Część I

Podstawy

Rozdział 1

Próbkowe odpowiedniki wielkości populacyjnych

1.1 Rozkład empiryczny

Statystyka matematyczna opiera się na założeniu, że dane są wynikiem pewnego „doświadczenia losowego". Przypuśćmy, że dane mają postać ciągu liczb $x_1, x_2, \ldots, x_n$. Zakładamy, że mamy do czynienia ze *zmiennymi losowymi* $X_1, X_2, \ldots, X_n$ określonymi na przestrzeni probabilistycznej $(\Omega, \mathcal{F}, \mathbb{P})$ i dane są realizacjami (wartościami) tych zmiennych losowych, czyli $x_1 = X_1(\omega), \ldots, x_n = X_n(\omega)$ dla pewnego $\omega \in \Omega$. *Nie znamy* rozkładu prawdopodobieństwa $\mathbb{P}$ na przestrzeni Ω, który „rządzi" zachowaniem zmiennych losowych i chcemy się dowiedzieć czegoś o tym rozkładzie na podstawie obserwacji $x_1, x_2, \ldots, x_n$. Rozważmy najpierw prostą sytuację, kiedy obserwacje są realizacjami niezależnych zmiennych losowych o jednakowym rozkładzie.

1.1.1 DEFINICJA. ***Próbką*** *z rozkładu prawdopodobieństwa o dystrybuancie F nazywamy ciąg niezależnych zmiennych losowych $X_1, X_2, \ldots, X_n$ o jednakowym rozkładzie, $\mathbb{P}(X_i \leqslant x) = F(x)$ dla $i = 1, 2, \ldots, n$. Będziemy używali oznaczenia*

$$X_1, X_2, \ldots, X_n \sim_{\text{iid}} F.$$

W powyższej definicji dystrybuanta jest tylko pewnym sposobem opisu rozkładu prawdopodobieństwa. Mówiąc na przykład o próbce z rozkładu normalnego, napiszemy $X_1, \ldots X_n \sim_{\text{iid}} \mathrm{N}(\mu, \sigma^2)$. Mówi się także, że $X_1, X_2, \ldots, X_n$ jest próbką z rozkładu fikcyjnej zmiennej losowej $X \sim F$.

Uwaga. W statystycznych badaniach reprezentacyjnych stosuje się różne schematy losowania z populacji skończonej. W Definicji 1.1.1 żądamy niezależności, zatem ta definicja *nie obejmuje* próbki wylosowanej *bez zwracania*.

1.1.2 DEFINICJA. *Niech $X_1, X_1, \ldots, X_n$ będzie próbką z rozkładu o dystrybuancie F. Funkcję*

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leqslant x)$$

nazywamy ***dystrybuantą empiryczną****.*

Gdy chcemy podkreślić, że próbka ma rozmiar n, to piszemy $\hat{F}_n$ zamiast $\hat{F}$. Traktujemy $\hat{F}$ jako „empiryczny odpowiednik" nieznanej dystrybuanty F.

1.1.3 Przykład (Waga noworodków)**.** Powiedzmy, że wylosowano 114 noworodków[1] w celu poznania cech fizycznych dzieci urodzonych w Warszawie w roku 2009. Waga noworodków była taka:

