Opowieści statystyczne

Paweł Czyż

Zadania w formie interaktywnego zeszytu ćwiczeń dostępne są pod tym adresem.

Zadania można wysyłać przez portal w następującym formacie:

- (Opcja rekomendowana) Wszystkie zadania można wykonać w interaktywnym zeszycie ćwiczeń. Rozwiązania można wówczas wydrukować do pliku PDF ("Plik" oraz "Drukuj" lub skrót klawiszowy "Ctrl + P").
- (Opcja mniej polecana) PDF powstały z wydrukowania zadania trzeciego, jak opisano powyżej, oraz rozwiązań pozostałych zadań (np. napisanych w LaTeXu, Markdown lub zeskanowanych rozwiązań napisanych czytelnie na kartce). Chociaż zachęcam do myślenia nad zadaniami przy kartce papieru (zamiast komputera), końcowe rozwiązania warto przepisać do LaTeXa lub Markdown, by wyeliminować możliwość tego, że nie dam rady przeczytać innego stylu pisma niż swój własny.

Zadanie 1: Trzy pytania do Andrzeja [10 punktów]

Andrzej został wezwany do tablicy, gdzie dostanie trzy pytania z rachunku prawdopodobieństwa. Na każde z nich może odpowiedzieć poprawnie (co oznaczamy jako $X_n = 1$ dla pytania numer n) lub niepoprawnie (czyli $X_n = 0$).

Na pierwsze pytanie Andrzej odpowie poprawnie z prawdopodobieństwem $s = 0.8$, co zapisujemy jako

$$P(X_1 = x_1) = \text{Bernoulli}(x_1 \mid s) = s^{x_1}(1-s)^{1-x_1}, \quad x_1 \in \{0, 1\}.$$

Odpowiedź Andrzeja na drugie pytanie zależy od tego, jak poszło mu w pierwszym pytaniu. Zachodzi:

$$P(X_2 = x_2 \mid X_1 = x_1) = w\,\mathbf{1}[x_2 = x_1] + (1-w)\,\text{Bernoulli}(x_2 \mid s),$$

gdzie $w = 0.6$, a funkcja $\mathbf{1}[\phi]$ przyjmuje wartość 1 gdy wyrażenie ϕ jest prawdziwe oraz wartość 0 gdy wyrażenie ϕ jest fałszywe.

Innymi słowy, z prawdopodobieństwem w Andrzej udzieli odpowiedzi tak samo poprawnej jak na pierwsze pytanie. Natomiast jest też możliwość (z prawdopodobieństwem $1-w$), że podejdzie do drugiego pytania na chłodno i odpowie na nie poprawnie z prawdopodobieństwem s.

Odpowiedź Andrzeja na trzecie pytanie zależy od tego jak odpowiedział na drugie pytanie. Natomiast, nie zależy bezpośrednio od tego jak sobie poradził w pytaniu pierwszym:

$$\begin{aligned} P(X_3 = x_3 \mid X_2 = x_2, X_1 = x_1) &= P(X_3 = x_3 \mid X_2 = x_2) \\ &= w\,\mathbf{1}[x_3 = x_2] + (1-w)\,\text{Bernoulli}(x_3 \mid s). \end{aligned}$$

Model ten nazywamy łańcuchem Markowa i często oznaczamy jako $X_1 \to X_2 \to X_3$, co wskazuje na to, że zmienne X_1 i X_3 stają się niezależne, gdy znamy wartość zmiennej X_2.

a) Jakie jest prawdopodobieństwo, że Andrzej udzieli poprawnej odpowiedzi na wszystkie trzy pytania? Innymi słowy, jakie jest $P(X_1 + X_2 + X_3 = 3) = P(X_1 = 1, X_2 = 1, X_3 = 1)$?
b) Jakie jest prawdopodobieństwo na to, że Andrzej udzieli poprawnej odpowiedzi na trzecie pytanie, jeśli udzielił poprawnej odpowiedzi na pierwsze pytanie? Innymi słowy, jakie jest $P(X_3 = 1 \mid X_1 = 1)$? [Podpowiedź: to liczba różna od s: w tym pytaniu nie znamy wartości X_2.]
c) Jakie jest prawdopodobieństwo na to, że Andrzej odpowie poprawnie na dokładnie jedno pytanie, to znaczy $P(X_1 + X_2 + X_3 = 1)$?
d) Opisz w jaki sposób można zweryfikować powyższe odpowiedzi mając dostęp do urządzeń pozwalających na szybkie losowanie wielu liczb (jak kostki, monety czy komputer)?

