dr Adam Ćmiel, A3-A4 p.311a, tel. (617) 31-72,

cmiel@agh.edu.pl

http://home.agh.edu.pl/~cmiel

https://cri.agh.edu.pl/oprogramowanie - jak zainstalować program Statistica

Literatura

Koronacki J., Mielniczuk J., Statystyka dla kierunków technicznych i przyrodniczych, WNT 2001.
Klonecki W., Statystyka dla inżynierów, PWN 1991.
Gajek L., Wnioskowanie statystyczne,WNT 1998.
Mendenhall W., Beaver R.J., Beaver B.M., Introduction to Probability and Statistics, Duxbury Press 2005

Statystyka jest nauką zajmującą się najogólniej mówiąc zbieraniem danych i wydobywaniem informacji zawartej w tych danych. Statystykę można podzielić na dwie części:

- statystykę opisową,
- statystykę matematyczną.

Statystyka opisowa zajmuje się gromadzeniem danych oraz informacji dotyczących sposobu ich uzyskania (np. informacji dotyczących sposobu przeprowadzenia eksperymentu laboratoryjnego) oraz ich wstępną obróbką, przez którą rozumiemy **sortowanie danych**, ich **prezentację graficzną** a także wyznaczenie pewnych charakterystyk liczbowych. **Statystyka opisowa nie używa aparatu probabilistycznego**. Jej celem jest sformułowanie pewnych hipotez badawczych mających intuicyjne uzasadnienie w zgromadzonych danych. W raz z powszechną dostępnością komputerów i wzrostem ich mocy obliczeniowej metody statystyki opisowej wzbogaciły się o nowe techniki zwane **eksploracyjną analizą danych**. Połączenie eksploracyjnej analizy danych z techniką systemów uczących i sztucznej inteligencji zaowocowały powstaniem nowej dziedziny: **inteligentnej analizy danych (data mining)**.

Statystyka matematyczna jest częścią probabilistyki powstałą na gruncie rachunku prawdopodobieństwa. Istotna jest jednak różnica między zadaniami rachunku prawdopodobieństwa a zadaniami statystyki matematycznej. Sens tej różnicy ilustruje następujący przykład. Rzucamy 10 razy monetą. Zadanie z rachunku prawdopodobieństwa polega np. na obliczeniu prawdopodobieństwa uzyskania czterech orłów jeżeli wiadomo, że prawdopodobieństwo uzyskania orła w jednym rzucie jest równe 0,5. Statystykę matematyczną interesuje zagadnienie w pewnym sensie odwrotne. Jakie jest prawdopodobieństwo uzyskania orła w jednym rzucie, jeżeli w dziesięciu rzutach uzyskano np. cztery orły? **Zadaniem statystyki matematycznej jest więc określenie nieznanych prawdopodobieństw w przyjętym modelu doświadczenia**.

Cztery typy skal pomiarowych

W naukach empirycznych podstawowym sposobem zdobywania informacji jest **pomiar**, który rozumiemy jako operację przypisywania rzeczywistości eksperymentalnej liczb. Choć pomiar jest równie stary jak nauka, to jego logicznymi podstawami zajęto się dopiero w początku XX wieku, kiedy to **Hölder** (1901) podał aksjomatyzację pomiaru masy. Pewne cechy jak **długość** czy **masa** mierzymy metodami , które rozwinęły się w ciągu stuleci i wydają się obecnie całkiem naturalne. Przy mierzeniu takich cech jak **użyteczność** czy **inteligencja** wykorzystuje się metody, które mogą się wydawać dość arbitralne i mniej naturalne. Kiedy rozważymy problem pomiaru **zdolności twórczych** czy **poczucia szczęścia** pojawia się oczywiste pytanie:

Czy te wielkości można mierzyć?

Odpowiedzi na tak stawiane pytania może udzielić **aksjomatyczna teoria pomiaru**, która zajmuje się m.in. **uzasadnianiem logicznym różnych procedur pomiarowych oraz badaniem sensu wyników uzyskiwanych w wyniku zastosowania tych procedur**.

Oczywiste wydaje się jednak to, że istnieją różne rodzaje pomiarów różniące się między sobą strukturą, "stopniem dokładności" czy "ilością uzyskiwanej informacji".

