarrow P(s_0 \to \{\boxdot\}) = \boldsymbol{x}(s_0) = 1/3$

   - By symmetry, $P(s_0 \to \{\boxdot\}) = 1/3$ as well.

   - Since $\boxdot, \boxdot$ and $\boxdot$ are the only BSCCs, then $P(s_0 \to \{\boxdot\}) = 1 - P(s_0 \to \{\boxdot\}) - P(s_0 \to \{\boxdot\}) = 1/3$

Solution

2. Efficiency = expected number of tosses to execute the protocol. Let $T = \{\boxdot, \boxdot, \boxdot\}$ and let $\boldsymbol{e}(s)$ denote the expected number of steps to reach T from s. Then:

$$\boldsymbol{e}(\boxdot) = \boldsymbol{e}(\boxdot) = \boldsymbol{e}(\boxdot) = 0$$

$$\boldsymbol{e}(s_0) = 1 + \frac{1}{2} \cdot \boldsymbol{e}(s_1) + \frac{1}{2} \cdot \boldsymbol{e}(s_2)$$

$$\boldsymbol{e}(s_1) = 1 + \frac{1}{2} \cdot \boldsymbol{e}(\boxdot) + \frac{1}{2} \cdot \boldsymbol{e}(\boxdot)$$

$$\boldsymbol{e}(s_2) = 1 + \frac{1}{2} \cdot \boldsymbol{e}(s_0) + \frac{1}{2} \cdot \boldsymbol{e}(\boxdot)$$

$\Rightarrow \boldsymbol{e}(s_0) = 8/3$

Solution

3

$$\boldsymbol{t}_0(s_0, s_1, s_2, \boxdot, \boxdot, \boxdot) = (1, 0, 0, 0, 0, 0)$$

$$\boldsymbol{t}_1 = \boldsymbol{t}_0 \cdot \boldsymbol{P} = \left(0, \frac{1}{2}, \frac{1}{2}, 0, 0, 0\right)$$

$$\boldsymbol{t}_2 = \boldsymbol{t}_1 \cdot \boldsymbol{P} = \left(\frac{1}{4}, 0, 0, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$$

$$\boldsymbol{t}_3 = \boldsymbol{t}_2 \cdot \boldsymbol{P} = \left(0, \frac{1}{8}, \frac{1}{8}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$$

$$\boldsymbol{t}_4 = \boldsymbol{t}_3 \cdot \boldsymbol{P} = \left(\frac{1}{16}, 0, 0, \frac{5}{16}, \frac{5}{16}, \frac{5}{16}\right)$$

$$P(\text{protocol is running after } 4 \text{ tosses}) = \boldsymbol{t}_4(s_0) + \boldsymbol{t}_4(s_1) + \boldsymbol{t}_4(s_2) = 1/16$$

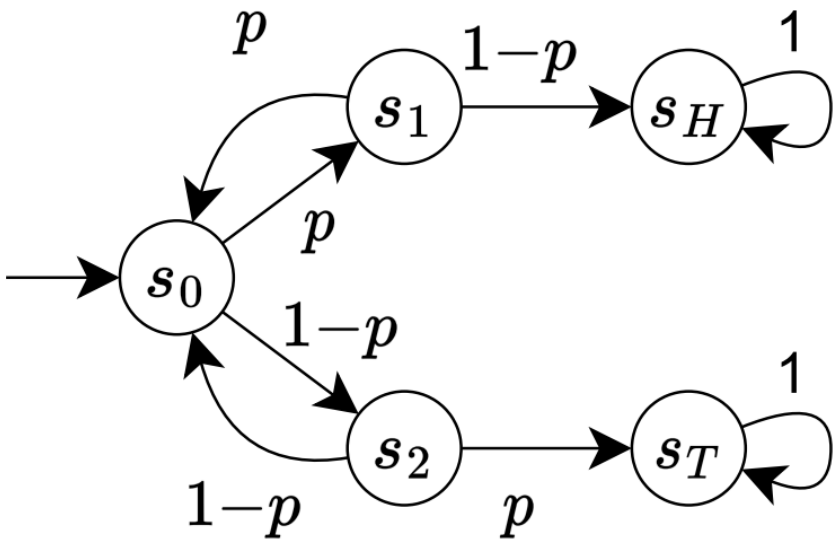
Exercise 4 (homework). Design a protocol that uses a fair coin to simulate a toss of a 5-sided dice. Ensure that your Markov chain does not have unnecessary states/transitions.

1. Verify correctness of your protocol.
2. Analyze efficiency of your protocol.
3. Compute the probability that the protocol produces result 1 or 5 within first 5 steps.

Exercise 5. Design a protocol that uses an unfair coin with bias p to simulate a toss of a fair coin.

1. Verify correctness of your protocol.
2. Compute expected number $t(p)$ of tosses until protocol termination. Plot $t(p)$ and interpret the result.

Solution



Solution

2. Correctness: $P(s_0 \to \{s_H\}) = P(s_0 \to \{s_T\}) = 1/2$.

   Let $T = \{s_H\}$, $\boldsymbol{x}(s) \coloneqq P(s \to T)$. $S_0 = \{s_T\}$.

$$
\begin{aligned}
\boldsymbol{x}(s_H) &= 1 \\
\boldsymbol{x}(s_T) &= 0 \\
\boldsymbol{x}(s_0) &= p \cdot \boldsymbol{x}(s_1) + (1-p) \cdot \boldsymbol{x}(s_2) \\
\boldsymbol{x}(s_1) &= p \cdot \boldsymbol{x}(s_0) + (1-p) \cdot \boldsymbol{x}(s_H) \\
\boldsymbol{x}(s_2) &= p \cdot \boldsymbol{x}(s_T) + (1-p) \cdot \boldsymbol{x}(s_0)
\end{aligned}
$$

   $\Rightarrow P(s_0 \to \{s_H\}) = \boldsymbol{x}(s_0) = 1/2$. Since s_H and s_T are the only BSCCs, it must hold that $P(s_0 \to \{s_T\}) = 1 - P(s_0 \to \{s_H\}) = 1/2$.

Solution

2. Let $T = \{s_H, s_T\}$ and let $\boldsymbol{e}(s)$ denote the expected number of steps to reach T from s. Then:

$$\begin{aligned}
\boldsymbol{e}(s_H) &= \boldsymbol{e}(s_T) = 0 \\
\boldsymbol{e}(s_0) &= 1 + p \cdot \boldsymbol{e}(s_1) + (1-p) \cdot \boldsymbol{e}(s_2) \\
\boldsymbol{e}(s_1) &= 1 + p \cdot \boldsymbol{e}(s_0) + (1-p) \cdot \boldsymbol{e}(s_H) \\
\boldsymbol{e}(s_2) &= 1 + p \cdot \boldsymbol{e}(s_T) + (1-p) \cdot \boldsymbol{e}(s_0)
\end{aligned}$$

$$\Rightarrow t(p) = \boldsymbol{e}(s_0) = \frac{1}{p(1-p)}$$

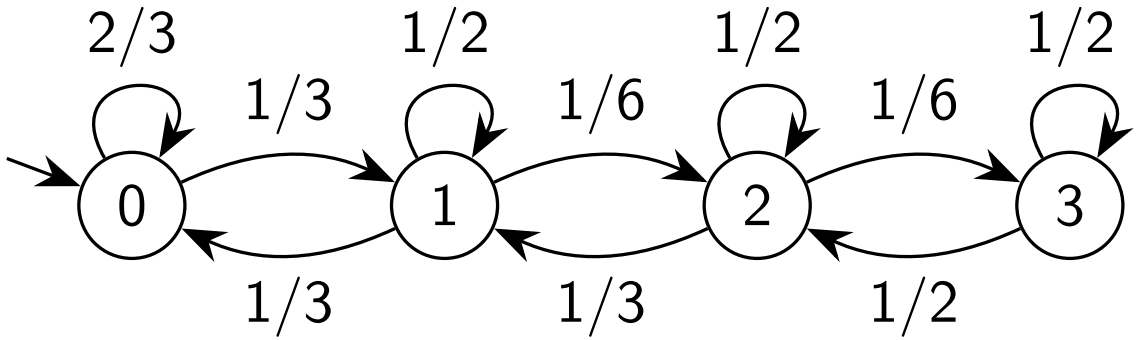
Exercise 6 (2022 midterm test). Consider a router that processes and forwards incoming packets. Every millisecond, the following *independent* events may occur:

- With probability 1/3, there is an incoming packet that is stored in the buffer. The buffer can store up to 3 packets and is initially empty. If the buffer is full at the beginning of the millisecond, the incoming packet is lost.
- If the buffer is not empty at the beginning of the millisecond, with probability 1/2 the router processes and forwards the first packet in the buffer.

1. Construct the Markov chain that models such a system.
2. Compute the probability that after 3ms the buffer contains at least 2 packets.
3. Compute the expected time until the buffer becomes full.

Solution

1. State of the DTMC encodes the number of packets in the buffer.



Solution

2

$$\boldsymbol{t}_0(0,1,2,3) = (1,0,0,0)$$

$$\boldsymbol{t}_1 = \left(\frac{2}{3}, \frac{1}{3}, 0, 0\right)$$

$$\boldsymbol{t}_2 = \left(\frac{5}{9}, \frac{7}{18}, \frac{1}{18}, 0\right)$$

$$\boldsymbol{t}_3 = \left(\frac{1}{2}, \frac{43}{108}, \frac{10}{108}, \frac{1}{108}\right)$$

$$P(2+ \text{ packets after } 3\text{ms}) = \boldsymbol{t}_3(2) + \boldsymbol{t}_3(3) = \frac{11}{108}$$

Solution

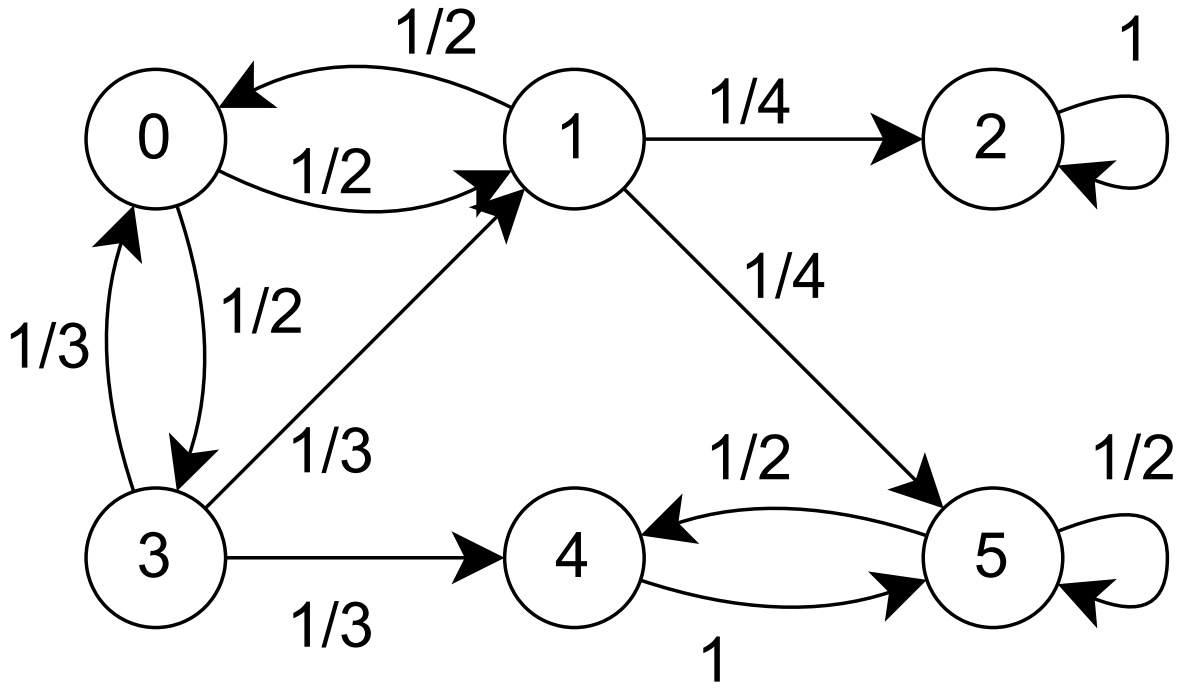
3. Expected time until the buffer is full = expected number of milliseconds (steps) to reach state 3. Let $T = \{3\}$ be the set of target states and let $\boldsymbol{e}(s)$ denote the expected number of steps to reach T from s.
$P(0 \to T) = 1$ since the DTMC contains no transient states, thus, $\boldsymbol{e}(0)$ is defined and is obtained by solving the following system:

$$\begin{aligned}
\boldsymbol{e}(0) &= 1 + \frac{2}{3} \cdot \boldsymbol{e}(0) + \frac{1}{3} \cdot \boldsymbol{e}(1) \\
\boldsymbol{e}(1) &= 1 + \frac{1}{3} \cdot \boldsymbol{e}(0) + \frac{1}{2} \cdot \boldsymbol{e}(1) + \frac{1}{6} \cdot \boldsymbol{e}(2) \\
\boldsymbol{e}(2) &= 1 + \frac{1}{3} \cdot \boldsymbol{e}(1) + \frac{1}{2} \cdot \boldsymbol{e}(2) + \frac{1}{6} \cdot \boldsymbol{e}(3) \\
\boldsymbol{e}(3) &= 0
\end{aligned}$$

$\Rightarrow \boldsymbol{e}(0) = 45$ ms

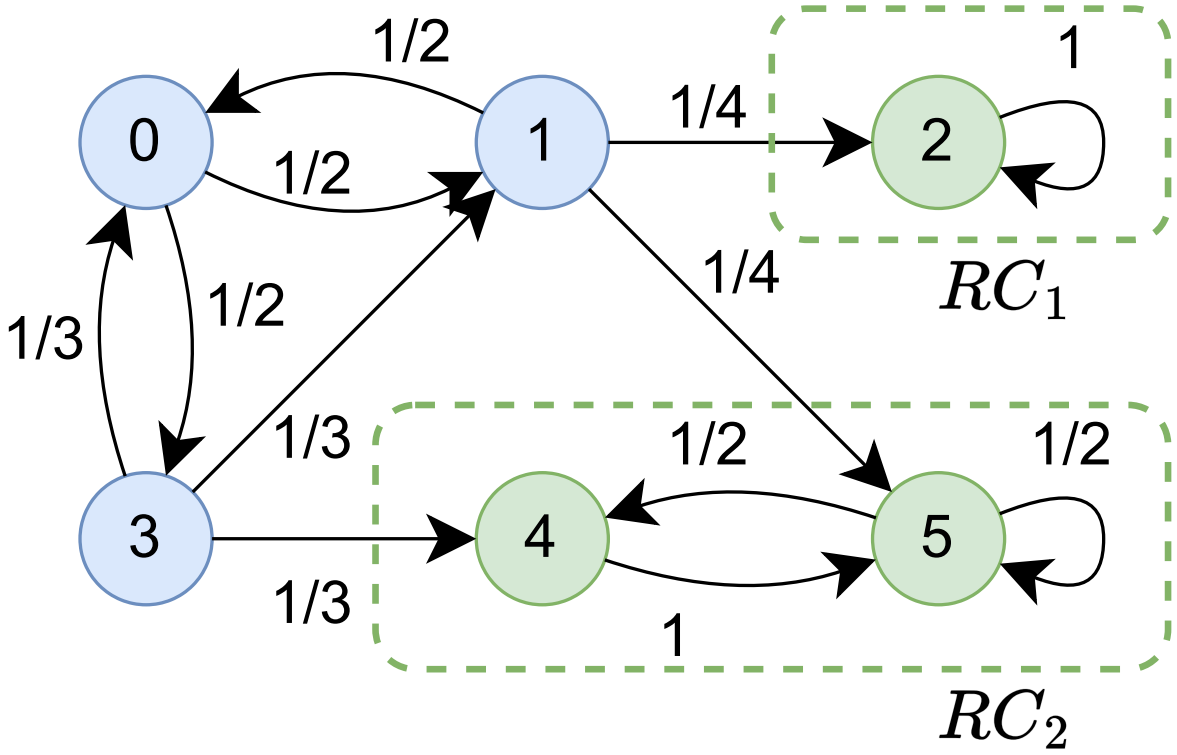
MSP 2024 - Tutorial 2

Exercise 1. Assume DTMC below.