[Podpowiedź: Można wypisać tabelkę $P(X_1 = x_1, X_2 = x_2, X_3 = x_3)$.]

[*Miejsce na Twoje odpowiedzi.*]

Zadanie 2: Jak szybko tupta jeż? [10 punktów]

Mamy do dyspozycji Specjalny Miernik Prędkości, który pozwala nam na mierzenie prędkości zbliżającego się do nas zwierzęcia, wyrażonej w kilometrach na godzinę. Jeśli wskazanie miernika jest ujemne, oznacza to, że zwierzę się od nas oddala.

Jeśli obiekt zbliża się do nas z prędkością μ, to zastosowanie Specjalnego Miernika Prędkości zwróci oszacowanie X z rozkładu normalnego zadanego gęstością prawdopodobieństwa

$$\text{Normal}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

gdzie parametr $\sigma = 10$ km/h jest znanym odchyleniem standardowym rozkładu, opisującym jak bardzo może mylić się miernik. (Zauważ, że średnią rozkładu jest μ).

Wyobraźmy sobie populację zwierząt, taką że losowo wybrane zwierzę zbliża się do nas z prędkością μ wylosowaną z rozkładu normalnego $\text{Normal}(\mu \mid 0, \tau^2)$, gdzie τ określa pewną prędkość charakterystyczną danego gatunku: ponad 99% zwierząt z tej populacji ma prędkość μ między -3τ a 3τ.

Rozważmy model, w którym:

1. Najpierw losujemy prędkość zwierzęcia z powyższego rozkładu, co oznaczamy jako $\mu \sim \text{Normal}(0, \tau^2)$.
2. Następnie korzystamy ze Specjalnego Miernika Prędkości by dostać oszacowanie $X \mid \mu \sim \text{Normal}(\mu, \sigma^2)$.

Powyższy model oznacza, że gęstość prawdopodobieństwa wartości pomiaru oraz prawdziwej prędkości jest równa

$$p(x, \mu \mid \sigma^2, \tau^2) = \text{Normal}(\mu \mid 0, \tau^2) \cdot \text{Normal}(x \mid \mu, \sigma^2).$$

Będziemy korzystać z tego, że warunkowa gęstość prawdopodobieństwa samego wyniku pomiaru jest zadana poprzez

$$p(x \mid \sigma^2, \tau^2) = \text{Normal}(x \mid 0, \sigma^2 + \tau^2).$$

(Nie ma obowiązku dowodzenia tego wzoru – można założyć, że tak po prostu jest.)

a) Pokaż, że warunkowa gęstość prawdopodobieństwa, zdefiniowana jako

$$p(\mu \mid x, \sigma^2, \tau^2) := \frac{p(x, \mu \mid \sigma^2, \tau^2)}{p(x \mid \sigma^2, \tau^2)},$$

jest tożsama z rozkładem normalnym $\text{Normal}\big(\mu \mid \frac{x}{1+r}, \frac{\sigma^2}{1+r}\big)$, gdzie $r = \sigma^2/\tau^2$.

b) Widzimy biegnącego geparda. Jeśli $\tau = 40$ km/h dla gepardów oraz pomiar wskazał $x = 70$ km/h, to jak oszacujesz prawdziwą prędkość geparda? (Pamiętaj, że $\sigma = 10$ km/h.)

c) Widzimy tuptającego jeża. Jeśli $\tau = 0.5$ km/h dla jeży, a pomiar wskazał $x = 70$ km/h, to jak oszacujesz prawdziwą prędkość jeża?

d) Widzimy innego tuptającego jeża. Pomiar wskazał $x = -0.25$ km/h. Jak sądzisz, czy jeż się od nas oddala?

[*Uwaga matematyczna, którą można pominąć i która nie jest potrzebna do rozwiązania zadania.* Powyższe wyrażenie na $p(x \mid \sigma^2, \tau^2)$ wynika z równości:

$$p(x \mid \sigma^2, \tau^2) = \int_{-\infty}^{+\infty} p(x, \mu \mid \sigma^2, \tau^2)\, \mathrm{d}\mu = \text{Normal}(x \mid 0, \sigma^2 + \tau^2).$$

Jest to szczególny przypadek tożsamości podanej tutaj. Można też ją uzasadnić zapisując zmienną X w postaci $X = \mu + \epsilon$, gdzie zmienne $\mu \sim \text{Normal}(0, \tau^2)$ i $\epsilon \sim \text{Normal}(0, \sigma^2)$ są niezależne.]