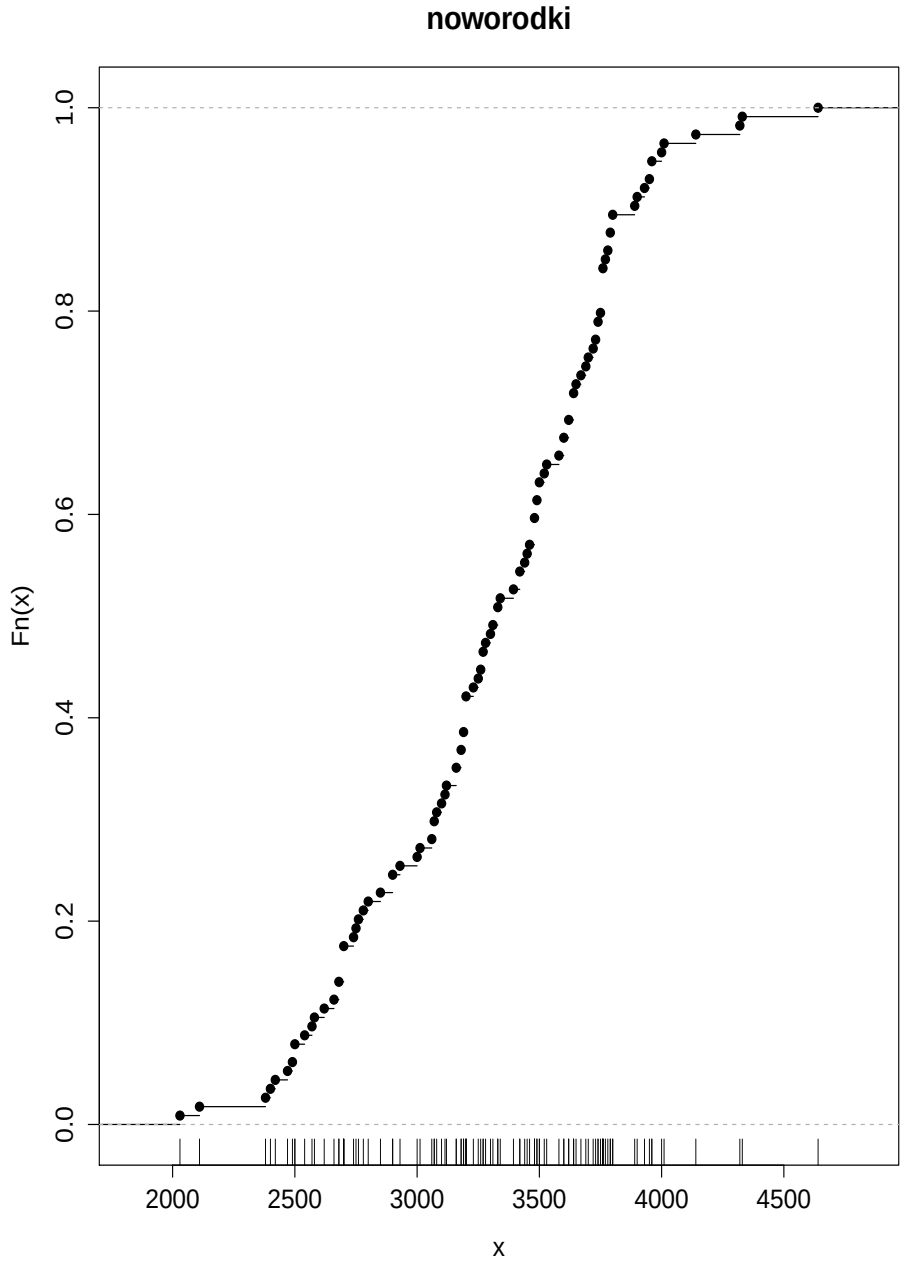
3080	3650	3250	4000	3180	3480	4140	3930	3950	2700
3720	3520	3200	3700	3500	3790	3900	3760	3740	3200
3280	3960	3300	2490	3260	3780	3600	3060	2850	3490
2620	3690	3200	3070	4640	3760	3190	3180	3760	3670
3310	3770	2580	2700	3740	2700	3760	3960	2800	3500
3460	3800	3394	3640	2680	3490	3000	2900	4320	3450
3200	3530	3330	2680	2700	3580	2500	2660	3600	3114
3760	3640	2780	2760	3480	2420	2110	2930	3160	3012
2900	3750	4010	3230	2570	3480	3340	3420	3330	2030
3730	3640	3420	4330	3790	3120	3890	3070	3270	2750
2470	3620	2740	3800	3440	3160	3620	3190	2380	3100
2400	2500	2540	3270						

Dane traktujemy jako próbkę z rozkładu prawdopodobieństwa zmiennej losowej $X =$ „waga noworodka losowo wybranego z populacji". Rysunek 1.1 przedstawia dystrybuantę empiryczną $\hat{F}$ odpowiadającą tej próbce. ◊