Wyróżnia się 4 **podstawowe skale pomiarowe**

• nominalną	skale jakościowe
• porządkową	
• interwałową	skale ilościowe
• ilorazową	

Z pomiarem na **skali nominalnej** mamy do czynienia wtedy, gdy każdy obiekt z rozważanego zbioru może być jednoznacznie sklasyfikowany ze względu na pewną cechę. Zbiór obiektów jest wiec rozbity na sumę pewnej ilości (zwykle skończonej) rozłącznych podzbiorów- **klas**. Formalnie rozbicie zbioru obiektów na klasy jest równoważne zadaniu w zbiorze obiektów pewnej **relacji równoważnościowej**, która dzieli zbiór obiektów na klasy równoważności. Klasom równoważności przypisujemy pewne liczby, które traktujemy jako **etykiety**. Dowolny obiekt trafia do jednej z klas równoważności, której odpowiada liczbowa etykieta. Obiektowi w ten sposób przypisujemy liczbę. Oczywiście informacja o obiekcie nie zmieni się jeżeli klasom przypiszemy inne etykiety. Pomiar jest tu wyznaczony z dokładnością do bijektywnego przekształcenia $f : R \xleftrightarrow{1:1} R$. Liczb **nie można tu porównywać**, ani wykonywać na nich żadnych **operacji algebraicznych** (w szczególności uśredniać). Każda rozsądna procedura analizy takich danych powinna być niezmiennicza ze względu na bijekcje $f : R \xleftrightarrow{1:1} R$ zbioru liczb rzeczywistych na siebie.

Przykłady zmiennych nominalnych: **płeć, wyznanie , grupa krwi**.

Z pomiarem na skali **porządkowej** mamy do czynienia wtedy, gdy **w rozważanym zbiorze ilorazowym (czyli zbiorze klas równoważności obiektów) zadana jest pewna relacja liniowego porządku** (zwrotna, antysymetryczna, przechodnia i spójna). Rozważmy dowolną rzeczywistą i

monotoniczną funkcję określoną na zbiorze klas równoważności. Każdy obiekt należy do jednej z klas równoważności, której monotoniczna funkcja przypisuje liczbę. Pomiar polega tu więc na przypisywaniu obiektom liczb z zachowaniem porządku i jest wyznaczony z dokładnością do **rosnących bijekcji** $f : R \xleftrightarrow{1:1} R$. Każda **procedura** analizy danych porządkowych powinna być **niezmiennicza względem rosnących bijekcji**. W statystyce są to np. procedury rangowe.
Przykłady zmiennej porządkowej: **poparcie dla partii rządzącej (**zdecydowanie nie, raczej nie, obojętny, raczej tak, zdecydowanie tak)

Z pomiarem na **skali interwałowej** mamy do czynienia wtedy, gdy w rozważanym zbiorze obiektów jest określona relacja liniowego porządku oraz w iloczynie kartezjańskim zbioru obiektów przez siebie określona jest druga relacja liniowego porządku porządkująca "różnice pomiędzy parami obiektów". Jeżeli obiektom przypiszemy liczby **z zachowaniem obu porządków**, to pomiar jest tu wyznaczony z dokładnością do rosnącego przekształcenia afinicznego $x \to ax+b$; $a>0$, czyli z dokładnością do wyboru zera i jednostki. Procedury analizy takich danych powinny być niezmiennicze względem zmiany położenia i skali (scale and location invariant).
Przykład Pomiar temperatury na skali °C lub °F. Nie ma większego sensu mówić, że temperatura obiektu *A* jest 2 razy wyższa od temperatury obiektu *B*. Można jedynie mówić, że temperatura obiektu *A* jest wyższa od temperatury obiektu *B* o pewną liczbę jednostek.

Pomiar na **skali ilorazowej** jest podobny do pomiaru na skali interwałowej przy czym jest on wyznaczony z dokładnością do rosnącego przekształcenia liniowego $x \to ax$; $x>0$, $a>0$ to znaczy z dokładnością do wyboru jednostki. Zero jest tu naturalnie ustalone. Procedura analizy takich danych powinna być niezmiennicza względem zmiany skali (scale invariant)

Przykład. Pomiar temperatury na skali °K.