[*Miejsce na Twoje odpowiedzi.*]

Zadanie 3: Czy to lekarstwo szkodzi? [10 punktów]

[*Uwaga: Poniższe zadanie jest podzielone na części (a)–(g) oraz wymaga odrobiny programowania w Pythonie. Można korzystać ze wszelkich źródeł dostępnych w internecie i chatbotów, w celu nauczenia się składni: ten poziom Pythona będzie naszym punktem wyjściowym na warsztatach, gdzie będziemy uczyć się nieco trudniejszych rzeczy.*

Wysyłając rozwiązanie, skorzystaj z interaktywnego zeszytu ćwiczeń dostępnego pod tym adresem i upewnij się, że widoczne są kod, rozumowanie jak i wykresy.]

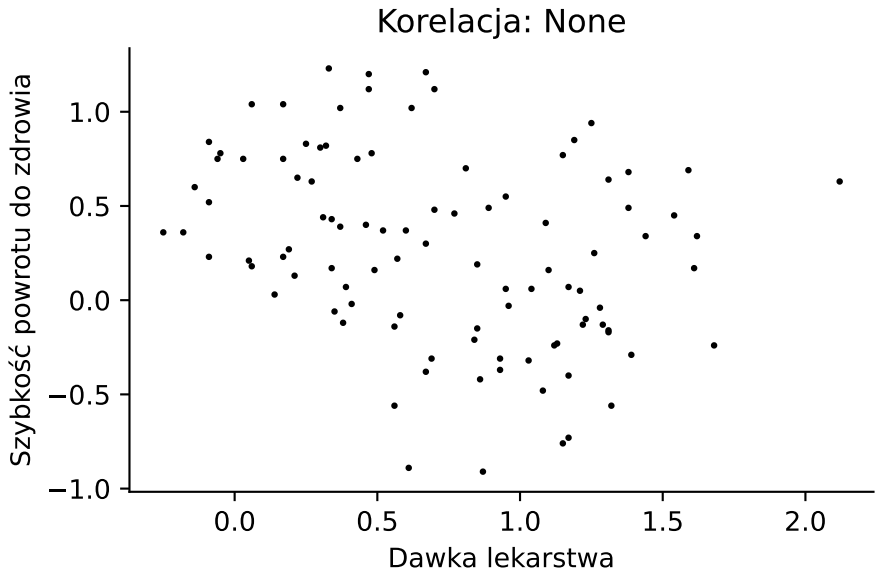
[*Uwaga: Historia jest wymyślona, a dane zostały zasymulowane. Nie stanowią porady lekarskiej ani weterynaryjnej: do tego kompetencje mają lekarze, a nie statystycy!*]

Weterynarz Bajtocjusz przez rok notował w swoim zeszycie dawki lekarstwa podawanego chorym króliczkom (im większa wartość, tym więcej lekarstwa zostało podane) oraz to jak szybko króliczki wróciły do stanu zdrowia (im większa wartość, tym szybciej króliczek wyzdrowiał).

Część (a): napisz funkcję w Pythonie, która obliczy korelację. [1 punkt]

```
import numpy as np
import matplotlib.pyplot as plt

def correlation(x: list[float], y: list[float]) -> float:
  return None  # Zadanie: zastąp te linijki swoim kodem. [Podpowiedź: spójrz na dokumentac


lek = [1.28, 0.39, 1.17, 1.44, 0.03, 1.68, 0.86, 1.12, -0.09, 0.57, 0.81, 0.87, 1.1, 1.31,
zdrowienie = [-0.04, 0.07, -0.4, 0.34, 0.75, -0.24, -0.42, -0.24, 0.52, 0.22, 0.7, -0.91,

korelacja = correlation(lek, zdrowienie)

fig, ax = plt.subplots(figsize=(5, 3), dpi=120)

ax.scatter(lek, zdrowienie, s=2, c='k')
ax.set_xlabel("Dawka lekarstwa")
ax.set_ylabel("Szybkość powrotu do zdrowia")
ax.set_title(f"Korelacja: {korelacja}")
```

```
ax.spines[["top", "right"]].set_visible(False)
```