Dystrybuanta empiryczna jest funkcją pary argumentów (x, ω), czyli $\hat{F} : \mathbb{R} \times \Omega \to [0, 1]$, ale wygodnie jest pomijać argument ω. *Dla ustalonego* $\omega \in \Omega$ dystrybuanta empiryczna jest funkcją $\mathbb{R} \to [0, 1]$, która argumentowi x przyporządkowuje liczbę $\sum \mathbb{1}(X_i(\omega) \leqslant x)/n$. *Dla ustalonego* $a \in \mathbb{R}$ wartość dystrybuanty empirycznej jest *zmienną losową*, $\hat{F}(a) : \Omega \to [0, 1]$. Ciąg indykatorów odpowiada schematowi Bernoulliego z prawdopodobieństwem sukcesu $F(a)$ i dlatego zmienna losowa $\hat{F}(a)$ ma następujący rozkład prawdopodobieństwa:

$$\mathbb{P}(\hat{F}(a) = k/n) = \binom{n}{k} F(a)^k (1 - F(a))^{n-k} \qquad (k = 0, 1, \ldots, n).$$

[1] W istocie, dane pochodzą z dwóch numerów „Gazety Wyborczej", („Gazeta Stołeczna", 29 sierpnia 2009 i 5 września 2009).



Rysunek 1.1: Dystrybuanta empiryczna wagi noworodków. Dane z Przykładu 1.1.3.

1.1.4 DEFINICJA. *Rozważmy próbkę $X_1, X_2, \ldots, X_n$. Dla każdego $\omega \in \Omega$, niech $X_{1:n}(\omega) \leqslant X_{2:n}(\omega) \leqslant \cdots \leqslant X_{n:n}(\omega)$ będzie ciągiem liczb $X_1(\omega), X_2(\omega), \ldots, X_n(\omega)$ uporządkowanym w kolejności rosnącej. Określone w ten sposób zmienne losowe $X_{1:n}, X_{2:n}, \ldots, X_{n:n}$ nazywamy* ***statystykami pozycyjnymi****.*

W szczególności, $X_{1:n} = \min(X_1, \ldots, X_n)$ i $X_{n:n} = \max(X_1, \ldots, X_n)$; pierwsza i ostatnia statystyka pozycyjna to, odpowiednio, najmniejsza i największa obserwacja w próbce.

Dystrybuanta empiryczna $\hat{F}$ jest funkcją „schodkową": jest stała na każdym z przedziałów pomiędzy statystykami pozycyjnymi $[X_{i:n}, X_{i+1:n}[$. Widać, że

$$\begin{aligned}
&\text{dla } x < X_{1:n} \text{ mamy } \hat{F}(x) = 0;\\
&\text{dla } X_{i:n} \leqslant x < X_{i+1:n} \text{ mamy } \hat{F}(x) = \frac{i}{n};\\
&\text{dla } x \geqslant X_{n:n} \text{ mamy } \hat{F}(x) = 1.
\end{aligned}$$

W punktach $X_{i:n}$ funkcja $\hat{F}$ ma nieciągłości (skacze w górę). Jeśli *teoretyczna* dystrybuanta F jest ciągła, to $\mathbb{P}(X_{1:n} < X_{2:n} < \cdots < X_{n:n}) = 1$, a więc, z prawdopodobieństwem 1, mamy $\hat{F}(X_{i:n}) = i/n$ i każdy skok dystrybuanty empirycznej ma wielkość $1/n$. Jeśli teoretyczna dystrybuanta jest dyskretna, to z niezerowym prawdopodobieństwem niektóre statystyki pozycyjne będą się pokrywać i dystrybuanta empiryczna będzie miała skoki wysokości $2/n$ lub $3/n$ i tak dalej.