1. Identify recurrent classes and transient states of this DTMC.
2. Compute the limiting distribution of this DTMC given that initially the chain starts in state:
   a) 2, b) 4, c) 5, d) 0, e) 1

Exercise 1. Assume DTMC below.



1. Identify recurrent classes and transient states of this DTMC.
2. Compute the limiting distribution of this DTMC given that initially the chain starts in state:
   a) 2, b) 4, c) 5, d) 0, e) 1

Solution

1. recurrent classes: $RC_1 = \{2\},\ RC_2 = \{4, 5\}$; transient states: $\{0, 1, 3\}$.

Let $\boldsymbol{t}_\infty^s$ denote the limiting distribution when starting in state s.

2. a) $\boldsymbol{t}_\infty^2 = (0, 0, 1, 0, 0, 0) =: \boldsymbol{\pi}^{RC_1}$
   b,c) Computing steady-state for RC_2:

$$\pi(4) = \frac{1}{2} \cdot \pi(5)$$

$$\pi(5) = \pi(4) + \frac{1}{2} \cdot \pi(5)$$

$$\pi(4) + \pi(5) = 1$$

$$\boldsymbol{t}_\infty^4 = \boldsymbol{t}_\infty^5 = \left(0, 0, 0, 0, \tfrac{1}{3}, \tfrac{2}{3}\right) =: \boldsymbol{\pi}^{RC_2}$$

Solution

Reachability to RC_1: $T = RC_1,\ x(s) \coloneqq P(s \to T)\,.\ S_0 = \{4, 5\}$

$$\boldsymbol{x}(2) = 1, \quad \boldsymbol{x}(4) = \boldsymbol{x}(5) = 0$$

$$\boldsymbol{x}(0) = \frac{1}{2} \cdot \boldsymbol{x}(1) + \frac{1}{2} \cdot \boldsymbol{x}(3)$$

$$\boldsymbol{x}(1) = \frac{1}{2} \cdot \boldsymbol{x}(0) + \frac{1}{4} \cdot \boldsymbol{x}(2) + \frac{1}{4} \cdot \boldsymbol{x}(5)$$

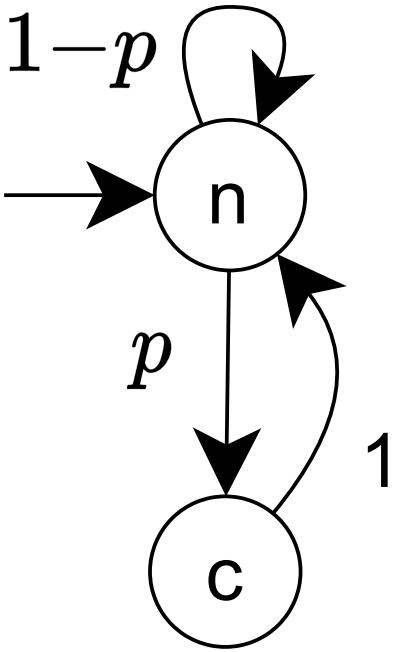
$$\boldsymbol{x}(3) = \frac{1}{3} \cdot \boldsymbol{x}(0) + \frac{1}{3} \cdot \boldsymbol{x}(1) + \frac{1}{3} \cdot \boldsymbol{x}(4)$$

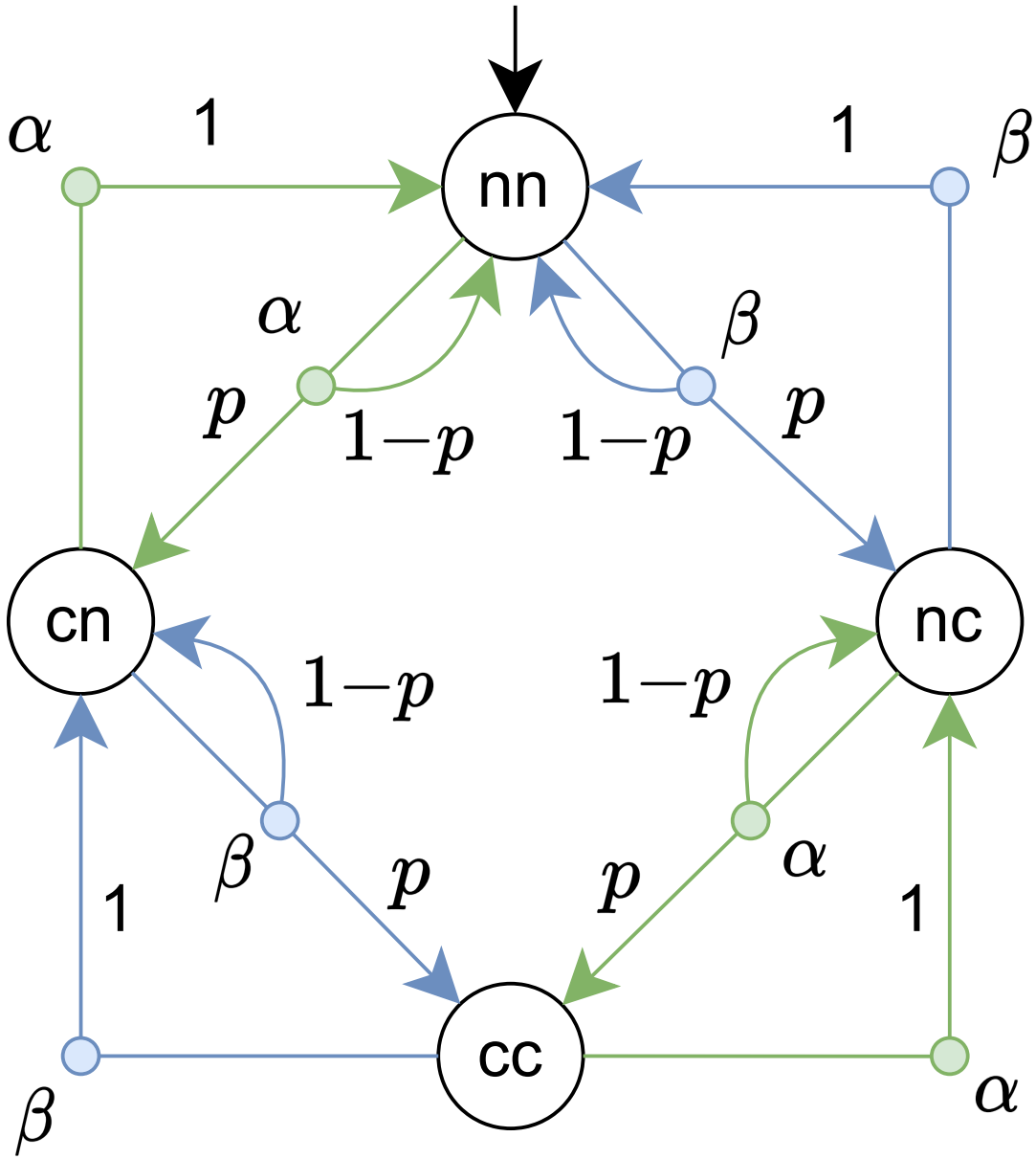
$\Rightarrow P(0 \to RC_1) = 1/3$ and $P(1 \to RC_1) = 5/12$. Then,
$P(0 \to RC_2) = 1 - P(0 \to RC_1) = 2/3$ and $P(1 \to RC_2) = 7/12$.
d)$\boldsymbol{t}^0_\infty = P(0 \to RC_1) \cdot \boldsymbol{\pi}^{RC_1} + P(0 \to RC_2) \cdot \boldsymbol{\pi}^{RC_2} = \left(0, 0, \frac{1}{3}, 0, \frac{2}{9}, \frac{4}{9}\right)$
e)$\boldsymbol{t}^1_\infty = P(1 \to RC_1) \cdot \boldsymbol{\pi}^{RC_1} + P(1 \to RC_2) \cdot \boldsymbol{\pi}^{RC_2} = \left(0, 0, \frac{5}{12}, 0, \frac{7}{36}, \frac{14}{36}\right)$

Exercise 2. During each time step, a process enters a critical section (CS) with probability p ($0 < p < 1$). Once in the CS, the process leaves it before the next time step.

1. Model such a process as a Markov chain.
2. Model a concurrent execution of two such processes on a single-core CPU as a Markov decision process. Non-deterministic actions correspond to the choice of the process to be run on a CPU during the next time step.

Solution


$1-p$
n
p
1
c


α
1
nn
1
β
α
β
p
$1-p$
$1-p$
p
cn
nc
$1-p$
$1-p$
β
p
p
α
1
1
cc
β
α

Exercise 2 (cont.). We say that the scheduler guarantees

- safety: the two processes can never both be in the CS
- liveness: with probability 1, every process can enter and leave CS infinitely often

3. Find scheduler that minimizes/maximizes probability of both processes being in the CS.
4. Show that no deterministic memoryless scheduler can guarantee safety and liveness.
5. Find a randomized memoryless scheduler that guarantees safety and liveness.
6. Find a deterministic non-memoryless scheduler that guarantees safety and liveness.

Solution

3. • minimizing scheduler: $\sigma_{\min}(nn) = \sigma_{\min}(cn) = \alpha$.

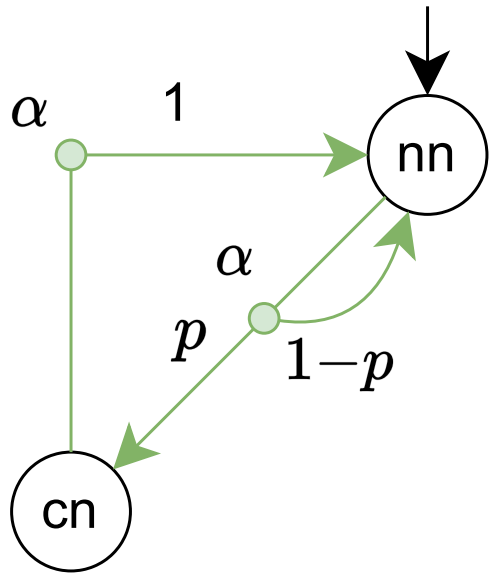


Figure: DTMC $\mathcal{M}_{\sigma_{\min}}$ induced by $\sigma_{\min}$

$$P_{\sigma_{\min}}(nn \to \{cc\}) = 0$$

Solution

❸ • maximizing scheduler: $\sigma_{\max}(nn) = \alpha,\ \sigma_{\max}(cn) = \beta,\ \sigma_{\max}(cc) = *$