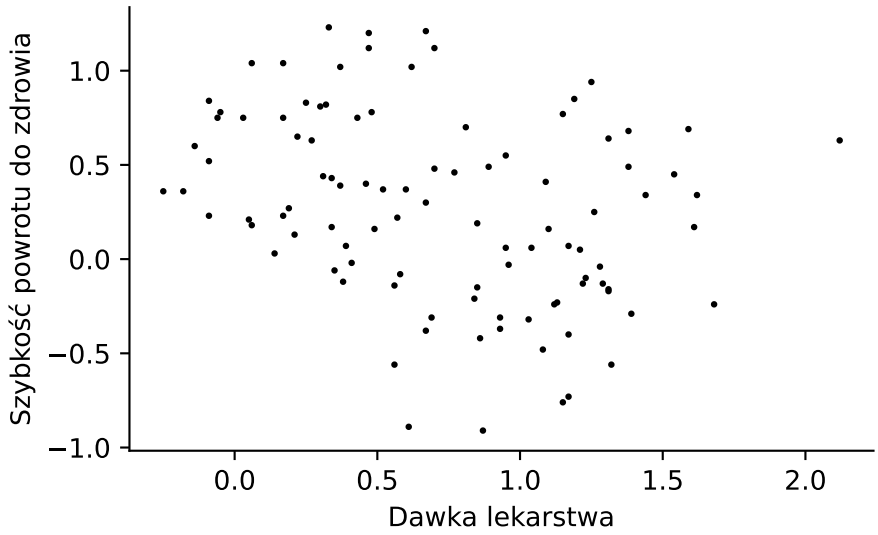


Część (b): Czy korelacja jest dodatnia czy ujemna? Jak ją zinterpretować? [1 punkt]

[*Miejsce na Twoją odpowiedź.*]

Część (c): Na wykresie narysuj prostą najlepszego dopasowania. [2 punkty]

```
fig, ax = plt.subplots(figsize=(5, 3), dpi=120)

ax.scatter(lek, zdrowienie, s=2, c='k')
ax.set_xlabel("Dawka lekarstwa")
ax.set_ylabel("Szybkość powrotu do zdrowia")
ax.spines[["top", "right"]].set_visible(False)


# Zadanie: to jest miejsce na Twój kod :)
# [Podpowiedź: potrzebny wzór można znaleźć na przykład tutaj:
# https://jianghaochu.github.io/ordinary-least-squares-regression-in-python-from-scratch.h
```



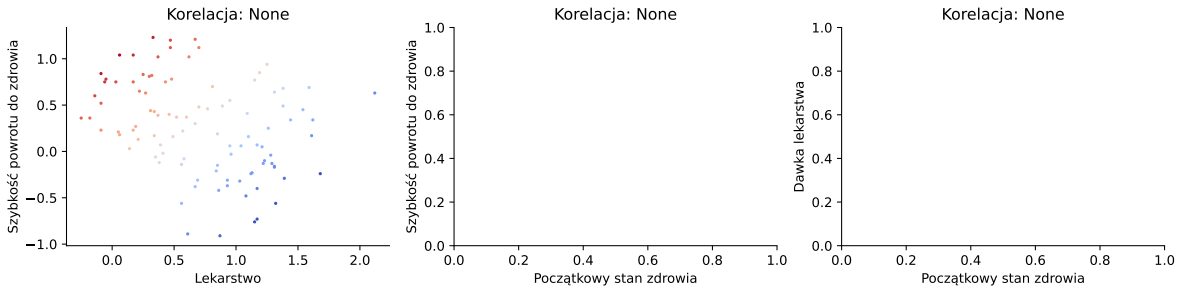
Bajtocjusz nie mógł zrozumieć powyższego wyniku. Spojrzał do swojego notatnika i zauważył, że notował początkowy stan zdrowia każdego króliczka.