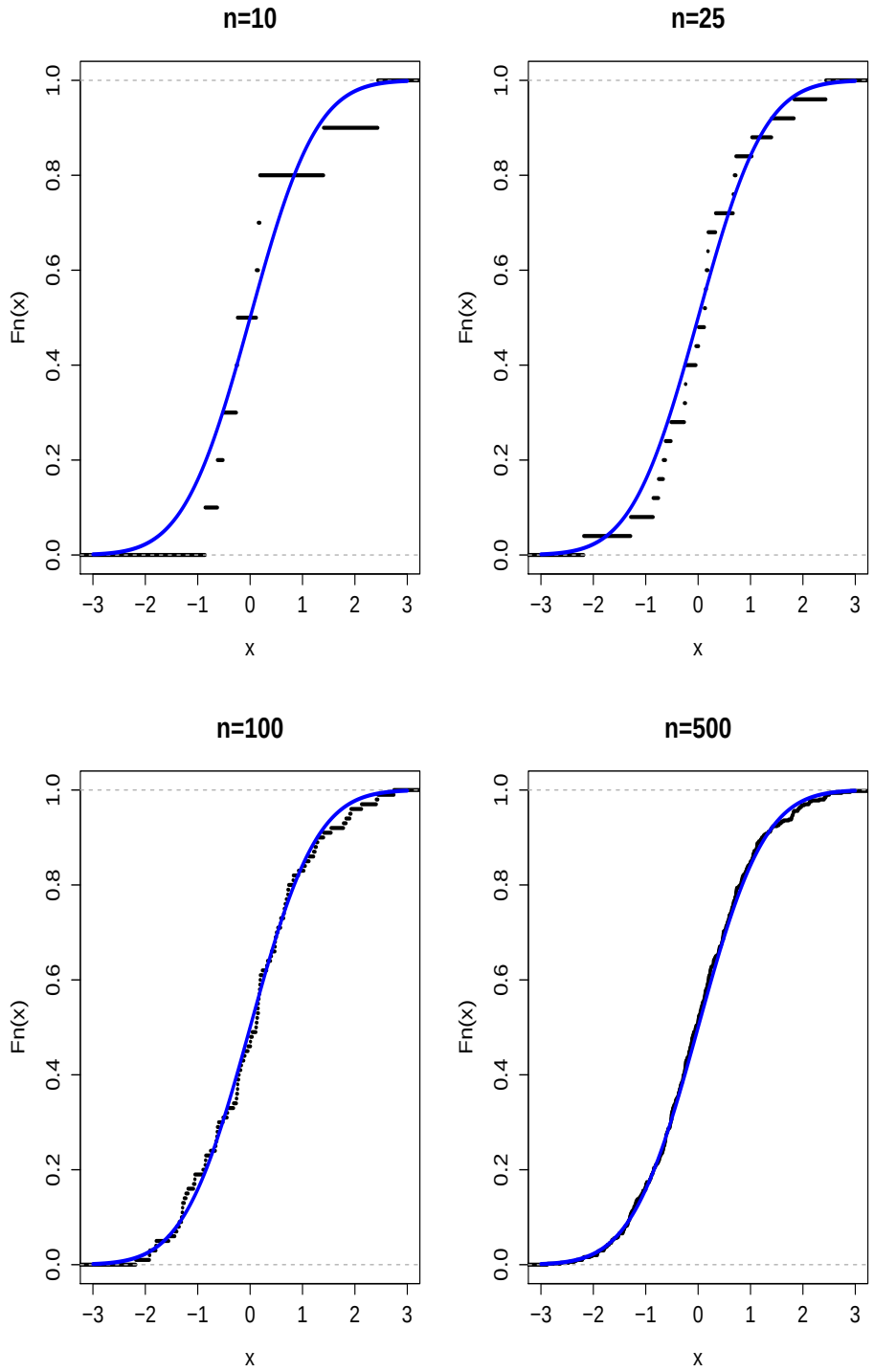
W poniższym stwierdzeniu będziemy mieli do czynienia z *nieskończoną* próbką, czyli z ciągiem zmiennych losowych $X_1, X_2, \ldots, X_n, \ldots$, które są niezależne i mają jednakowy rozkład prawdopodobieństwa. Możemy sobie wyobrazić, że wciąż dodajemy do próbki nowe zmienne losowe. Dystrybuanta empiryczna $\hat{F}_n$ jest określona tak jak w Definicji 1.1.2, to znaczy, zależy od *początkowych* zmiennych $X_1, \ldots, X_n$. Rozpatrujemy teraz *ciąg* dystrybuant empirycznych $\hat{F}_1, \hat{F}_2, \ldots, \hat{F}_n, \ldots$.