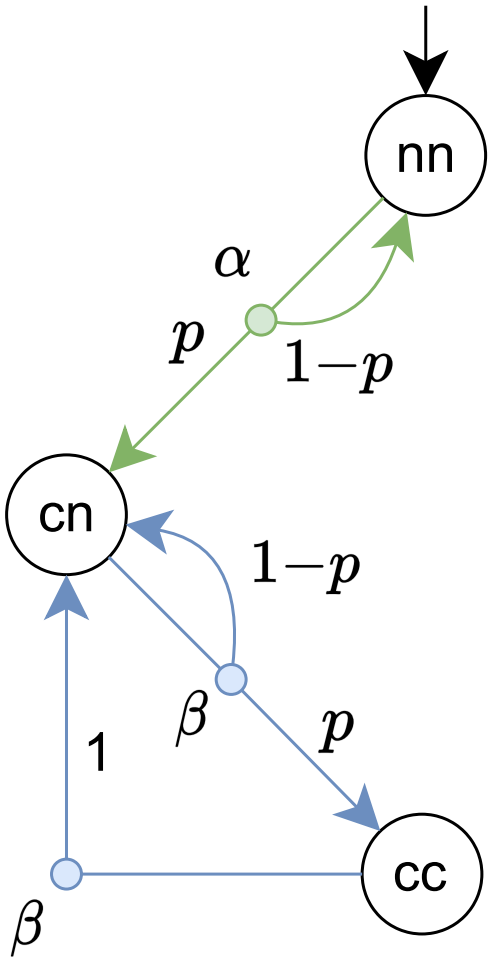
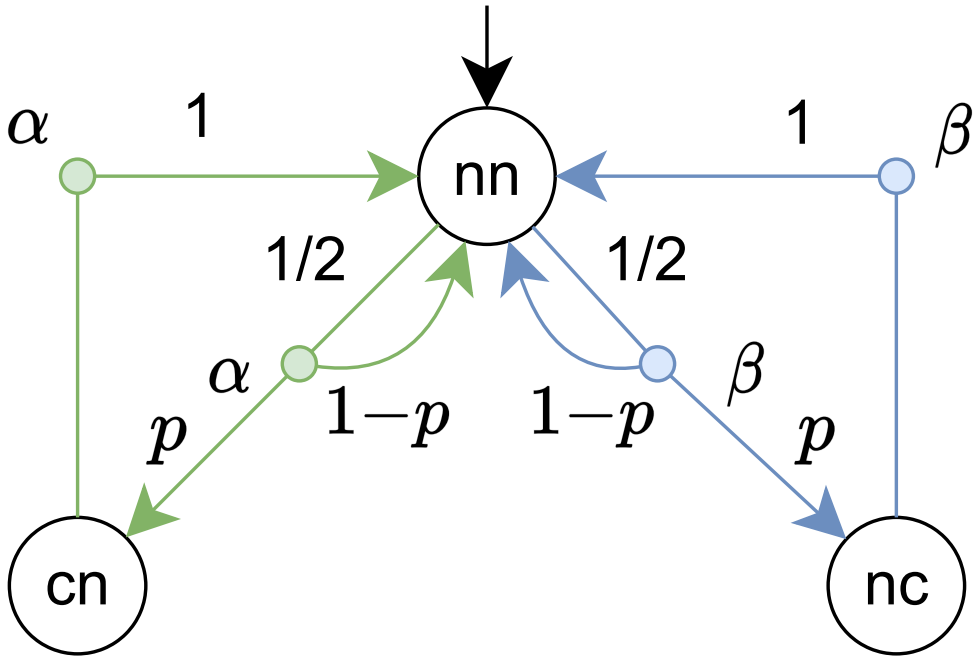


Figure: DTMC $\mathcal{M}_{\sigma_{\max}}$ induced by $\sigma_{\max}$

Solution

4. hint: show using scheduler enumeration
5. gist: when neither of the processes is in the CS, randomize which one to schedule

$$\sigma(nn) = 1/2 : \alpha + 1/2 : \beta,\ \sigma(cn) = 1 : \alpha,\ \sigma(nc) = 1 : \beta$$

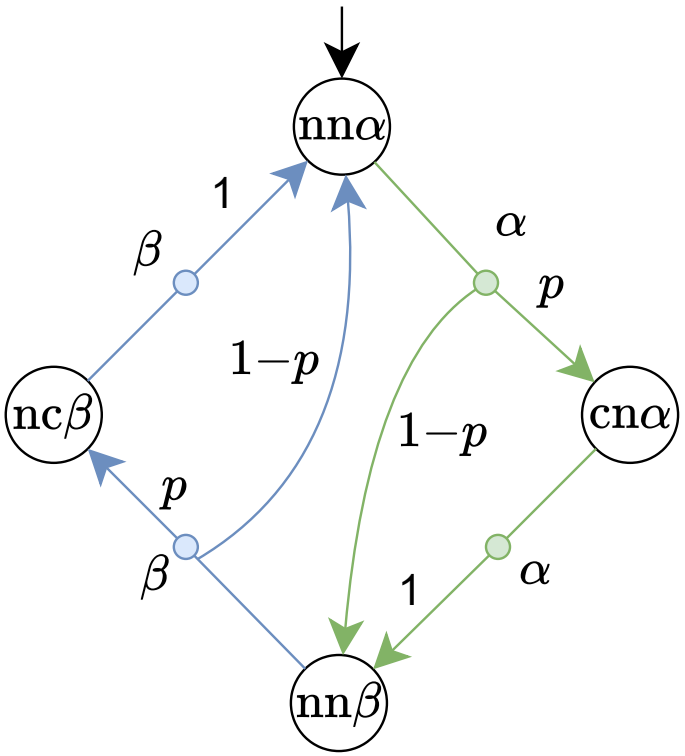


Solution

6. gist: alternate the scheduling of the two processes

$$\sigma(nn) = \sigma\left([S\,Act]^* cn\right) = \sigma\left(S\,[Act\,S]^* \beta nn\right) = \alpha$$
$$\sigma\left([S\,Act]^* nc\right) = \sigma\left(S\,[Act\,S]^* \alpha nn\right) = \beta$$

to construct the induced DTMC, introduce additional variable (memory) encoding "*which process has the CS priority*"

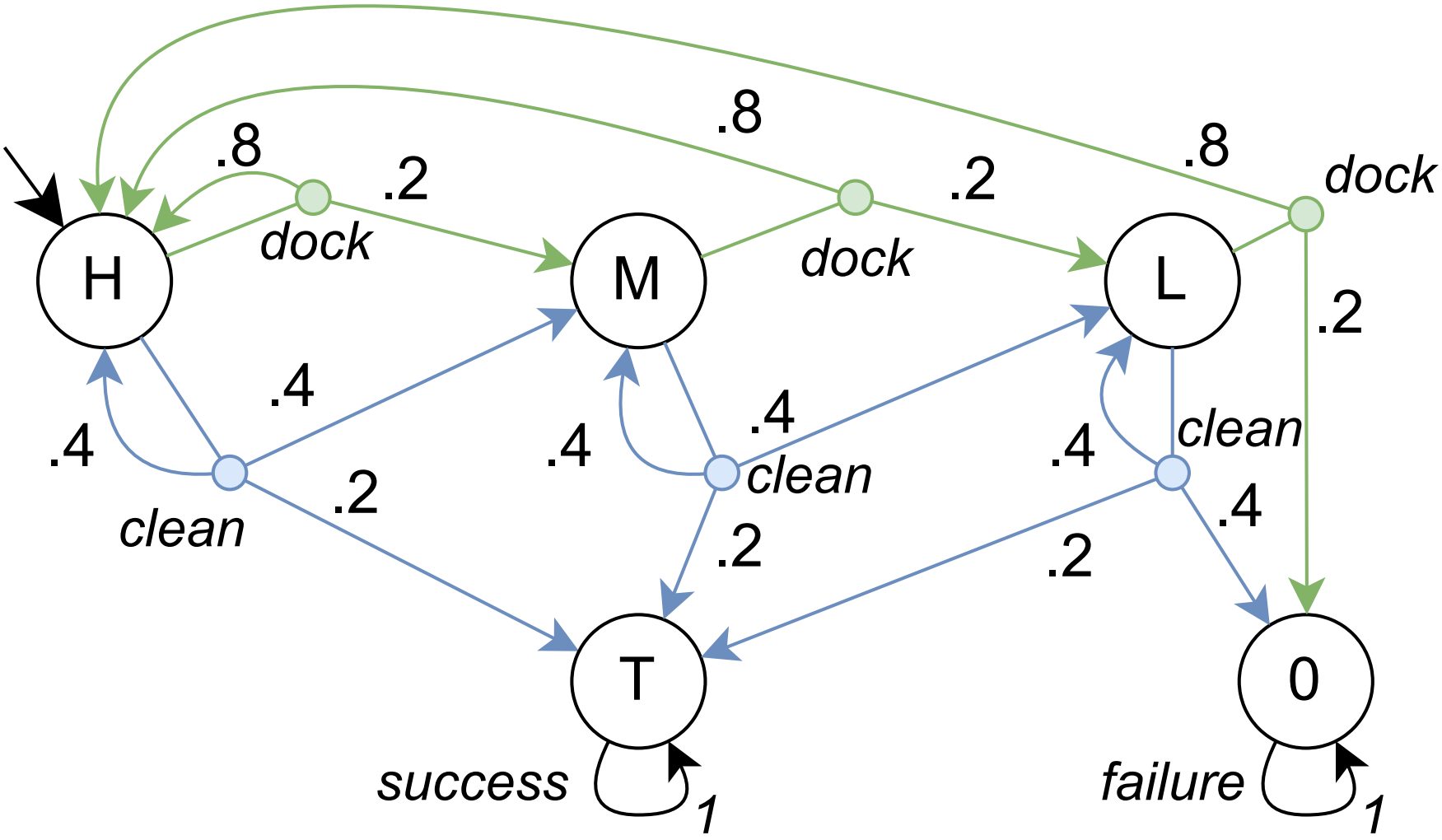


Exercise 3. Assume a cleaning robot has four distinct battery levels: high, medium, low, and zero. During each step, the robot can decide whether to resume cleaning or return to the charging dock.

- When the robot decides to resume cleaning, it will successfully finish the task with probability $p_{clean} = 0.2$. With each cleaning attempt, regardless of the success, the battery will be either drained by one level (with probability $p_{drain} = 0.5$) or remain the same. When the battery level drops to zero, the robot is cannot continue.
- If the robot decides to charge, it will succeed in finding the dock with probability $p_{dock} = 0.8$, fully charging its battery. Otherwise, the robot gets lost and loses one battery level.

1. Find a strategy that maximizes the robot's chances of cleaning the room before running out of battery.
    - Use all three techniques presented in the lecture (scheduler enumeration, LP, value iteration)

Solution


H
M
L
T
0
dock
dock
dock
.8
.8
.8
.2
.2
.2
.4
.4
.4
.4
.4
.4
clean
clean
clean
.2
.2
.2
success
1
failure
1

Solution

- scheduler enumeration = investigate all 8 possible schedulers:

$$\sigma_1 = \{H \mapsto \text{dock},\ M \mapsto \text{dock},\ L \mapsto \text{dock},\}, P_{\sigma_1}(H \to \{T\}) = 0$$
$$\sigma_2 = \{H \mapsto \text{dock},\ M \mapsto \text{dock},\ L \mapsto \text{clean},\}, P_{\sigma_2}(H \to \{T\}) \approx .333$$
$$\sigma_3 = \{H \mapsto \text{dock},\ M \mapsto \text{clean},\ L \mapsto \text{dock},\}, P_{\sigma_3}(H \to \{T\}) \approx .714$$
$$\sigma_4 = \{H \mapsto \text{dock},\ M \mapsto \text{clean},\ L \mapsto \text{clean},\}, P_{\sigma_4}(H \to \{T\}) \approx .555$$
$$\sigma_5 = \{H \mapsto \text{clean},\ M \mapsto \text{dock},\ L \mapsto \text{dock},\}, P_{\sigma_5}(H \to \{T\}) \approx .925$$
$$\sigma_6 = \{H \mapsto \text{clean},\ M \mapsto \text{dock},\ L \mapsto \text{clean},\}, P_{\sigma_6}(H \to \{T\}) \approx .809$$
$$\sigma_7 = \{H \mapsto \text{clean},\ M \mapsto \text{clean},\ L \mapsto \text{dock},\}, P_{\sigma_7}(H \to \{T\}) \approx .862$$
$$\sigma_8 = \{H \mapsto \text{clean},\ M \mapsto \text{clean},\ L \mapsto \text{clean},\}, P_{\sigma_8}(H \to \{T\}) \approx .703$$

$$\Rightarrow \sigma_{\max} = \sigma_5$$

Solution

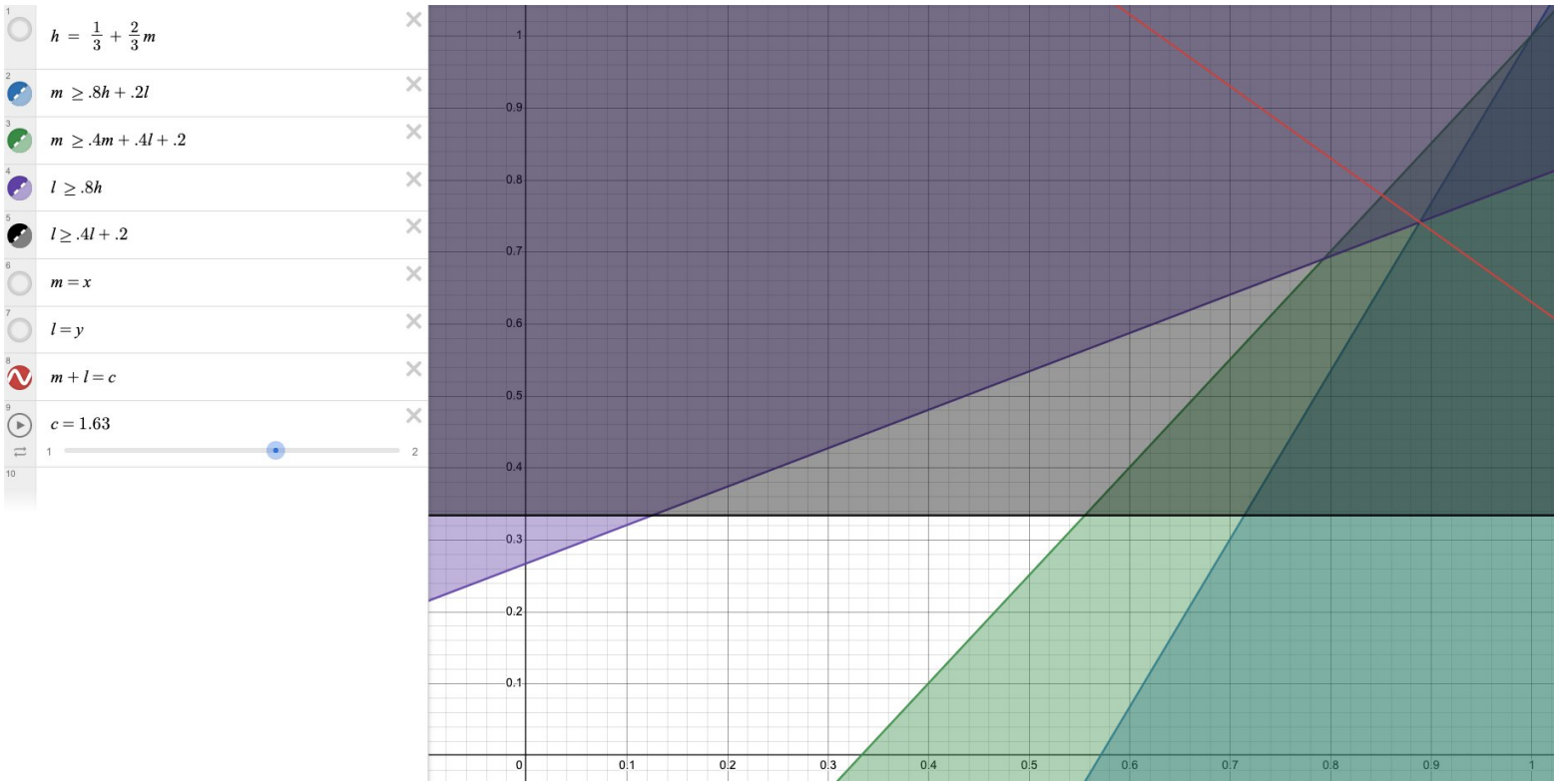
- using linear programming:
  $S_1 = \{T\} \Rightarrow \boldsymbol{x}(T) = 1 \quad S_0 = \{0\} \Rightarrow \boldsymbol{x}(0) = 0$
  - solve the following LP: minimize $\boldsymbol{x}(H) + \boldsymbol{x}(M) + \boldsymbol{x}(L)$ subject to