Dla danego typu danych pomiarowych mamy do dyspozycji standardowe procedury statystyczne. Dla danych pomiarowych dostępnych na skali wyższego typu można stosować procedury analizy danych niższego typu. Oczywiście taka analiza wiąże się z częściową utratą informacji. **Należy wyraźnie podkreślić, że stosowanie metod przeznaczonych dla wyższych skal pomiarowych do danych dostępnych na niższych skalach jest nieuprawnioną manipulacją**.

Wstępna analiza danych jakościowych

Metody liczbowe

- tabele liczności,
- tabele wielodzielcze

	1 Marichuana	2 Poglądy
1	nigdy	postępowe
2	nigdy	postępowe

Metody graficzne:

- wykresy słupkowe i kołowe
- histogramy skategoryzowane,
- interakcje liczności,
- histogramy 3W
- wykresy obrazkowe (koła, gwiazdy, promienie, twarze Chernoffa, wielokąty, profile itd.)

Wstępna analiza danych ilościowych

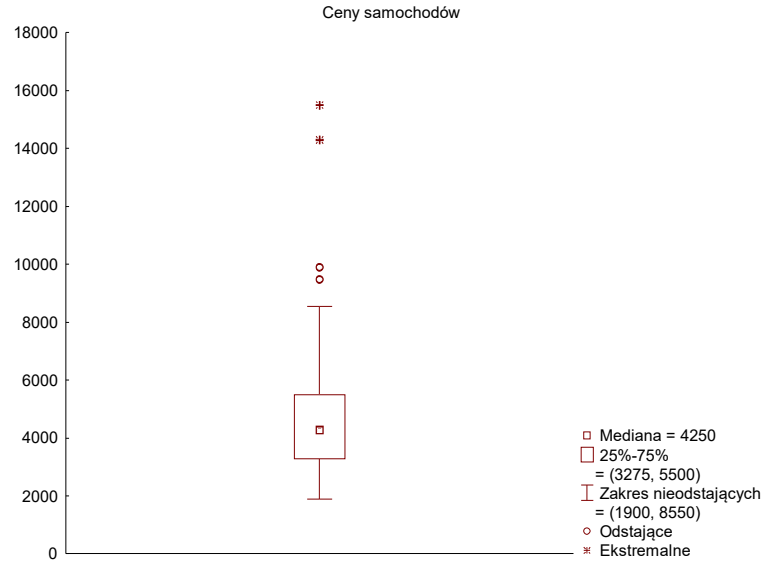
Niech $x_1,...,x_n$ będzie ciągiem danych ilościowych. Oznaczmy przez $x_{(1)} \le ... \le x_{(n)}$ uporządkowane rosnąco dane.

Metody graficzne

- histogramy liczności i częstości - uwagi o wyborze długości przedziału i początku histogramu –jedna z możliwości: długość przedziału $h_0 = 2{,}64 \cdot IQR \cdot n^{-1/3}$ a początek wybieramy tak, aby najmniejsza obserwacja była środkiem pierwszego przedziału (IQR wyjaśnimy poniżej)

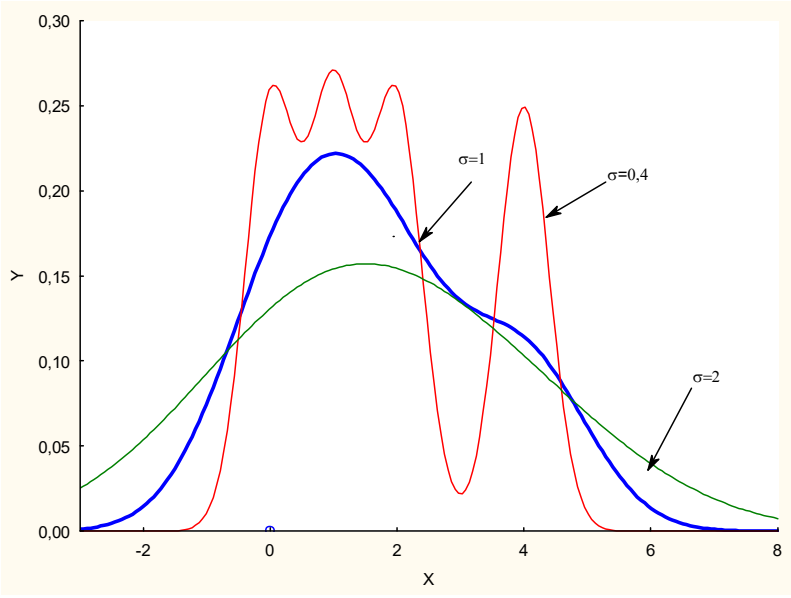
	Czas pracy wierteł
	1 Czas pracy
1	504
2	507
3	522