1.1.5 Stwierdzenie. *Jeśli $X_1, \ldots, X_n, \ldots$ jest próbką z rozkładu o dystrybuancie F, to dla każdego $x \in \mathbb{R}$,*

$$\hat{F}_n(x) \to_{\text{p.n.}} F(x), \quad (n \to \infty).$$

Dowód. Zmienne losowe $\mathbb{1}(X_1 \leqslant x), \ldots, \mathbb{1}(X_n \leqslant x), \ldots$ są niezależne i mają jednakowy rozkład prawdopodobieństwa: $\mathbb{1}(X_n \leqslant x)$ przyjmuje wartość 1 z prawdopodobieństwem $F(x)$ lub wartość 0 z prawdopodobieństwem $1 - F(x)$. Oczywiście, $\mathbb{E}\mathbb{1}(X_n \leqslant x) = F(x)$. Z Mocnego Prawa Wielkich Liczb (MPWL) dla schematu Bernoulliego wynika, że zdarzenie $\lim_{n\to\infty} \hat{F}_n(x) = F(x)$ zachodzi z prawdopodobieństwem 1. To znaczy, że ciąg zmiennych losowych $\hat{F}_n(x)$ jest zbieżny *prawie na pewno* do liczby $F(x)$. □

Istnieje mocniejsza wersja poprzedniego stwierdzenia, którą przytoczymy bez dowodu. Można pokazać, że zbieżność $\hat{F} \to F$ zachodzi *jednostajnie* z prawdopodobieństwem 1.



Rysunek 1.2: Zbieżność dystrybuant empirycznych do dystrybuanty.

1.1.6 TWIERDZENIE (Gliwienko-Cantelli)**.** *Jeżeli $X_1, \ldots, X_n, \ldots$ jest próbką z rozkładu o dystrybuancie F to*

$$\sup_{-\infty<x<\infty} |\hat{F}_n(x) - F(x)| \to_{\text{p.n.}} 0 \quad (n \to \infty).$$

Jeśli mamy możliwość nieograniczonego powiększania próbki, to możemy poznać rozkład prawdopodobieństwa z dowolną dokładnością.

Zamiast dowodu Twierdzenia Gliwienki-Cantelliego przytoczymy wyniki przykładowych symulacji komputerowych. Na Rysunku 1.2 widać dystrybuanty empiryczne F_{10}, F_{25}, F_{100} i F_{500}, dla próbki z rozkładu normalnego $\mathrm{N}(0,1)$ – na tle teoretycznej dystrybuanty tego rozkładu (ciągła, niebieska krzywa).

Skoncentrowaliśmy uwagę na dystrybuancie empirycznej, ale podobnie można zdefiniować o empiryczny rozkład prawdopodobieństwa. Rozważmy zbiór borelowski $B \subseteq \mathbb{R}$ i próbkę $X_1, X_2, \ldots, X_n$ z rozkładu zmiennej losowej X. Przybliżeniem nieznanej liczby $P(B) = \mathbb{P}(X \in B)$ jest **prawdopodobieństwo empiryczne**

$$\hat{P}(B) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \in B).$$

Określone w ten sposób odwzorowanie $\hat{P} : \mathcal{B} \times \Omega \to \mathbb{R}$, gdzie $\mathcal{B}$ oznacza rodzinę zbiorów borelowskich, nazywane jest **empirycznym rozkładem prawdopodobieństwa**. Dla ustalonego $\omega \in \Omega$ jest to dyskretny rozkład prawdopodobieństwa; jeśli wartości $x_1 = X_1(\omega), \ldots, x_n = X_n(\omega)$ są różnymi liczbami to $\hat{P}(\{x_i\}) = 1/n$ dla $i = 1, 2, \ldots, n$, czyli empiryczny rozkład prawdopodobieństwa jest rozkładem równomiernym na zbiorze $\{x_1, \ldots, x_n\}$. Z drugiej strony $\hat{P}(B)$ jest, dla ustalonego zbioru B, *zmienną losową* (a nie liczbą). Oczywiście, $\hat{P}(]-\infty, x]) = \hat{F}(x)$.

1.1.7 Przykład (Statystyczna kontrola jakości)**.** Producent chce się dowiedzieć, jaki procent wytwarzanych przez niego wyrobów jest wadliwych. Sprawdza dokładnie pewną liczbę sztuk. Powiedzmy, że badaniu poddano 50 sztuk i wyniki są takie (zakodujemy „wyrób prawidłowy" jako liczbę „1" i „wadliwy" jako „0"):

1 0 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1

Potraktujemy ten ciąg jako próbkę z pewnego rozkładu prawdopodobieństwa na zbiorze dwupunktowym $\{\text{prawidłowy}, \text{wadliwy}\} = \{1, 0\}$. Producenta interesuje liczba

$$P(0) = P(\text{wadliwy}) = \%\text{ sztuk wadliwych wśród } \textit{wszystkich} \text{ wyrobów.}$$