$$
\begin{aligned}
\boldsymbol{x}(H) &\geq 0.8 \cdot \boldsymbol{x}(H) + 0.2 \cdot \boldsymbol{x}(M) && \textit{h\_dock} \\
\boldsymbol{x}(H) &\geq 0.4 \cdot \boldsymbol{x}(H) + 0.4 \cdot \boldsymbol{x}(M) + 0.2 \cdot \boldsymbol{x}(T) && \textit{h\_clean} \\
\boldsymbol{x}(M) &\geq 0.8 \cdot \boldsymbol{x}(H) + 0.2 \cdot \boldsymbol{x}(L) && \textit{m\_dock} \\
\boldsymbol{x}(M) &\geq 0.4 \cdot \boldsymbol{x}(M) + 0.4 \cdot \boldsymbol{x}(L) + 0.2 \cdot \boldsymbol{x}(T) && \textit{m\_clean} \\
\boldsymbol{x}(L) &\geq 0.8 \cdot \boldsymbol{x}(H) + 0.2 \cdot \boldsymbol{x}(0) && \textit{l\_dock} \\
\boldsymbol{x}(L) &\geq 0.4 \cdot \boldsymbol{x}(L) + 0.4 \cdot \boldsymbol{x}(0) + 0.2 \cdot \boldsymbol{x}(T) && \textit{l\_clean} \\
0 \leq \boldsymbol{x}(H), \boldsymbol{x}(M), \boldsymbol{x}(L) &\leq 1
\end{aligned}
$$

  $\Rightarrow \boldsymbol{x}(H) \approx 0.925, \ \boldsymbol{x}(M) \approx 0.888, \ \boldsymbol{x}(L) \approx 0.740$
  - to deduce the optimizing scheduler, plug $\boldsymbol{x}$ into constraints above and see that constraints *h_clean*, *m_dock*, *l_dock* yield equality (at lower bound, slack=0): $\Rightarrow \sigma_{\max} = \{H \mapsto \text{clean}, \ M \mapsto \text{dock}, \ L \mapsto \text{dock}\}$
  - try out this linear program in LP solver

Solution

- graphical solution using graphing calculator:
  assume $\sigma(H) = \mathit{clean}$ (why?), then
  $\boldsymbol{x}(H) = 0.4 \cdot \boldsymbol{x}(H) + 0.4 \cdot \boldsymbol{x}(M) + 0.2 \Rightarrow \boldsymbol{x}(H) = \frac{1}{3} + \frac{2}{3} \cdot \boldsymbol{x}(M)$
  let $h := \boldsymbol{x}(H),\ x = m := \boldsymbol{x}(M),\ y = l := \boldsymbol{x}(L)$



- value iteration: see matlab/python script

Exercise 4 (homework). Engineers designed a budget version of the cleaning robot with a weaker navigation system, which allows the robot to find the charging dock with probability $p_{dock} = 0.5$ during each attempt.

1. Find a strategy that maximizes the robot's chances of cleaning the room before running out of battery.

Solution: $\sigma_{\max}(H) = \sigma_{\max}(M) = \text{clean}$, $\sigma_{\max}(L) = \text{dock}$, $P_{\max} \approx 0.714$

MSP 2024 - Tutorial 3

Exercise 1. Consider a stream of n elements where n is not known in advance.

1. Write a program in $\mathcal{O}(1)$ space that returns a random stream element, where all elements have equal probability of being picked.
2. Show correctness.

Solution

```
stream_random_element(stream):
1.  item = stream.next()
2.  item_count = 1
3.  while not stream.empty():
4.    next = stream.next()
5.    item_count += 1
6.    r = random_float(0,1)
7.    if r < 1/item_count:
8.      item = next
9.  return item
```

Correctness:

$P(\text{i-th element is returned}) = P(\text{i-th element is stored to variable } \texttt{item}) \cdot P(\texttt{item} \text{ is not rewritten afterwards}) = \frac{1}{i} \cdot \left(\frac{i}{i+1} \cdot \frac{i+1}{i+2} \cdots \frac{n-2}{n-1} \cdot \frac{n-1}{n} \right) = \frac{1}{n}$

Exercise 2. Assume an array with n elements.

1. Write a program that generates a random permutation of this array in $\mathcal{O}(n)$ time.
2. Show correctness.

Solution

```
array_shuffle(arr, n):
1.  for i from 1 to n-1:
2.     j = random_int(i,n)
3.     swap arr[i] and arr[j]
```

Correctness: $P($`j`-th element ends up on position `k`$) =$
$P($`arr[j]` is not swapped first (`k`-1) times$) \cdot$
$P($`arr[j]` is swapped during `k`-th iteration$) =$
$\left(\frac{n-1}{n} \cdot \frac{n-2}{n-1} \cdots \frac{n-k+1}{n-k+2}\right) \cdot \frac{1}{n-k+1} = \frac{1}{n}$

Exercise 3. Assume the following modification of the hiring problem, where the company must pay severance to the fired worker if they have just been hired:

```
Hire-Assistant(n)
1.  hire 1, best = 1
2.  for i from 2 to n:
3.    if candidate i is better than best:
4.      fire best
5.      if best == i-1:
6.        pay severance to best
7.      hire i, best = i
```

Assume that candidates arrive at the interview in random order. Compute how many times the severance will be paid:

a) in the best case, b) in the worst case, c) on average.

Solution

a. 0 times, e.g. if candidate 1 is the best one

b. $n-1$ times iff the candidates are sorted from worst to best

c. Let X_i, $2 \leq i \leq n$, be the indicator variable that attains value 1 if the severance is paid during i-th iteration, and 0 otherwise. The severance is paid during i-th iteration when (i-1)-th as well as i-th candidates were hired, that is, if candidate #(i-1) is the best among first (i-1) candidates and candidate #i is the best among first i candidates. Thus, $P(X_i = 1) = \frac{1}{i-1} \cdot \frac{1}{i}$.

The expected total number of paid severances is then

$$\mathrm{E}\left[\sum_{i=2}^{n} X_i\right] = \sum_{i=2}^{n} \mathrm{E}\left[X_i\right] = \sum_{i=2}^{n} P(X_i = 1) = \sum_{i=2}^{n} \frac{1}{i-1} \cdot \frac{1}{i} = \sum_{i=2}^{n} \frac{1}{i-1} - \frac{1}{i} =$$

$$= \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \cdots + \left(\frac{1}{n-2} - \frac{1}{n-1}\right) + \left(\frac{1}{n-1} - \frac{1}{n}\right) =$$

$$= 1 - 1/n$$

Exercise 4. Consider the following algorithm describing one pass of the *bubble-sort* algorithm. Assume that the input array a (indexed from 1) contains numbers from 1 to n in random order (all permutations have equal probability).

```
bubble-sort-one-pass(a,n):
1.  for i from 1 to n-1:
2.    if a[i] > a[i+1]:
3.      swap a[i] and a[i+1]
```

Determine:

1. Probability that no elements are swapped (best-case behaviour).
2. Probability that $n-1$ swaps are executed (worst-case behaviour).
3. The expected number of swaps (average-case behaviour) – it is sufficient to give the asymptotic number of swaps including an explanation.

Solution

1. No elements are swapped when array is sorted: $P = \frac{1}{n!}$
2. $n-1$ swaps are executed when `a[1]` contains n: $P = \frac{1}{n}$
3. Let X_i, $1 \leq i \leq n-1$, be the indicator variable that attains value 1 if a swap is performed during i-th iteration, and 0 otherwise. Right before the i-th iteration, `a[i]` contains the maximum value between `a[1]`...`a[i]`. Then, X_i is 0 when `a[i+1]` is the largest number between `a[1]`...`a[i+1]`, i.e. with probability $\frac{1}{i+1}$. Thus, $P(X_i = 1) = 1 - \frac{1}{i+1}$.

   The expected total number of swaps is then

$$\mathrm{E}\left[\sum_{i=1}^{n-1} X_i\right] = \sum_{i=1}^{n-1} \mathrm{E}\left[X_i\right] = \sum_{i=1}^{n-1} P(X_i = 1) = \sum_{i=1}^{n-1} 1 - \frac{1}{i+1} =$$

$$= \sum_{i=1}^{n-1} 1 - \sum_{i=1}^{n-1} \frac{1}{i+1} = n - 1 - \sum_{i=1}^{n-1} \frac{1}{i+1} =$$

$$= \mathcal{O}(n) - \mathcal{O}(\log n) = \mathcal{O}(n)$$

Exercise 5. Consider the following randomised algorithm `all_even` that tests whether array `arr` of size `len` > 0 (indexed from 1) contains only even numbers. Function `rand_int(1,len)` returns a random integer in range 1 to `len` with the uniform probability.

```
all_even(arr,len,n):
1. for i from 1 to n:
2.     k = rand_int(1,len)
3.     if arr[k] is odd:
4.         return false
5. return true
```

1. Decide and justify whether `all_even` is a Las Vegas or a Monte Carlo algorithm.
2. Construct function $\mathcal{P}(\texttt{len}, \texttt{n})$ that returns, for the given `len` and `n` > 0, the upper bound on the probability (with respect to all input arrays of size `len`), that `all_even(arr,len,n)` returns a wrong result.
3. Determine the smallest `n` such that the worst-case probability that `all_even(arr,10,n)` returns a wrong result is smaller than 50%.

Solution

1. It is a Monte Carlo algorithm because there is a non-zero probability that for a given input the algorithm returns a wrong result.
2. Clearly, if $\texttt{len} = 1$, then $\mathcal{P}(\texttt{len}, \texttt{n}) = 0$. It is easy to see that the worst-case input contains only a single odd number. For such inputs $\mathcal{P}(\texttt{len}, \texttt{n}) = \left(\frac{len-1}{len}\right)^n$.
3. Since $P(10, \texttt{n}) = \left(\frac{9}{10}\right)^n$, we require that $\left(\frac{9}{10}\right)^n < 0.5$, i.e. $\log_{0.9}(0.9^n) > \log_{0.9}(0.5) \Rightarrow n > 6.5$. Hence $n = 7$.

MSP_4_cviko

pondělí 7. října 2024 18:20

Derive MLE estimator for parameters of normal distribution. Estimate μ and σ^2 for observed values: 10,2; 11,5; 9,7; 10,8; 11,1; 9,9 and 10,6.
PDF for normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Consider a mixture of two normal distributions with $\mu_1 \neq \mu_2$ how would the likelihood function change?

$\hat{\mu} = 10,5429$

$\hat{\sigma}^2 = 0,362449$

Formulate likelihood and log-likelihood function for n IID 2-parametric exponentially distributed random variables. Derive MLE estimator for both parameters and estimate them for observed values: 6,5; 2,9; 1,3; 3,7; 3,4; 7,8; 7,1; 4,1; 4,8; 6,1.
PDF for exponential distribution:

$$f(x) = \begin{cases} \lambda e^{-\lambda(x-a)} & x > a \\ 0 & otherwise \end{cases}$$

Exponential distribution can be used as a model for "time to failure". Find a way to incorporate surviving parts into likelihood function.

$\bar{x} = 4,77 \rightarrow \hat{\lambda} = 0,288$

$\min(x_i) = 1,3$

Amount of latecomers into MSP lessons can be viewed as a random variable with Poisson distribution (see PMF). Derive MLE for parameter λ and estimate lambda given following observations from last year: 1; 2; 4; 2; 1; 4; 1; 1; 3; 5.

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Where $x \in \mathbb{N} \cup \{0\}$

$$1)\ L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \qquad \ell(\lambda) = \sum_{i=1}^{n} x_i \ln(\lambda) - \sum_{i=1}^{n} \ln(x_i!) - n\lambda$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{\sum x_i}{\lambda} - n \rightarrow \hat{\lambda} = \bar{x}$$

$\bar{x} = 2{,}4$

Bonus

Try to estimate population size of some species via capture-recapture scheme. Capture-recapture uses hypergeometric distribution vith PMF:

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

python code

KÓD V ODKAZU, LOGIN DO GOOGLE

Derive MLE for parameter of Pareto distribution given by following PDF

$$f(x) = \begin{cases} \frac{\alpha}{x^{\alpha+1}} & x > 1 \\ 0 & \textit{otherwise.} \end{cases}$$

Discuss existence of moments (depending on values α) for this probability distribution.

1) $L(\alpha) = \prod_{i=1}^{n} \frac{\alpha}{x^{\alpha+1}} \rightarrow \ell(\alpha) = \sum_{i=1}^{n} \ln\left(\frac{\alpha}{x_i^{\alpha+1}}\right) = n \cdot \ln(\alpha) - \sum_{i=1}^{n} (\alpha+1) \cdot \ln(x_i)$

$\frac{\partial \ell}{\partial \alpha} = \frac{n}{\alpha} - \sum_{i=1}^{n} \ln(x_i) \rightarrow \hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \ln(x_i)}$

2) $m \in \mathbb{R} \Rightarrow \int_1^{\infty} x^k \cdot \frac{\alpha}{x^{\alpha+1}} dx \rightarrow$ dobrý pro jaké α existují momenty

Consider MLE for a parameter σ^2 of Normal distribution (with unknown μ).

- Find out whether the MLE is biased or not.
- If the MLE is biased, find an unbiased estimator.
- Discuss bias of any estimators for standard deviation.