- **Wykres ramkowy** (skrzynka z wąsami) . Długość wąsa nie powinna przekraczać 1,5×IQR. Jeśli obserwacja jest odległa od brzegu skrzynki więcej niż 1,5×IQR, to jest traktowana jako obserwacja odstająca (outlier). Jeśli obserwacja jest odległa od brzegu skrzynki więcej niż 3×IQR, to jest traktowana jako ekstremalnie odstająca.



- łamane częstości i krzywe estymatora jądrowego

  Przykład. Załóżmy że zaobserwowano 4 wartości zmiennej losowej : 0, 1, 2 i 4. Narysować wykres funkcji będącej średnią arytmetyczną funkcji gęstości rozkładu normalnego z wartościami oczekiwanymi równymi podanym obserwacjom i odchyleniu standardowym 1. Na tym samym rysunku narysować analogiczne wykresy dla parametru σ=0,4 i σ=2.

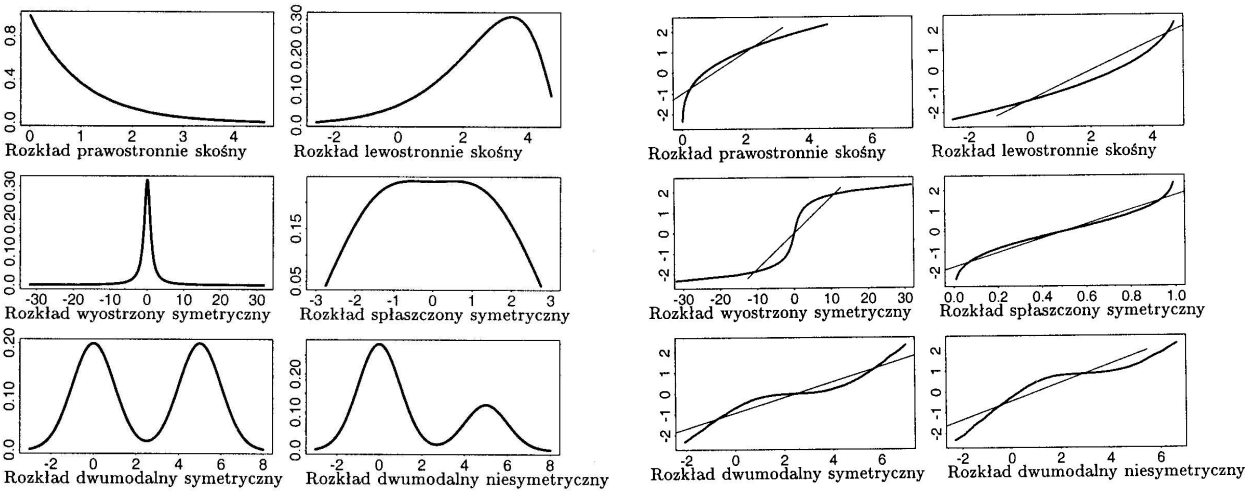


Narysowaliśmy wykres jądrowego estymatora funkcji gęstości dla trzech parametrów wygładzania 0,4 1 i 2.

- wykresy przebiegu dla danych chronologicznych - trendy, okresowość
- wykresy kwantylowe

  $z_{i/n}$ - kwantyl rzędu i/n rozkładu $N(0,1)$ (zwykle kwantyl rzędu $(i-3/8)/(n+1/4)$)

$x_{i/n}$ - kwantyl rzędu *i/n* badanego rozkładu. Jeżeli badany rozkład jest normalny $N(m,\sigma^2)$ to punkty ($x_{i/n}$, $z_{i/n}$) leżą na prostej $z_{i/n} = \dfrac{x_{i/n} - m}{\sigma}$