Na podstawie próbki możemy obliczyć prawdopodobieństwo *empiryczne*

$$\hat{P}(0) = \hat{P}(\text{wadliwy}) = \%\text{ sztuk wadliwych wśród 50 } \textit{zbadanych} \text{ wyrobów}$$

$$= \frac{5}{50} = 0.10.$$

Przykład jest trywialny. Chodzi tylko o to, żeby podkreślić różnicę między *nieznaną*, interesującą nas liczbą $P(0)$ i *znaną ale losową* wielkością $\hat{P}(0)$. $\Diamond$

1.2 Momenty i kwantyle z próbki.

Określimy teraz *próbkowe odpowiedniki* pewnych wielkości, związanych z rozkładem prawdopodobieństwa. Będziemy postępowali w podobnym duchu jak w definicji dystrybuanty empirycznej. Cały czas $X_1, \ldots, X_n$ jest próbką. **Średnią z próbki** nazywamy zmienną losową

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Widać, że $\bar{X}$ jest wartością oczekiwaną rozkładu empirycznego. Podobnie, **wariancja z próbki**

$$\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

jest niczym innym, jak wariancją rozkładu empirycznego. Wyższego rzędu **momenty z próbki** (zwykłe i centralne) oznaczymy przez $\hat{a}_k$ i $\hat{m}_k$:

$$\hat{a}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \qquad \hat{m}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k.$$

Są to odpowiedniki momentów, czyli

$$a_k = \mathbb{E}X^k, \qquad m_k = \mathbb{E}(X_i - \mathbb{E}X)^k.$$

Wielkości a_k i m_k zależą od „prawdziwego", teoretycznego rozkładu zmiennej losowej X, podczas gdy $\hat{a}_k$ i $\hat{m}_k$ są obliczone dla rozkładu empirycznego. Oczywiście, $\hat{a}_1 = \bar{X}$ i $\hat{m}_2 = \tilde{S}^2$, ale te dwa momenty spotykać będziemy tak często, że zasługują na specjalne oznaczenie. Zauważmy jeszcze oczywisty związek $\hat{m}_2 = \hat{a}_2 - \hat{a}_1^2$ (Zadanie 1.4).

Kwantyle próbkowe określamy zgodnie z tym samym schematem. Po prostu zastępujemy rozkład prawdopodobieństwa rozkładem *empirycznym* i obliczamy kwantyle. Przypomnijmy najpierw definicję kwantyla. Niech $0 < q < 1$. Jeśli $\mathbb{P}(X < \xi_q) = F(\xi_q-) \leqslant q \leqslant F(\xi_q) = \mathbb{P}(X \leqslant \xi_q)$, to liczbę ξ_q nazywamy **kwantylem** rzędu q zmiennej losowej X. Taka liczba zawsze istnieje, ale nie musi być wyznaczona jednoznacznie. Jeśli istnieje dokładnie jedna liczba ξ_q taka, że $\mathbb{P}(X \leqslant \xi_q) = F(\xi_q) = q$ to oczywiście ξ_q jest q-tym kwantylem. Podobnie jest w przypadku gdy $F(\xi_q-) < q < F(\xi_q)$. Jeśli jednak $F(a) = F(b) = q$, to każda z liczb z przedziału $[a, b]$ jest kwantylem.

Liczbę $\hat{\xi}_q$ nazywamy **kwantylem empirycznym** rzędu q, jeśli

$$\hat{F}(\hat{\xi}_q-) \leqslant q \leqslant \hat{F}(\hat{\xi}_q).$$

Statystyka pozycyjna $X_{\lceil np\rceil:n}$ jest kwantylem empirycznym rzędu p ale niekoniecznie jedynym. Najlepiej widać to na przykładzie mediany (kwantyla rzędu $q = 1/2$). Jeśli rozmiar próbki n jest liczbą nieparzystą, to statystyka pozycyjna o numerze $(n + 1)/2$ jest *medianą z próbki*. Jeśli rozmiar próbki jest liczbą parzystą, to każda z liczb z przedziału $[X_{n/2:n}, X_{n/2+1,n}]$ jest medianą rozkładu empirycznego. W R i innych pakietach statystycznych, dla uniknięcia niejednoznaczności, zwykle podaje się środek przedziału median: $(X_{n/2:n} + X_{n/2+1:n})/2$. Przyjmiemy następujące oznaczenia na medianę i medianę z próbki:

$$\mathrm{med}(X) = \xi_{1/2}, \qquad \hat{\mathrm{med}} = \hat{\mathrm{med}}(X_1, \ldots, X_n) = \hat{\xi}_{1/2}.$$