$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2\sum x_i \mu + n\mu^2\right) - \left(\bar{x}^2 - 2\bar{x}\mu + \mu^2\right) =$

$= \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \mu)^2\right) - (\bar{x} - \mu)^2 \Rightarrow E\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = E\left(\frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \mu)^2 - (\bar{x} - \mu)^2\right)\right)$

$E\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right) \rightarrow x_i - \mu \sim N(0; \sigma^2) \Rightarrow E\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right) = \frac{1}{n} \cdot n\sigma^2 = \sigma^2$

$E\left((\bar{x} - \mu)^2\right) \rightarrow$ Rozptyl $\bar{x}$; $\bar{x} \sim N(\mu; \frac{\sigma^2}{n}) \Rightarrow E\left((\bar{x} - \mu)^2\right) = \frac{\sigma^2}{n}$

dohromady: $E\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n\sigma^2 - \sigma^2}{n} = \frac{n-1}{n} \cdot \sigma^2 \neq \sigma^2$

Nestranný odhad: $\frac{1}{n} \rightarrow \frac{1}{n-1} \Rightarrow \frac{n-1}{n-1}\sigma^2 = \sigma^2$

Směrodatná odchylka: $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \rightarrow \sqrt{E(x)} \neq E(\sqrt{x})$

$\sqrt{x}$ konkávní $\Rightarrow E\left(\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) < \sqrt{E\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)} = \sigma$

$\Rightarrow$ odhad je vychýlený a podhodnocuje

Derive MLE for a parameter of geometric distribution (number of Bernoulli trials before first success) given by following PMF

$$p(x) = (1-p)^{x-1}p$$

Discuss a bias of this estimator (if the MLE is biased try to use Jensen's inequality).
Show that geometric distribution is a part of exponential family.
Use factorization criterion to show that the resulting MLE is using sufficient statistic.

$L(p) = \prod_{i=1}^{n} (1-p)^{x_i - 1} \cdot p$

$\ell(p) = \sum_{i=1}^{n} \ln\left[(1-p)^{x_i - 1} \cdot p\right] = \sum_{i=1}^{n}\left[(x_i - 1) \cdot \ln(1-p) + \ln p\right] =$

$= \sum_{i=1}^{n} x_i \cdot \ln(1-p) - n\ln(1-p) + n\ln(p)$

$\frac{\partial \ell}{\partial p} = -\frac{\sum_{i=1}^{n} x_i}{1-p} + \frac{n}{1-p} + \frac{n}{p} = 0$

$\frac{\sum_{i=1}^{n} x_i}{n} = \frac{1-p}{1-p} + \frac{1-p}{p} = \frac{1}{p} \rightarrow \hat{p} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$

$E\left(\frac{1}{\bar{x}}\right) > \frac{1}{E(\bar{x})}$ protože $\frac{1}{x}$ je konvexní $\Rightarrow$

$\Rightarrow \hat{p}$ nadhodnocuje

$p(x) = (1-p)^{x-1}p \rightarrow e^{\ln\left((1-p)^{x-1}p\right)} = e^{(x-1)\cdot\ln(1-p) + \ln p}$

$\Rightarrow h(x) = 1$; $T(x) = x - 1$ $\eta(\theta) = \ln(1-p)$ $A(\theta) = \ln p$

$$\prod_{i=1}^{n}\left[(1-p)^{x_i-1}\cdot p\right] = (1-p)^{\sum_{i=1}^{n}(x_i-1)}\cdot p^n = (1-p)^{\sum_{i=1}^{n}x_i}\cdot\left(\frac{p}{1-p}\right)^n \Rightarrow$$

$$T(x) = \sum_{i=1}^{n} x_i \quad \text{is sufficient}$$

Find a (jointly) sufficient statistics for following distributions:

- Poisson distribution $p(x) = \frac{\lambda^x}{x!}e^{-\lambda}; x \in \{0; 1; ...\}$
- Normal distribution
- Gamma distribution $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}; x \geq 0$

Gamma function wiki

1) $$P(x,\lambda) = \prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}\cdot e^{-n\lambda} = \prod_{i=1}^{n}\left(\frac{1}{x_i!}\right)\cdot\left[\lambda^{\sum_{i=1}^{n}x_i}\cdot e^{-n\lambda}\right] \Rightarrow$$

$$u(x) = \prod_{i=1}^{n}\left(\frac{1}{x_i!}\right) \;;\; v(x_i,\lambda) = \lambda^{\sum_{i=1}^{n}x_i}\cdot e^{-n\lambda} \Rightarrow T_1(x) = \sum_{i=1}^{n}x_i$$

2) $$f(x,\mu,\sigma^2) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\cdot e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi}\cdot\sigma}\right)^n\cdot e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}} =$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\cdot e^{-\frac{1}{2\sigma^2}\cdot\sum_{i=1}^{n}x_i^2 + \frac{\mu}{\sigma^2}\cdot\sum_{i=1}^{n}x_i - \frac{\mu^2}{2\sigma^2}} \Rightarrow \begin{matrix} T_1 = \sum_{i=1}^{n}x_i \\ T_2 = \sum_{i=1}^{n}x_i^2 \end{matrix}$$

3) $$f(x,\alpha,\beta) = \prod_{i=1}^{n}\frac{\beta^\alpha}{\Gamma(\alpha)}x_i^{\alpha-1}e^{-\beta x_i} = \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n}\cdot e^{n(\alpha-1)\sum_{i=1}^{n}\ln(x_i)}\cdot e^{-\beta\sum_{i=1}^{n}x_i}$$

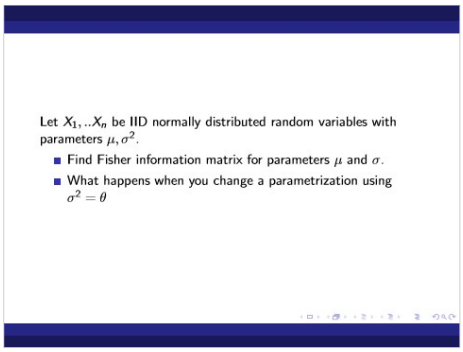
$$T_1(x) = \sum_{i=1}^{n}x_i \qquad T_2(x) = \sum_{i=1}^{n}\ln(x_i) \qquad \text{protože } \sum_{i=1}^{n}\ln(x_i) = \ln\left(\prod_{i=1}^{n}(x_i)\right)$$

nebo

$$T_2^*(x) = \prod_{i=1}^{n}(x_i)$$

MSP_6_cviko

úterý 22. října 2024 8:07



$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \rightarrow \ell(\mu,\sigma) = -\ln(\sqrt{2\pi}) - \ln(\sigma) - \frac{(x-\mu)^2}{2\sigma^2}$$

$$\frac{\partial \ell}{\partial \mu} = \frac{x-\mu}{\sigma^2} \qquad \frac{\partial \ell}{\partial \sigma} = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}$$

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{1}{\sigma^2} \qquad \frac{\partial^2 \ell}{\partial \mu \partial \sigma} = -\frac{2}{\sigma^3}(x-\mu)$$

$$\frac{\partial^2 \ell}{\partial \sigma^2} = \frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4}$$

$$E(X) = \mu$$

$$-E\left[-\frac{1}{\sigma^2}\right] = \frac{1}{\sigma^2} \qquad -E\left(-\frac{2}{\sigma^3}(x-\mu)\right) = \frac{2}{\sigma^3} \cdot \int (x-\mu) f(x)\, dx = 0$$

$$-E\left[\frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4}\right] = -\frac{1}{\sigma^2} + \frac{3}{\sigma^4} \cdot E\left((x-\mu)^2\right) = -\frac{1}{\sigma^2} + \frac{3\sigma^2}{\sigma^4} = \frac{2}{\sigma^2}$$

$$E\left[(X - E(X))^2\right] = \sigma^2$$

$$J(\mu,\sigma) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}$$

$$f(x) = \frac{1}{\sqrt{2\pi\theta}} \cdot e^{-\frac{(x-\mu)^2}{2\theta}} \rightarrow \ell(\mu,\theta) = -\frac{1}{2}\ln(\theta) - \frac{1}{2}\ln(2\pi) - \frac{(x-\mu)^2}{2\theta}$$

$$\frac{\partial \ell}{\partial \mu} = \frac{x-\mu}{\theta} \qquad \frac{\partial \ell}{\partial \theta} = -\frac{1}{2\theta} + \frac{(x-\mu)^2}{2\theta^2}$$

$$\frac{\partial \ell}{\partial \mu^2} = -\frac{1}{\theta} \qquad \frac{\partial^2 \ell}{\partial \mu \partial \theta} = -\frac{x-\mu}{\theta^2} \qquad -E\left[-\frac{x-\mu}{\theta^2}\right] = 0$$

$$\frac{\partial^2 \ell}{\partial^2 \theta} = \frac{1}{2\theta^2} - \frac{(x-\mu)^2}{\theta^3} \qquad -E\left[+\frac{1}{2\theta^2} - \frac{(x-\mu)^2}{\theta^3}\right] = -\frac{1}{2\theta^2} + \frac{\theta}{\theta^3} = \frac{1}{2\theta^2}$$

$$J(\mu,\theta) = \begin{pmatrix} \frac{1}{\theta} & 0 \\ 0 & \frac{1}{2\theta^2} \end{pmatrix}$$

Let $X_1, \dots X_n$ be IID distributed random variables with PMF:

$$P(x) = \binom{x-1}{r-1}\pi^r(1-\pi)^{x-r}; x \in \{r, r+1, r+2, \dots\}.$$

Negative binomial distribution - number of observed Bernoulli trials with probability of success π to get r successes. Assume that r is known.

- Find a MLE of π for known r.
- Decide whether this Negative binomial distribution belongs to exponential family.
- Try to find a sufficient statistic for a parameter π
- Compute Fisher information for a parameter π
- You observed n NB trials with $r = 3$. $\sum_{i=1}^{100} x_i = 2999$. Find the 95% asymptotic CI for π

$$\ell(\pi) = \sum \ln\binom{x_i - 1}{r-1} + r r \cdot \ln(\pi) + \sum_{i=1}^{n} (x_i - r)\ln(1-\pi) =$$

$$= \sum \ln\binom{x_i - 1}{r-1} + \sum x_i \ln(1-\pi) + n \cdot r \cdot \ln(\pi) - n r \ln(1-\pi)$$

$$\frac{\partial \ell}{\partial \pi} = -\frac{\sum_{i=1}^{n} x_i}{1-\pi} + \frac{n \cdot r}{\pi} + \frac{n \cdot r}{1-\pi} = 0 \Rightarrow \frac{\sum_{i=1}^{n} x_i}{n \cdot r} = \frac{1}{\pi}$$

$$\hat{\pi} = \frac{n \cdot r}{\sum_{i=1}^{n} x_i}$$

$$\binom{x-1}{r-1} \cdot e^{r \ln(\pi) + (x-r)\cdot \ln(1-\pi)} = \binom{x-1}{r-1} \cdot e^{x \cdot \ln(1-\pi) + r \ln\left(\frac{\pi}{1-\pi}\right)}$$

pro $n = 1$

$$\binom{x_i - 1}{r-1}^n \cdot e^{\sum x_i \ln(1-\pi) + r n \ln\left(\frac{\pi}{1-\pi}\right)} \Rightarrow T(x) = \sum_{i=1}^{n} x_i \quad \Big\| \quad \frac{\partial^2 \ell}{\partial \pi^2} = -\frac{x}{(1-\pi)^2} - \frac{r}{\pi^2} + \frac{r}{(1-\pi)^2}$$

$$-E\left(\frac{-x}{(1-\pi)^2}\right) = \frac{1}{(1-\pi)^2} \cdot E(x) = \frac{r}{\pi(1-\pi)^2} + \frac{r}{\pi^2} - \frac{r}{(1-\pi)^2} = \frac{r(\pi + (1-\pi)^2 - \pi^2)}{\pi^2(1-\pi)^2} = \frac{r(\pi + 1 - 2\pi + \pi^2 - \pi^2)}{\pi^2(1-\pi)^2} =$$

$$= \frac{r}{\pi^2(1-\pi)} = J(\pi)$$

$$CI: \left\langle \hat{\pi} - u_{1-\alpha/2}\sqrt{\frac{1}{J_n(\pi)}} ; \hat{\pi} + u_{1-\alpha/2}\sqrt{\frac{1}{J_n(\pi)}} \right\rangle =$$

$$\hat{\pi} \doteq 0{,}1 \Rightarrow \left\langle 0{,}1 - 1{,}96 \cdot \sqrt{\frac{1}{\frac{100 \cdot 3}{0{,}1^2(1-0{,}1)}}} ; 0{,}1 + 0{,}01074 \right\rangle = \langle 0{,}0893; 0{,}1107 \rangle$$

Let $X_1, \dots X_n$ be IID random variables following a Poisson distribution with a parameter λ.

- Compute Fisher information for λ.
- You observed this random variable $n = 150$ times with a $\sum_{i=1}^{150} x_i = 351$. Find the 99% asymptotic CI for λ.
- Use likelihood ratio to test a hypothesis $H_0 : \lambda = 2$.

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda} \Rightarrow \ell(\lambda) = x \ln(\lambda) - \ln(x!) - \lambda$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{x}{\lambda} - 1 \qquad \frac{\partial^2 \ell}{\partial \lambda^2} = -\frac{x}{\lambda^2} \qquad -E\left(-\frac{x}{\lambda^2}\right) = \frac{1}{\lambda} = J(\lambda)$$

$$J_n(\lambda) = \frac{n}{\lambda} \qquad \bar{x} = 2{,}34 \Rightarrow \lambda \in \left\langle 2{,}34 - 2{,}576 \cdot \sqrt{\frac{2{,}34}{150}} ; 2{,}34 + 0{,}322 \right\rangle$$

$$\lambda \in \langle 2{,}012 ; 2{,}662 \rangle$$

$$LR = 2\left[\ell(\hat{\lambda}) - \ell(\lambda_0)\right] =$$

$$2\left[\sum x_i \ln(\hat{\lambda}) - \ln\left(\prod_{i=1}^{n} x_i!\right) - n\hat{\lambda} - \left(\sum x_i \ln(\lambda_0) - \ln\left(\prod_{i=1}^{n} x_i!\right) - n\lambda_0\right)\right] =$$

$$= 2n\left[\bar{x} \cdot (\ln(\bar{x}) - \ln(\lambda_0)) - (\bar{x} - \lambda_0)\right] = 8{,}216$$

$$\chi^2_{0,95}(1) = 3{,}841 \qquad LR > 3{,}841 \Rightarrow \text{zamítám } H_0$$

Let Y be a normally distributed random variable. Find a PDF for a random variable X created by a formula $e^Y = X$.