Wskaźniki sumaryczne

położenia

- wartość średnia w próbie (arytmetyczna, geometryczna, harmoniczna)
- mediana $x_{med} = \begin{cases} x_{((n+1)/2)} & n \ \ nieparzyste \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & n \ parzyste \end{cases}$
- średnia ucinana (trimmed) $\bar{x}_{tk} = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} x_{(i)}$
- średnia winsorowska $\bar{x}_{wk} = \frac{1}{n}\left((k+1)x_{(k+1)} + \sum_{i=k+2}^{n-k-1} x_{(i)} + (k+1)x_{(n-k)}\right)$ -*k* najmniejszych wartości $x_{(1)},...,x_{(k)}$ zastępujemy wartościami $x_{(k+1)}$ natomiast *k* największych wartości $x_{(n-k+1)},...,x_{(n)}$ zastępujemy wartościami $x_{(n-k)}$

rozproszenia (rozrzutu)

- rozstęp próby $R = x_{(n)} - x_{(1)}$
- wariancja w próbie $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ i odchylenie standardowe $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- odchylenie przeciętne $d_1 = \frac{1}{n}\sum_{i=1}^{n}| x_i - \bar{x} |$
- kwantyle (w szczególności kwartyle)
- rozstęp międzykwartylowy $IQR = Q_3 - Q_1$

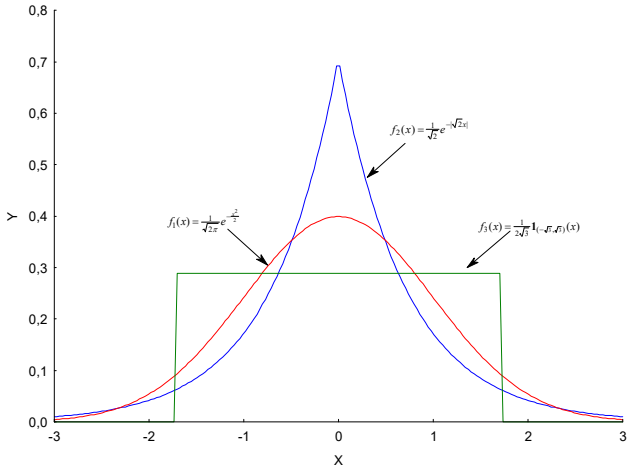
kształtu

- Skośność $= \dfrac{n\sum_{i=1}^{n}(x_i-\bar{x})^3}{(n-1)(n-2)s^3}$ (ocena $E\left(\dfrac{X-\mu}{\sigma}\right)^3$)
- Kurtoza $= \dfrac{n(n+1)\sum_{i=1}^{n}(x_i-\bar{x})^4}{(n-1)(n-2)(n-3)s^4} - 3\dfrac{(n-1)^2}{(n-2)(n-3)}$ (ocena $E\left(\dfrac{X-\mu}{\sigma}\right)^4 - 3$)

Rozważmy trzy symetryczne standaryzowane (o zerowej wartości oczekiwanej i jednostkowej wariancji) rozkłady: normalny, Laplace'a (dwustronny wykładniczy) i jednostajny o funkcjach gęstości odpowiednio:

$f_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$; $f_2(x) = \frac{1}{\sqrt{2}} e^{-|\sqrt{2}x|}$; $f_3(x) = \frac{1}{2\sqrt{3}} \mathbf{1}_{(-\sqrt{3},\sqrt{3})}(x)$

Kurtozy dla tych rozkładów są odpowiednio równe : 0; 3; $-\frac{6}{5}$.



Przestrzeń statystyczna jako model eksperymentu

Z formalnego punktu widzenia eksperyment statystyczny jest opisywany za pomocą przestrzeni statystycznej $(\mathcal{X}, \mathcal{B}, \mathcal{P}=\{P_\theta : \theta\in\Theta\})$, gdzie

- $\mathcal{X}$ jest tzw. przestrzenią prób, czyli zbiorem możliwych wyników eksperymentu,
- $\mathcal{B}$ jest σ-ciałem podzbiorów przestrzeni prób $\mathcal{X}$
- $\mathcal{P}=\{P_\theta : \theta\in\Theta\}$, jest rodziną rozkładów na σ-ciele $\mathcal{B}$ parametryzowaną parametrem θ ze zbioru parametrów Θ.