Część (d): Narysuj wykres reprezentujący początkowy stan zdrowia wraz z szybkością powrotu do zdrowia oraz wykres reprezentujący początkowy stan zdrowia i przepisaną dawkę leku. [2 punkty]

```
stan = [0.68, 1.68, 0.42, 0.91, 2.72, 0.07, 0.72, 0.64, 2.62, 1.65, 1.89, 0.22, 1.07, 0.53

fig, axs = plt.subplots(1, 3, figsize=(5*3, 3), dpi=120)
for ax in axs:
  ax.spines[["top", "right"]].set_visible(False)

ax = axs[0]
ax.scatter(lek, zdrowienie, s=2, c=stan, cmap="coolwarm")
ax.set_xlabel("Lekarstwo")
ax.set_ylabel("Szybkość powrotu do zdrowia")
ax.set_title(f"Korelacja: {correlation(lek, zdrowienie)}")

ax = axs[1]
ax.set_xlabel("Początkowy stan zdrowia")
ax.set_ylabel("Szybkość powrotu do zdrowia")
ax.set_title(f"Korelacja: {correlation(stan, zdrowienie)}")

# Zadanie: to miejsce na Twój kod :)
```

```
ax = axs[2]
ax.set_xlabel("Początkowy stan zdrowia")
ax.set_ylabel("Dawka lekarstwa")
ax.set_title(f"Korelacja: {correlation(stan, lek)}")


# Zadanie: to miejsce na Twój kod :)
```

```
Text(0.5, 1.0, 'Korelacja: None')
```



Bajtocjusz zauważył, że najciężej chore króliczki dostawały najwięcej lekarstwa. Chciałby podzielić króliczki na trzy grupy. Króliczek n trafia do grupy $g \in \{0, 1, 2\}$, jeśli jego początkowy stan mieści się w przedziale $\text{stan}[n] \in [g, g + 1)$.

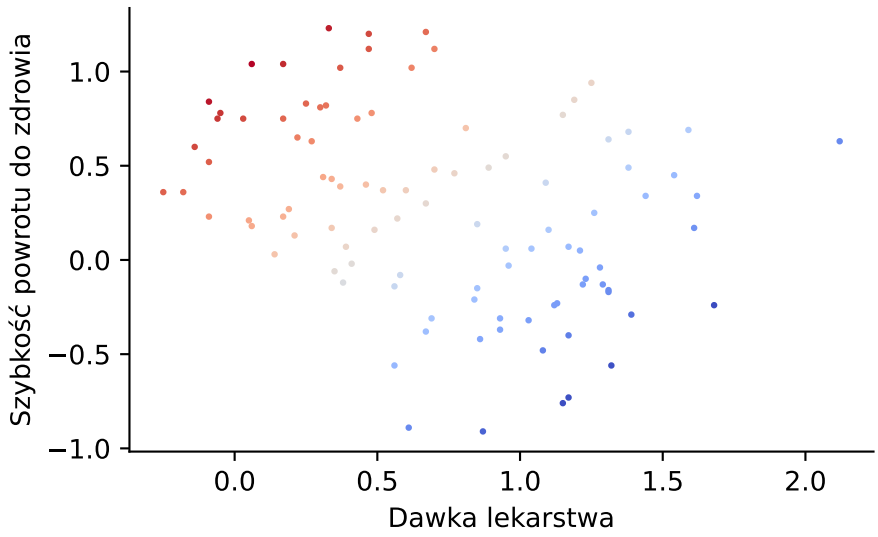
Część (e): Korzystając z pętli `for` policz i wypisz korelację między dawką lekarstwa a szybkością zdrowienia w każdej z trzech grup. [1 punkt]

```
# Zadanie: to miejsce na Twój kod!
```

Część (f): Na poniższym wykresie narysuj prostą najlepszego dopasowania dla każdej grupy. Zinterpretuj otrzymany wynik. [2 punkty]:

```
fig, ax = plt.subplots(figsize=(5, 3), dpi=120)

ax.scatter(lek, zdrowienie, s=2, c=stan, cmap='coolwarm')
ax.set_xlabel("Dawka lekarstwa")
ax.set_ylabel("Szybkość powrotu do zdrowia")
ax.spines[["top", "right"]].set_visible(False)
```

```
for g in range(3):
  ...  # Zadanie: to miejsce na Twój kod :)
```



Część (g): Powyższa analiza podzieliła stan zdrowia na trzy grupy w dość arbitralny sposób. Czy można przenanalizować dane w taki sposób, by uniknąć takiego podziału? [1 punkt]

[*Miejsce na Twoją odpowiedź.*]

Powodzenia!