- Find a MLE for μ, σ^2 using observations of X.
- Prove that the Fisher information matrix will not change by the transformation.

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(y-\mu)^2}{2\sigma^2}} \quad \rightarrow \quad e^y = x$$

$$y = \ln(x)$$

$$dy = \frac{dx}{x}$$

$$-\infty \rightarrow 0$$

$$\infty \rightarrow \infty$$

$$f(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma\cdot x} \cdot e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$$

$$l(\mu,\sigma) = -n\ln(\sqrt{2\pi}) - n\ln(\sigma) - n\cdot\ln\left(\prod x_i\right) - \frac{\sum_{i=1}^{n}(\ln(x_i)-\mu)^2}{\sigma^2}$$

$$\frac{\partial l}{\partial \mu} = + \frac{\sum_{i=1}^{n}\ln(x_i) - n\mu}{\sigma^2} \quad \rightarrow \quad \hat{\mu} = \frac{\sum_{i=1}^{n}\ln(x_i)}{n}$$

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(\ln(x_i)-\mu)^2}{\sigma^3} \quad \rightarrow \quad \hat{\sigma}^2 = \frac{\sum(\ln(x_i)-\hat{\mu})^2}{n}$$

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{1}{\sigma^2} \qquad \frac{\partial^2 l}{\partial \mu\,\partial\sigma} = -\frac{2}{\sigma^3}(\ln(x)-\mu)$$

$$\frac{\partial^2 l}{\partial \sigma^2} = \frac{1}{\sigma^2} + \frac{3}{\sigma^4}\cdot(\ln(x)-\mu)^2$$

$$\int_{-\infty}^{\infty}(\ln(x)-\mu)\cdot\frac{1}{x}\cdot\frac{1}{\sqrt{2\pi}\sigma}\cdot e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}dx \quad \left| \begin{array}{l} y = \ln(x) \\ dy = \frac{dx}{x} \\ x\cdot dy = dx \end{array} \right| = \int_{0}^{\infty}(y-\mu)\cdot\frac{1}{\sqrt{2\pi}\sigma}\cdot e^{-\frac{(y-\mu)^2}{2\sigma^2}}dy$$

$\downarrow$

NORMÁLNÍ ROZDĚLENÍ

MSP_7_cviko

pondělí 28. října 2024 18:32

Assume that $X_1, ..., X_{n_1}$ are IID binomially distributed variables, and $Y_1, ..., Y_{n_2}$ are IID geometrically distributed variables.

- find a mle estimate for probability parameter from just binomial part
- find a MLE estimate for probability parameter from joint sample $X_1, ...X_{n_1}, Y_1, ..., Y_{n_2}$
- let the parameter n for binomial distribution be $n = 100$, you observed $n_1 = 100$ binomial trials with $\sum_{i=1}^{n_1} x_i = 1008$, then you observed $n_2 = 100$ geometric trials with $\sum_{j=1}^{n_2} y_j = 1247$. Use likelihood ratio test to decide whether the value of a probability parameter changed. Use $\alpha = 5\%$

$X_i \sim Bi(n, \pi_B)$ $\quad \pi_B = \pi_G = \pi \Rightarrow$

$Y_i \sim Ge(\pi_G)$ $\quad L(\pi) = \prod_{i=1}^{n_1} \binom{n}{x_i} \pi^{x_i}(1-\pi)^{n-x_i} \cdot \prod_{j=1}^{n_2} \pi(1-\pi)^{y_j - 1}$

$$l(\pi) = \ln\left(\prod_{i=1}^{n_1}\binom{n}{x_i}\right) + \sum_{i=1}^{n_1} x_i \cdot \ln(\pi) + \sum_{i=1}^{n_1}(n - x_i)\cdot \ln(1-\pi) + n_2 \ln(\pi) + \sum_{j=1}^{n_2}(y_j - 1)\ln(1-\pi)$$

$$\frac{\partial l}{\partial \pi} = \frac{\sum_{i=1}^{n_1} x_i}{\pi} - \frac{n_1 n}{1-\pi} + \frac{\sum_{i=1}^{n_1} x_i}{1-\pi} + \frac{n_2}{\pi} - \frac{\sum_{j=1}^{n_2} y_j - n_2}{1-\pi} = 0$$

$$(1-\pi)\sum_{i=1}^{n_1} x_i - \pi n_1 n + \pi \sum_{i=1}^{n_1} x_i + (1-\pi) n_2 - \pi \sum_{j=1}^{n_2} y_j + \pi n_2 = 0$$

$$\sum_{i=1}^{n_1} x_i - \pi n_1 n + n_2 - \pi \sum_{j=1}^{n_2} y_j = 0 \Rightarrow \hat{\pi} = \frac{\sum_{i=1}^{n_1} x_i + n_2}{n_1 n + \sum_{j=1}^{n_2} y_j} \doteq 0{,}0995$$

$$\frac{\partial l}{\partial \pi_B} = \frac{\sum x_i}{\pi_B} - \frac{n_1 n}{1-\pi_B} + \frac{\sum x_i}{1-\pi_B} = 0 \Rightarrow \sum x_i - \pi n_1 n = 0$$

$$\hat{\pi}_B = \frac{\sum x_i}{n_1 n} = 0{,}1008$$

$$\hat{\pi}_G = \frac{1}{\bar{y}} \doteq 0{,}0802$$

$$LR = 2\left[\left(\ln\left(\prod_{i=1}^{n_1}\binom{x_i}{n}\right) + \sum_{i=1}^{n_1} x_i \ln(\hat{\pi}_B) + \left(n_1 n - \sum_{i=1}^{n_1} x_i\right)\ln(1-\hat{\pi}_B) + n_2 \ln(\hat{\pi}_G) + \sum_{j=1}^{n_2} y_j \ln(1-\hat{\pi}_G) - n_2 \ln(1-\hat{\pi}_G)\right) - \left(\ln\left(\prod_{i=1}^{n_1}\binom{x_i}{n}\right) + \sum_{i=1}^{n_1} x_i \cdot \ln(\hat{\pi}) + \left(n_1 n - \sum_{i=1}^{n_1} x_i\right)\ln(1-\hat{\pi}) + n_2 \ln(\hat{\pi}) + \sum_{j=1}^{n_2} y_j \ln(1-\hat{\pi}) - n_2 \ln(1-\hat{\pi})\right)\right] =$$

$$= 2\left[\sum x_i \ln\left(\frac{\hat{\pi}_B}{\hat{\pi}}\right) + \left(n_1 n - \sum_{i=1}^{n_1} x_i\right)\ln\left(\frac{1-\hat{\pi}_B}{1-\hat{\pi}}\right) + n_2 \ln\left(\frac{\hat{\pi}_G}{\hat{\pi}}\right) + \left(\sum_{j=1}^{n_2} y_j - n_2\right)\ln\left(\frac{1-\hat{\pi}_G}{1-\hat{\pi}}\right)\right] =$$

$$\doteq 5{,}586 \qquad \bar{W}_{0,05} = \langle 0; \chi^2_{0,95}(1)) = \langle 0; 3{,}841 \rangle$$

$LR \notin \bar{W}_{0,05} \Rightarrow$ ZAMÍTÁM

Let $X_1, ..., X_9$ be IID distributed Bernoulli trials.

- Identify a distribution of $\sum_{i=1}^{9} x_i$
- If $\sum_{i=1}^{9} x_i = 3$ test a hypothesis $H_0 : \pi = 1/2$ against $H_1 : \pi < 1/2$ for $\alpha = 0, 15$
- Estimate β and power, if the critical value of π to detect is $\pi = 1/4$

$\sum X_i \sim Bi(9; \pi)$

$x_i = 3 \Rightarrow P(X \le 3 | \pi = \frac{1}{2}) \doteq 0{,}254 > \alpha \Rightarrow$ NEZAMÍTÁM H_0

$W_\alpha = \{0; 1; 2\}$ - protože $p(0) + p(1) + p(2) < \alpha$

$\beta \Rightarrow$ NEZAMÍTÁM $\wedge$ NEPLATÍ $\Rightarrow P(X \in \bar{W}_\alpha | H_1) = P(X \ge 3 | \pi = \frac{1}{4}) \doteq 0{,}4$

Assume that the observations
11, 5; 6, 7; 7, 8; 7, 1; 10, 3; 12, 2; 11, 4; 12, 6; 9, 8; 10, 9 follow a normal distribution with unknown parameters μ, σ.

- Compute 95% CI's for both μ and σ.
- Using $\alpha = 5\%$ test a hypothesis that $H_0 : \mu = 11$ against $H_1 : \mu \neq 11$.
- Derive a test for $H_0 : \sigma^2 = \sigma_0^2$.
- Test $H_0 : \sigma^2 = 2$ against $H_1 : \sigma^2 \neq 2$ using $\alpha = 1\%$

$\bar{x} = 10{,}03 \quad s^2(x) = 4{,}542$

$s(x) = \sqrt{\frac{\hat{}}{n-1} \dots} \doteq 2{,}131$

$\mu \in \langle 8{,}505; 11{,}555 \rangle$

$\sigma^2 \in \langle 2{,}149; 15{,}140 \rangle$

$\sigma \in \langle 1{,}466; 3{,}891 \rangle$

$t = -1{,}44 \quad t_{0,975}(9) \doteq 2{,}262$

$t \in \bar{W}_\alpha \quad \bar{W}_\alpha = \langle -2{,}262; 2{,}262 \rangle$

$\Rightarrow$ NEZAMÍTÁM

1) Rozptyl nesáhnín na μ

2) pro $N(\mu; 1)$ platí $\frac{(n-1)S^2(X)}{1} \sim \chi^2(n-1)$

3) pro $N(\mu; \sigma_0^2) \Rightarrow \frac{(n-1)\cdot S^2(X)}{\sigma_0^2} \sim \chi^2(n-1)$

$\Rightarrow$ KDYŽ PLATÍ H_0 $\quad t = \frac{(n-1)S^2(X)}{\sigma_0^2} \sim \chi^2(n-1)$

$t = 20{,}44 \quad \bar{W}_{0,01} = \langle 1{,}735; 23{,}589 \rangle$

$t \in \bar{W} \Rightarrow$ NEZ.

MSP_8_cviko

neděle 3. listopadu 2024 10:05

Assume that the observations
11, 5; 6, 7; 7, 8; 7, 1; 10, 3; 12, 2; 11, 4; 12, 6; 9, 8; 10, 9 follow a normal distribution with unknown parameters μ, σ.

- Compute 95% CI's for both μ and σ.
- Using $\alpha = 5\%$ test a hypothesis that $H_0 : \mu = 11$ against $H_1 : \mu \neq 11$.
- Derive a test for $H_0 : \sigma^2 = \sigma_0^2$.
- Test $H_0 : \sigma^2 = 2$ against $H_1 : \sigma^2 \neq 2$ using $\alpha = 1\%$

$\bar{x} = 10,05 \quad s^2(x) = 4,542$

$s(x) = \sqrt{\frac{\hat{}}{n-1} \dots} = 2,131$

$\mu \in \langle 8,505 ; 11,555 \rangle$

$\sigma^2 \in \langle 2,149 ; 15,140 \rangle$

$\sigma \in \langle 1,466 ; 3,891 \rangle$

$t = -1,44 \quad t_{0,975}(9) = 2,262$

$t \in \bar{W}_\alpha \quad \bar{W}_\alpha = (-2,262 ; 2,262)$

$\Rightarrow$ NEZAMÍTÁM

1) Rozptyl neznámý na μ

2) pro $N(\mu; 1)$ platí $\frac{(n-1)S^2(X)}{1} \sim \chi^2(n-1)$

3) pro $N(\mu; \sigma_0^2)$ $\Rightarrow \frac{(n-1) \cdot S^2(X)}{\sigma_0^2} \sim \chi^2(n-1)$

$\Rightarrow$ KDYŽ PLATÍ H_0 $t = \frac{(n-1)S^2(X)}{\sigma_0^2} \sim \chi^2(n-1)$

$t = 20,144 \quad \bar{W}_{0,01} = (1,735 ; 23,589)$

$t \in \bar{W} \Rightarrow$ NEZ.

You are a doctor and want to determine if a new medication reduces blood pressure in patients. You have a group of patients whose blood pressure was measured before and after taking the medication. You want to compare their results before and after the treatment to see if there was a reduction in blood pressure. For any hypothesis testing use $\alpha = 5\%$

ID	1	2	3	4	5	6	7	8	9	10
Before	150	160	145	155	148	152	149	153	157	151
After	147	157	145	154	147	150	148	151	155	149

Table: Blood Pressure Measurements Before and After Treatment

Compute descriptive statistics for both datasets.

$\bar{B} = 152 \quad \bar{A} = 150,3$

$s^2(B) = 19,78 \quad s^2(A) = 18,34$

$s(B) = 4,45 \quad s(A) = 3,92$

$\bar{D} = 1,7$

$s(D) = 0,949$

$H_0 : \mu_d = 0 \quad D = B - A$

$H_1 : \mu_d > 0$

$t = 5,67 \quad \bar{W}_\alpha = (-\infty ; 1,833)$

Use Central limit theorem (wiki, Python illustration) to derive a test of a hypothesis about equality of probability parameters in 2 different sets of Bernoulli trials ($H_0 : \pi_x = \pi_y$, use joint standard deviation). Imagine that the probability of Bernoulli trials corresponds to the proportion of defective parts manufactured on 2 different manufacturing stations. You observed 11 defects out of 211 at station x and 23 out of 303 at station y.