Uwaga. Powyższy zapis nie ogranicza rodziny rozkładów, gdyż każda rodzina rozkładów może być sparametryzowana w trywialny sposób - parametrem może być sam rozkład (lub jego dystrybuanta). Jeżeli zbiór parametrów Θ jest podzbiorem skończenie wymiarowej przestrzeni euklidesowej R^m, to rodzinę $\mathcal{P}=\{P_\theta : \theta\in\Theta\}$ nazywamy **rodziną parametryczną** a rozważane problemy statystyczne

dotyczące tej rodziny nazywać będziemy **problemami parametrycznymi** np. estymacja parametryczna, testowanie hipotez parametrycznych. Jeżeli Θ nie jest podzbiorem żadnej skończenie wymiarowej przestrzeni euklidesowej, to rodzinę $\mathcal{P}=\{P_\theta : \theta\in\Theta\}$ nazywamy **rodziną nieparametryczną** a rozważane problemy statystyczne dotyczące tej rodziny nazywać będziemy **problemami nieparametrycznymi**.

Przykłady przestrzeni statystycznych

Niech $(X_1,...,X_n)$ będzie ciągiem niezależnych obserwacji zmiennej losowej X o rozkładzie P (czyli $X_1,...,X_n$ **iid** - independent identically distributed). Fakt ten wypowiadamy w statystyce matematycznej następująco: $(X_1,...,X_n)$ **jest próbą prostą z populacji o rozkładzie** ***P***. Wobec tego

$$(x_1,...,x_n) = (X_1(\omega),...,X_n(\omega)) \text{ dla pewnego } \omega$$

i $(x_1,...,x_n)$ traktujemy jako zaobserwowaną wartość (realizację) próby prostej $(X_1,...,X_n)$.

Przykład 1.Niech $\boldsymbol{X}=(X_1,...,X_n)$ będzie próbą prostą z populacji o rozkładzie $N(m,1)$. Przestrzenią statystyczną jest wówczas

$$(R^n,\mathcal{B}(R^n),\{f(x_1,...,x_n)=\frac{1}{(2\pi)^{n/2}}e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-m)^2}, m\in R\}),$$

która jest przestrzenią produktową i jest oznaczana również przez

$$(R, \mathcal{B}(R),\{f(x)=\frac{1}{(2\pi)^{1/2}}e^{-\frac{1}{2}(x-m)^2}, m\in R\})^n .$$

Jednoparametrowa rodzina rozkładów jest wyznaczona w tym przypadku przez rodzinę funkcji gęstości (względem miary Lebesgue'a). Przestrzenie produktowe otrzymujemy wówczas gdy mamy obserwacje typu **iid**. Zbiór parametrów Θ jest w tym przypadku równy R.

Przykład 2. Jeżeli w przykładzie 1 zastąpimy rozkład $N(m,1)$ rozkładem $N(m,\sigma^2)$, to otrzymamy następującą produktową przestrzeń statystyczną

$$(R,\mathcal{B}(R),\{f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}, (m,\sigma)\in R\times R_+\})^n,$$

Parametrem θ jest para (m,σ) a zbiorem parametrów Θ jest w tym przypadku $R\times R_+$.
Oczywiście w obu powyższych przykładach rodziny rozkładów są rodzinami parametrycznymi.

Przykład 3. Jeżeli w przykładzie 1 zastąpimy rozkład $N(m,1)$ rozkładem P_F o dystrybuancie F ze zbioru $\mathcal{F}$ wszystkich dystrybuant na prostej, to otrzymamy przestrzeń statystyczną
$(R, \mathcal{B}(R),\{P_F,F\in\mathcal{F}\})^n$. W tym przypadku $\theta=F$ a $\Theta=\mathcal{F}$. Zbiór $\mathcal{F}$ nie jest podzbiorem żadnej skończenie wymiarowej przestrzeni euklidesowej. Mamy więc do czynienia z nieparametryczną rodziną rozkładów.

Przykład 4. Wykonujemy ciąg niezależnych doświadczeń, z których każde kończy się sukcesem z nieznanym prawdopodobieństwem θ lub porażką z prawdopodobieństwem 1-θ. Doświadczenia te

wykonujemy dopóty, dopóki nie uzyskamy m sukcesów. Sformułować model statystyczny tego eksperymentu.