- Test a hypothesis $H_0 : \pi_x = \pi_y$ against $H_1 : \pi_x \neq \pi_y$ for $\alpha = 5\%$
- Derive a likelihood ratio test for the same H_0

1) CLT and $H_0 \Rightarrow \Sigma$ of Bernoulli trials with π

$\frac{\Sigma x_i}{n_x} - \frac{\Sigma y_i}{n_y} = \hat{\pi}_x - \hat{\pi}_y \to 0$ (podle H_0)

$E((X - E(X))^2) = \pi(1-\pi)$ $\Rightarrow$ Bernoulli trials Variance

$\frac{\Sigma x_i + \Sigma y_i}{n_x + n_y} = f = \hat{\pi}$

Linear model $X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \quad \beta = \begin{pmatrix} \hat{\pi}_x \\ \hat{\pi}_y \end{pmatrix} \quad V = \sigma^2 \mathbb{I}$

with known σ^2 $\frac{\frac{\Sigma x_i}{n_x} - \frac{\Sigma y_i}{n_y}}{\sigma \cdot \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$ as. $\sim N(0;1)$; for Bernoulli trials $\sigma^2 = \pi(1-\pi)$

$\Rightarrow$ use joint estimate f

po úpravě: $t = \frac{\hat{\pi}_x - \hat{\pi}_y}{\sqrt{f(1-f)}} \cdot \sqrt{\frac{n_x \cdot n_y}{n_x + n_y}} = -1,07$

$t \in \bar{W} \Rightarrow$ NEZ.

$\bar{W}_{0,05} = \langle -1,96 ; 1,96 \rangle$

$l(\pi) = \Sigma x_i \ln(\pi) + \Sigma (1 - x_i) \ln(1 - \pi)$

$LR = 2\Big[\Big(\sum_i x_i \ln(\bar{x}) + (n_x - \sum_i x_i) \ln(1 - \bar{x}) + \sum_j y_j \ln(\bar{y}) + (n_y - \Sigma y_j) \ln(1 - \bar{y}) \Big) -$

$- \Big(\Big(\sum_i x_i + \sum_j y_j \Big) \ln(f) + \Big((n_x - \sum_i x_i) + (n_y - \sum_j y_j) \Big) \ln(1 - f) \Big) = \dots$

Find if there is a statistically significant difference ($\alpha = 5\%$) among mean test scores of high school graduation exams in different regions of Czechia in 2023. (data are randomly generated). You only obtained self reported subset of scores from each region.

region	JM	O	V	Z
no. of obs.	199	129	60	~~113~~ 115
mean score	74,81	72,11	71,00	72,61
sum fo squares	1123833,76	677710,65	305452,98	600399,20

Assume that all necessary assumption for ANOVA hold. Arange your results into ANOVA table.

$n = 503 \quad \bar{x} = 73,1623 \quad \sum_{i=1}^{n} x_i^2 = 2707395,61$

$S_{0Z} = 540,27$

$S_O = 142,85$

$S_V = 280,53$

$S_Z = 34,47$

	S	df	MS	F-value
Region	998,72	3	332,907	11,74
ERROR	14140,534	499	28,338	
TOTAL	15139,317	502	30,158	

$\overline{W_\alpha} = \langle 0 ; F_{0,95}(3;499) \rangle = \langle 0; 2,6227 \rangle \quad F \notin \overline{W_\alpha} \Rightarrow H_0$ ZAM.

Find if there is a statistically significant difference ($\alpha = 5\%$) among mean test scores of high school graduation exams in different regions of Czechia in 2023. (data are randomly generated). You only obtained self reported subset of scores from each region.

region	JM	O	V	Z
no. of obs.	199	129	60	115
mean score	$74,81$	$72,11$	$71,00$	$72,61$
sum fo squares	$1123833,76$	$677710,65$	$305452,98$	$600399,20$

Assume that all necessary assumption for ANOVA hold. Arange your results into ANOVA table.

Assume you are trying to compare mean blood pressure of people from 3 different countries. You obtained following observations:

Country A	120	122	118	121	
Country B	130	128	132		
Country C	125	127	126	124	128

Is there a statistically significant difference ($\alpha = 5\%$) among means of blood pressure by Country. Find pairwise differences and test all assumptions of ANOVA.

$$S_T = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 196,917$$

$$S_A = \sum_{j=1}^{3} n_{\cdot j} (\bar{y}_{\cdot j} - \bar{y})^2 = 170,16$$

$$S_e = S_T - S_A = 26,75$$

$H_0: \mu_A = \mu_B = \mu_C$ $\quad \alpha = 0,05$

	S	df	MS	F-value
COUTRY	170,16	3-1	85,08	28,65
ERROR	26,75	12-3	2,972	
TOTAL	196,917	12-1	17,9	

$F_{0,95}(3-1; 12-3) = 4,2565$

$W_\alpha = \langle 0; 4,2565 \rangle$

F-value $\notin W_{0,05} \Rightarrow H_0$ ZAM.

$$\frac{(\bar{y}_{\cdot j} - \bar{y}_{\cdot k})^2}{k-1} \Big/ s_{RES}^2 \cdot \left(\frac{1}{n_j} + \frac{1}{n_k}\right)$$

	B	C
A	27,41	12,36
B		5,05

$s_{RES}^2 = MS_e$

$\mu_A \neq \mu_B$

$\mu_A \neq \mu_C$

$\mu_B \neq \mu_C$

$RES_i = y_i - \bar{y}_{\cdot j}$

$$\sum_{i=1}^{n} RES_i^2 = -5,625$$

$$\sum_{i=1}^{n} RES_i^4 = 101,33$$

$$SKEW = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} RES_i^3}{\left(\frac{1}{n} S_e\right)^{\frac{3}{2}}} = -0,14$$

$$KURT = \frac{\frac{1}{n} \sum RES_i^4}{\left(\frac{1}{n} S_e\right)^2} - 3 = -1,3$$

$$JB = \frac{n}{6}\left[SKEW^2 + \frac{1}{4} KURT^2\right] = 0,8856$$

$JB \sim \chi^2(2)$ $\quad W_{0,05} = \langle 0; \chi^2_{0,95}(2) \rangle = \langle 0; 5,991 \rangle$ $\quad JB \in W_{0,05} \Rightarrow$ NEZ.

$$S^2_{RES} = MS_e = 2{,}972$$

$$\frac{1}{n_j - 1}\sum_{i=1}^{n_j}(y_i - \bar{y}_j)^2 = S_j^2 = \frac{1}{n_j - 1}\left(\sum_{i=1}^{n_j} y_i^2 - n_j \bar{y}_j^2\right) \#\#$$

$$S_A^2 = 2{,}917$$

$$S_B^2 = 4$$

$$S_C^2 = 2{,}5$$

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2$$

$$\overline{W}_{0,05} = \langle 0; 5{,}991 \rangle$$

$$B \sim \chi^2(k-1)$$

$$C = 1 + \frac{1}{3(k-1)}\left[\sum_{j=1}^{k}\frac{1}{n_j - 1} - \frac{1}{n-k}\right] = 1{,}162$$

$$B = \frac{1}{C}\left[(n-k)\cdot \ln(S^2_{RES}) - \sum_{j=1}^{k}(n_j - 1)\cdot \ln(S_j^2)\right] \doteq 0{,}133$$

$$B \in \overline{W}_{0,05} \Rightarrow \text{NEZ } H_0$$

Gather 6 observations of adult human weight (y) and height (x). For $\alpha = 0,05$ and using these observations

- find estimates for β_1 and β_2 for regression line $y = \beta_1 + \beta_2 x$
- test $H_0 : \beta_2 = 0$ against $\beta_2 \neq 0$
- find a CI for β_1
- compute R-squared
- compute leverage for the height farthest from the mean height
- estimate the value of weight for height $x_0 = 180$ cm
- compute CI and PI for your predicted value

	x	y	x^2	y^2	xy
1	181	80	32761	6400	14480
2	175	70	30625	4900	12250
3	190	100	36100	10000	19000
4	191	105	36481	11025	20055
5	185	84	34225	7056	15540
6	165	60	27225	3600	9900
Σ	1087	499	197417	42981	91225

$\bar{x} = 181,1\bar{6}$

$\bar{y} = 83,1\bar{6}$

①

$$\det(X^TX) = n \cdot \Sigma x_i^2 - (\Sigma x_i)^2 = 2933$$

$$b_2 = \frac{n \Sigma x_i y_i}{\det(X^TX)} = 1,683$$

$$b_1 = \bar{y} - b_2 \bar{x} = -221,732$$

⑥

$$\hat{y} = -221,732 + 1,683x$$

$$\hat{y}(180) = -221,732 + 1,683 \cdot 180 = 81,208$$

$H_0: \beta_2 = 0$

$H_A: \beta_2 \neq 0$

$$t = \frac{b_2 - \beta_{20}}{s_{res}\sqrt{g^{22}}} \sim t(6-2)$$

$$t = 7,693$$

$G = (X^TX)^{-1}$

$$g_{11} = \frac{\Sigma x_i^2}{\det(X^TX)} = 67,389$$

$$g_{22} = \frac{n}{\det(X^TX)} = 0,002045$$

$$S_e = \Sigma y_i^2 - b_1 \Sigma y_i - b_2 \Sigma x_i y_i = 93,593$$

$$\Sigma (y_i - b_1 - b_2 x_i)^2$$

$$s_{res}^2 = \frac{S_e}{n-h} = \frac{S_e}{6-2} = 23,398$$

$$s_{res} = 4,837$$

② $\overline{W}_{0,05} = \langle -t_{0,975}(11) ; t_{0,975}(11) \rangle$

$\langle -2,776 ; 2,776$

$t \notin \overline{W} \Rightarrow H_0$ zamítáme

③

$\beta_1 \in \langle b_1 - t_{1-\frac{\alpha}{2}}(n-h) \cdot S_{res} \cdot \sqrt{g_{11}} ; b_1 + t_{1-\frac{\alpha}{2}}(n-h) \cdot S_{res} \sqrt{g_{11}} \rangle$

$\beta_1 \in \langle -311.96 ; -111.504 \rangle$

④ $R_2 = 1 - \frac{S_e}{\Sigma y_i^2 - n\bar{y}^2} = 0,937$

⑤ $h_{66} = \frac{1}{n} + \frac{(x_6 - \bar{x})^2}{\Sigma x_i^2 - n\bar{x}^2} = 0,1685$

$x_0 = 180$

$(0,16934)$

⑦ CI $x_0 = 180$

$y(180) \langle \hat{y}(180) - t_{1-\frac{\alpha}{2}}(n-h) s_{res} \sqrt{d(x_0)} ; \hat{y}(180) + t_{1-\frac{\alpha}{2}}(n-h) s_{res} \sqrt{d(x_0)} \rangle$

$y(180) \in \langle 75,682 ; 86.734 \rangle$

$d(x_0) = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\Sigma x_i^2 - n\bar{x}^2} = 0,16934$

PI

$x_0 = 180$

$y(180) \langle \hat{y}(180) - t_{1-\frac{\alpha}{2}}(n-h) s_{res} \sqrt{1+d(x_0)} ; \hat{y}(180) + t_{1-\frac{\alpha}{2}}(n-h) s_{res} \sqrt{1+d(x_0)} \rangle$

$y(180) \in \langle 66.688 ; 95.727 \rangle$

Box-Cox transformation is defined by the following equation:

$$y^*(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \ln(y), & \text{if } \lambda = 0 \end{cases}$$

- prove that it is continuous with respect to λ
- write down log-likelihood function and its partial derivative with respect to λ for regression line model driven random variable Y.

Důkaz spojitosti

$$\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \ln(y)$$

dosadit $\lambda = 0$

$$\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \left[\frac{0}{0}\right] = \lim_{\lambda \to 0} \frac{y^\lambda \cdot \ln(y)}{1} = \ln(y)$$

$$L(\sigma, \lambda, \beta_1, \beta_2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{\left(\frac{y_i^\lambda - 1}{\lambda} - (\beta_1 + \beta_2 y_i)\right)^2}{2\sigma^2}} \cdot y_i^{\lambda - 1} =$$

$$y^* = \frac{y^\lambda - 1}{\lambda}$$

Transformace nezachovává PDF = 1 $\int f(x) = 1$

$\Rightarrow$ substituce a oprava

$$dy^* = \frac{\lambda y^{\lambda-1}}{\lambda} dy$$

$$dy^* = y^{\lambda-1} dy$$

$$= -n \ln\left(\sqrt{2\pi}\sigma\right) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(\frac{y_i^\lambda - 1}{\lambda} - (\beta_1 + \beta_2 y_i)\right)^2 + (\lambda - 1) \sum_{i=1}^{n} \ln(y_i)$$

$$\frac{dL}{d\lambda} = \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2\left(\frac{y_i^\lambda - 1}{\lambda} - (\beta_1 + \beta_2 y_i)\right) \cdot \frac{y_i^\lambda \cdot \ln(y_i) \cdot \lambda - (y_i^\lambda - 1) \cdot 1}{\lambda^2} + \sum_{i=1}^{n} \ln(y_i)$$

You successfully fitted a quadratic regression model through a dataset with $n = 100$ observations for predictor values $\in (-1; 1)$. Assume model $y = \beta_1 + \beta_2 x + \beta_3 x^2$, you obtained 95% CI for all coefficients $\beta_1 \in (9{,}762;\ 10{,}327)$, $\beta_2 \in (1{,}636;\ 2{,}289)$, $\beta_3 \in (4{,}007;\ 5{,}272)$ and residual standard deviation $s_{res} = 0,950093$. $\sum_{i=1}^{n}(y_i - \bar{y})^2 = 407{,}1588$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0{,}022504 & 0 & -0{,}03752 \\ 0 & 0{,}030003 & 0 \\ -0{,}03752 & 0 & 0{,}112556 \end{bmatrix}$$

a

- Compute R^2 and test whether all terms in regression model are statistically significant. b
- For observations

c $x = [-0{,}65; 0{,}53; 0{,}89]$; $y = [-8{,}532; 14{,}431; 13{,}265]$ compute Leverages, Standardized and Studentized residuals and Cook's distance.

d
- Compute 95% confidence intervals for prediction and expected value of a model for $x_{00} = 0$ and $x_0 = 0{,}53$

(a) $R^2 = 1 - \frac{S_e}{\sum_{i=1}^{n}(y-\bar{y})^2} = 1 - \frac{}{407,1788} = 0,7849$

$S_{res}^2 = \frac{S_e}{n-k} \Rightarrow S_e = S_{res}^2 \cdot (100-3) = 87,5596$

(b) $H_0: \beta_i = 0 \quad H_A: \beta_i \neq 0$

$t = \frac{\beta_i}{s_{res}\sqrt{g^{ii}}}$

$\beta_1: b_1 = 10,0445 \quad t = \frac{\frac{9,762+10,327}{2}}{s_{res}\sqrt{g^{11}}} = 70,475$

$\beta_2: b_2 = 1,9625 \quad t = 11,925$

$\beta_3: b_3 = 4,6395 \quad t = 14,555$

$t = \frac{(0,1,0)\cdot\begin{pmatrix}b_1\\b_2\\b_3\end{pmatrix} - (0,1,0)\begin{pmatrix}b_1\\0\\b_3\end{pmatrix}}{s_{res}\sqrt{(0,1,0)(X^TX)^{-1}\begin{pmatrix}0\\1\\0\end{pmatrix}}}$

$\bar{W}_{0,05} = \langle -t_{0,975}(97); t_{0,975}(97)\rangle$

$\langle -1,985 ; 1,985\rangle$

$\Rightarrow \beta_i \notin \bar{W}$

$\Rightarrow$ vše se významně liší od 0

= všechny β_i jsou významné

(c) $x = -0,65 \quad y = -8,532$

$H = X(X^TX)^{-1}X^T \qquad X = \begin{pmatrix}1 & x_1 & x_1^2\\ \vdots & & \\ 1 & x_n & x_n^2\end{pmatrix}$

$h_{ii} = ?$ Leverage

$h_{ii} = (1; -0,65; 0,65^2)(X^TX)^{-1}\begin{pmatrix}1\\-0,65\\-0,65^2\end{pmatrix} =$

$= (0,0066; -0,0195; 0,01)\begin{pmatrix}1\\-0,65\\-0,65^2\end{pmatrix} = 0,0235 = h_{ii}$

$D_i = \frac{e_i^2}{k s_{res}^2} \cdot \frac{h_{ii}}{(1-h_{ii})^2} = 3,376$

$e_i = y_i - b_1 - b_2 x_i - b_3 x_i^2 = -19,2629 \qquad k = 3$

$$r_i = \frac{e_i}{\sqrt{s^2_{res}(1-h_{ii})}} = -20{,}52$$

$$r_{(i)} = e_i \sqrt{\frac{(n-k)\cdot e_i}{S_e(1-h_{ii})-e_i^2}} = -43{,}28$$!

↳ vychází < 0 ⇒ nejde ze zadaných hodnot

(d) $CI = \left(\hat{y}(x_0) - s_{res} t_{1-\frac{\alpha}{2}}(n-k)\sqrt{d^2(x_0)} \; ; \; \hat{y}(x_0) + \dots\right)$

$PI = \left(\hat{y}(x_0) - s_{res} t_{1-\frac{\alpha}{2}}(n-k)\sqrt{d^2(x_0)+1} \; ; \; \hat{y}(x_0) + \dots\right)$

$x_0 = 0$

$\hat{y}(0) = 10{,}0445 \qquad t_{0{,}975}(97) = 1{,}985$

$$d^2(x_0) = (1,0,0)(X^TX)^{-1}\begin{pmatrix}1\\0\\0\end{pmatrix} = 0{,}022504$$

$CI = \langle 9{,}76 \; ; \; 10{,}327 \rangle$

$PI = \langle 8{,}137 \; ; \; 11{,}95 \rangle$

$x_0 = 0{,}53$

$\hat{y}(0{,}53) = 12{,}3879$

$$d^2(0{,}53) = (1, 0{,}53, 0{,}53^2)(X^TX)^{-1}\begin{pmatrix}1\\0{,}53\\0{,}53^2\end{pmatrix} = 0{,}019$$

$$= (0{,}0119 \; ; \; 0{,}0159 \; ; \; -0{,}0059)\begin{pmatrix}1\\0{,}53\\0{,}53^2\end{pmatrix} = 0{,}0186$$

You collected survey data aiming to prove a connection between self-reported patriotism and political support for Andrej Danko in Slovakia. Responses are summarized in following frequency table:

Patriotic? \ Danko?	Yes	No
Yes	53	42
Don't know	28	41
No	10	40

Test ($\alpha = 0,05$) whether responses associated with being patriotic are independent on the support for Andrej Danko

f_i $C = 2$

X_i	Yes	No	
Yes	53	42	95
Dn	28	41	69
No	10	40	50
	91	123	214

$R = 3$

$\hat{f}_i$

E_i	Y	N	
Y	40.4	54.6	95
D	29.3	39.7	69
N	21.3	28.7	50
	91	123	214

	3.93	2.9	6.83
	0.06	0.04	0.1
	6	4.45	10.45
	9.99	7.39	17.38 = t

$$\frac{(f_i - \hat{f}_i)^2}{\hat{f}_i}$$

$$\bar{W}_{0.95} = \langle 0; \chi^2((R-1)(C-1)) \rangle$$
$$= \langle 0; \chi^2(2 \cdot 1) \rangle = \langle 0; 5.9914 \rangle$$

$t \notin \bar{W}_\alpha \Rightarrow$ zamítáme H_0
(nezávislost)

For low sample sizes (expected frequencies < 5) and 2x2 table (so no further aggregation is possible) you can use Hypergeometric probabilities and Fisher's exact test. Suppose you want to find out whether some TV show will be polarizing to different age groups of audiences and you have just observations out of representative "test audience" (in following Frequency table)

Age \ Liked show	Yes	No
12-18	5	3
30-45	2	7

Derive some form of Fisher's exact test and test whether age and liking a show are independent (at $\alpha \leq 0,1$)

	Y	N	
12-18	5	3	8
30-45	2	7	9
	7	10	17

	Y	N	
			8
			9
7	10	17	

$X \sim H(17, 8, 7)$

$x = 0$: 0 z 8, 7 z 2

$p(X=0) = p(0) = \frac{\binom{8}{0}\binom{9}{7}}{\binom{17}{7}} = 0.002$

$x = 1$: 1 z 7, 6 z 3

$p(x=1) = p(1) = \frac{\binom{8}{1}\binom{9}{6}}{\binom{17}{7}} = 0.035$

$p(2) = 0.181$

$p(3) = 0.369$

$p(4) = 0.302$

$p(5) = 0.104$

$p(6) = 0.013$

$p(7) = 0.0004$

$\alpha \leq 0,1$

podle rostoucí p.

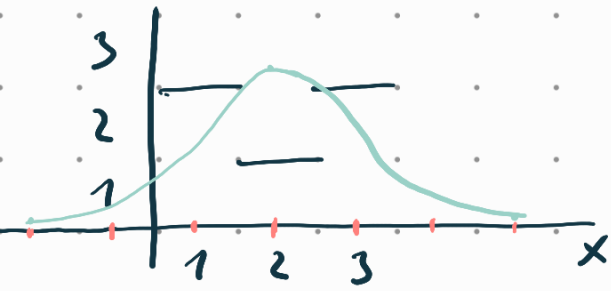
$\{7, 0, 6, 1\} = W_{0,1}$

0.04477... skutečné α

$\bar{W}_{0,1} = \{2, 3, 4, 5\}$

$5 \in \bar{W}_{0,1} \Rightarrow$ nezamítám H_0

Using IID observations $1; 2; 3; 1; 3$ of some random variable X test, whether $X \sim N(\mu, \sigma^2)$. Use Lilliefors test with significance level $\alpha = 0,05$



$\bar{x} = 2$

$s(x)^2 = \frac{1}{4}(1+1+1+1) = 1$

$x \to u$

$1 \to \frac{1-\bar{x}}{s^2} = -1$

$2 \to 0$

$3 \to 1$

u_i ; u_{i+1}	$F(u)$	$\Phi(u_i)$	$\Phi(u_{i+1})$	D_i^-	D_i^+
$(-\infty ; 1)$	$\frac{0}{5} = 0$	0	0.1587	0	0.1587
$(1 ; 2)$	$\frac{2}{5} = 0.4$	0.1587	0.5	0.2413	0.1
$(2 ; 3)$	$\frac{3}{5} = 0.6$	0.5	0.8413	0.1	0.2413
$(3 ; \infty)$	$\frac{5}{5} = 1$	0.8413	1	0.1587	0

$\max\{D_i^- ; D_i^+\} = 0.2413 = t$

$\bar{W}_\alpha = \langle 0 ; 0.3427 \rangle$ $\quad t \in \bar{W} \Rightarrow$ nezamítáme H_0

(že výběr pochází z N rozd.)

$H_0: x_{0.5} = 75 = c$ počet + = 3 $n = 10$

$d_i = x_i - c$ $\operatorname{sign}(d_i)$ $B: (10; 0.5)$

$X(0) = 0.000\,97 = \binom{10}{0} 0.5^0 0.5^{10}$

1 0.0097

2 0.0439

⋮ ⋮

8 0.0439

9 0.0097

10 0.000 97

You obtained final results (points) of some exam at BUT: ~~98,7~~; ~~63,7~~; ~~81,3~~; ~~63,0~~; 68,3; ~~62,0~~; ~~83,9~~; ~~63,0~~; ~~74,3~~; 64,1. Test whether the median of an underlying random variable is equal to 75 at $\alpha \leq 0,05$.

$W_{0.05} = \{0, 10, 1, 9\}$

$t = 3$ $t \in \overline{W}_{0.05} \Rightarrow H_0$ nezamítám

You want to compare median exam scores of groups with different lecturers of the same course. For this purpose you obtained 20 exam scores for each lecturer (sorted values are in the table bellow)

X	**Lecturer 1**									
	60,2	60,6	61,1	62,4	63,3	63,5	63,7	64,5	65,0	66,4
	66,5	67,3	68,2	69,2	73,1	73,4	73,5	76,8	80,7	83,1
	Lecturer 2									
Y	52,1	52,2	52,7	54,3	55,1	55,6	56,0	56,3	58,1	58,1
	60,7	66,3	67,2	68,0	76,0	79,2	85,1	86,9	98,0	100,0

Use Mood's median test to test $H_0 : X_{0,5} = Y_{0,5}$ at $\alpha = 0,05$

$X_{0.5} = 66.45$

$Y_{0.5} = 59.4$

$z_{0.5}$ median z obou souborů

$z_{0.5} = 65.65$

spočítaná

	1	2	
pod	9	11	20
nad	11	9	20
	20	20	40

f_j

očekávaná

10	10	20
10	10	20
20	20	40

$\hat{f}_j$

$$\sum \frac{(f_j - \hat{f}_j)^2}{f_j} = \frac{4}{10} = t$$

$$W_{0.05} = \langle 0; \chi_{0.95}(1) \rangle = \langle 0; 3.841 \rangle$$

$t \in W_\alpha \Rightarrow$ H_0 nezamítáme

You want to test whether therapy sessions decrease stress levels of patients. For this purpose you obtained 10 measurements of stress on the scale from 0 to 100. Assuming symmetrical (but not normal) probability distribution (of differences), find out if stress levels change after therapy. Use $\alpha = 0,05$

Patient	1	2	3	4	5	6	7	8	9	10
Before	75	68	83	90	70	98	72	99	75	98
After	68	70	60	95	51	90	75	88	60	80
$B_i - A_i$	7	-2	23	-5	19	8	-3	11	15	18

$H_0: x_{0.5} = c = 0$ $\alpha = 0.05$

$H_A: x_{0.5} \neq 0$

2	3	5	7	8	11	15	18	19	23
1	2	3	4	5	6	7	8	9	10

Σ pořadí, které náleží $d_i > 0$

$S^+ = 49$ 8 je hodnota z tabulky

$\bar{W}_\alpha = (8; \frac{n(n+1)}{2} - 8) = (8; 47)$

$S^+ \notin \bar{W}_\alpha \Rightarrow$ zamítáme H_0

došlo ke změně hl. stresu

You want to compare monthly income similarly employed people in 2 different regions. Since you can't account for different occupations, you expect the underlying probability distributions to be heavily skewed. Using $\alpha = 0,05$ and following data [1000 Kč], test whether the median income is the same.

X	73	11.5	4	1.5	5	16	17	7	8	3	
a	R 1	43,6	41,2	40,7	41,6	47,2	49,0	42,1	42,2	40,8	
10	R 2	43,7	49,5	43,1	43,4	45,2	47,0	43,6	40,7	41,8	55,2
Y	106	13	18	9	10	14	15	11.5	1.5	6	19

$H_0: X_{0.5} = Y_{0.5}$

$H_A: X_{0.5} \neq Y_{0.5}$

- Data dohromady => velký soubor
- u spojitých dat
- Suma pořadí pro prvky z X

$$S^X = 73$$

$$U_x = n_x n_y + \frac{n_x(n_x+1)}{2} - S_x = 90 + 45 - 73 = 135 - 73 = \underline{62}$$

$$\min\{U_x ; \underbrace{n_x(n_x+n_y+1)}_{9\ 9\ 10} - U_x\} = \min\{62; \underbrace{180}_{-62} - 62\} = \min\{62; 118\} = \underline{62} = t$$

$$\bar{W}_\alpha = (20; \text{pravá hranice} - 20)$$

$$t \in \bar{W}_\alpha \Rightarrow \underline{H_0 \text{ nezamítáme}}$$

Test whether there is some monotonous relation between the amount of points obtained before the exam and during the exam of some course taught at BUT. Use $\alpha = 0,05$ and following observations.

		10	8	2	5	4	6	7	9	3	1
R_x	before	40	35	21	31,5	28,5	33,5	34	38	25	20
R_y	during	60	51	42	40	43,5	47	52	58,5	55	40
		10	6	3	1.5	4	5	7	9	8	1.5

$n = 10$

$\bar{R}_x = \frac{n(n+1)}{2n} = \frac{11}{2} = 5.5 = \bar{R}_y$ $\quad 10 \cdot 5.5 \cdot 5.5 = 302.5$

$\bar{R} - R$

4.5 2.5 3.5 0.5 1.5 0.5 1.5 3.5 2.5 4.5

4.5 0.5 2.5 4 1.5 0.5 1.5 3.5 2.5 4

$s(R_x) = 3,0277$

$s(R_y) = 3,0185$

$r(R_x, R_y) = 0.7356$

$t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} = 3,07 \Rightarrow t \notin \overline{W_\alpha}$ $\Rightarrow$ zamít. H_0

$\overline{W_\alpha} = \langle 0, t_{1-\frac{\alpha}{2}}(n-1) \rangle = (0, 2.306)$

BONUS

Logistic regression model