Aby eksperyment zakończył się w k-tej próbie $k \geq m$ musimy uzyskać w próbie k sukces, a w poprzednich k-1 próbach dokładnie m-1 sukcesów. Oznaczając jak zwykle 1 - sukces, 0 - porażka przestrzeń prób można zapisać następująco:

$$\mathcal{X}=\{ (x_1,...,x_k): k\geq m,\ x_i=1 \text{ lub } x_i=0,\ x_k=1,\ \sum_{i=1}^{k} x_i = m \}.$$

Ponieważ z prawdopodobieństwem 1 taki eksperyment zakończy się w skończonej liczbie prób pomijamy nieskończone ciągi zawierające mniej niż m jedynek. Jest ich przeliczalna ilość. Przestrzeń prób składa się wiec z ciągów skończonych o długości co najmniej m, kończących się jedynką i zawierających dokładnie m jedynek.

Rodzina $\mathcal{B}$ jest w tym przypadku rodziną wszystkich podzbiorów zbioru $\mathcal{X}$, którą oznaczać będziemy $2^{\mathcal{X}}$. Rodzina rozkładów jest następująca:

$$\mathcal{P}=\{p_\theta(x_1,...,x_k)=\prod_{i=1}^{k} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^m(1-\theta)^{k-m}\ \theta\in(0,1),\ k\geq m \}.$$

Tak skonstruowana przestrzeń **nie jest przestrzenią produktową**, ale rodzina rozkładów jest znów rodziną parametryczną.

Możliwa jest też konstrukcja innej przestrzeni

$$\mathcal{X}_1=\{m,m+1,...\},\ \mathcal{B}_1=2^{\mathcal{X}_1},\ \mathcal{P}_1=\{p_\theta(t)=\binom{t-1}{m-1}\theta^m(1-\theta)^{t-m},\ \theta\in(0,1)\},\ t=m,m+1... .$$

Powstaje pytanie: **Czy przeniesienie rozważań z przestrzeni** $(\mathcal{X},\mathcal{B},\mathcal{P})$ **do znacznie uboższej przestrzeni** $(\mathcal{X}_1,\mathcal{B}_1,\mathcal{P}_1)$ **pociąga za sobą utratę informacji**.

Odpowiedź w tym przypadku jest negatywna. Jej uzasadnienie prowadzi do pojęcia **statystyk dostatecznych**.

Zadania

1. Wykonujemy n doświadczeń losowych z których każde kończy się sukcesem z prawdopodobieństwem θ. Wiadomo, że $\theta\in[\theta_1,\theta_2]$, gdzie $\theta_1,\theta_2\in[0,1]$ są ustalone. Sformułować model statystyczny tego eksperymentu.
2. Pewne urządzenie techniczne pracuje dopóty, dopóki nie uszkodzi się któryś z k elementów typu A lub któryś z l elementów typu B. Czas życia elementów typu A jest zmienną losową o rozkładzie wykładniczym z gęstością $f_\alpha(x)=\alpha^{-1}\exp(-x/\alpha)$, a czas życia elementów typu B jest zmienną losową o rozkładzie wykładniczym z gęstością $f_\beta(x)=\beta^{-1}\exp(-x/\beta)$. Obserwuje się czas życia T całego urządzenia. Sformułować model statystyczny tej obserwacji.
3. Przeprowadza się $n=\sum_{i=1}^{k} n_i$ eksperymentów w taki sposób, że n_i eksperymentów wykonuje się na poziomie x_i, $i=1,...,k$. Prawdopodobieństwo sukcesu w eksperymencie przeprowadzonym na

poziomie x jest równe $p(x)=\frac{1}{1+e^{-(\alpha+\beta x)}}$, $\alpha\in R$, $\beta>0$, gdzie(α,β) jest nieznanym parametrem. Sformułować model statystyczny tego eksperymentu.

4. **Pewna optymalna własność mediany**. Medianą zmiennej losowej o rozkładzie P nazywamy liczbę m_e taką, że $P(X\le m_e)\ge\frac{1}{2}$ i $P(X\ge m_e)\ge\frac{1}{2}$. Niech X będzie całkowalną zmienną losową (tzn. $E|X|<\infty$). Pokazać, ze funkcja $\varphi(c)= E|X-c|$ osiąga najmniejszą wartość, gdy $c= m_e$.

   Wskazówka Rozważyć 2 przypadki 1) $c<m_e$ 2) $c>m_e$ i w każdym z nich pokazać że

   $\varphi(c)- \varphi(m_e)= E|X-c|- E|X-m_e|=\psi(c)$ przy czym $\psi(c)\ge 0$ dla każdego c i $\psi(m_e)=0$.