Kwantyle rzędu 1/4 i 3/4 noszą nazwę kwartyli i bywają oznaczane Q_1 i Q_3

1.2.1 Przykład (Waga noworodków, kontynuacja)**.** Dla naszej „niemowlęcej" próbki z Przykładu 1.1.3 mamy

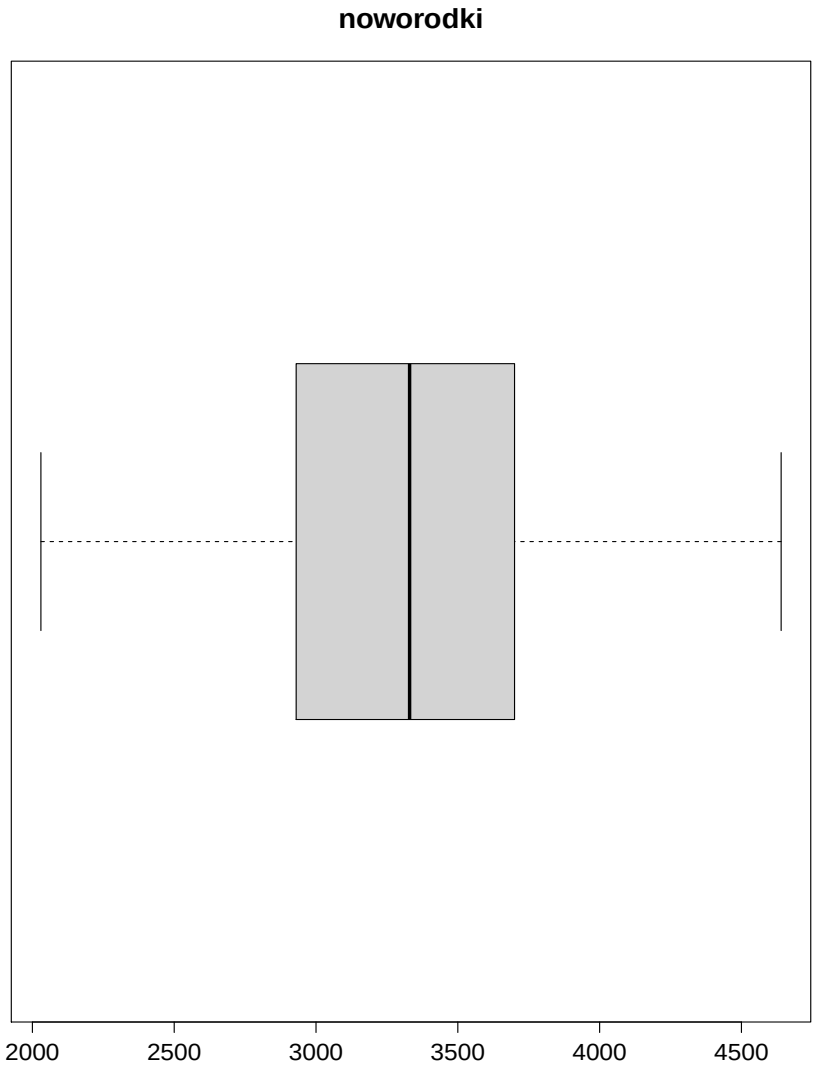
$$\bar{X} = 3302.105, \qquad \tilde{S} = 502.5677$$

Jak już zauważyliśmy poprzednio, $\hat{\mathrm{med}} = 3330$. Kwartyle próbkowe, zgodnie z naszą definicją, są równe $Q_1 = \hat{\xi}_{1/4} = X_{29:114} = 2930$ i $Q_3 = \hat{\xi}_{3/4} = X_{86:114} = 3700$ [2]. ♢

Medianę, kwartyle, minimum i maksimum próbki przedstawia tak zwany „wykres pudełkowy" (ang. *Box and Whiskers Plot*, Rysunek 1.3). Boki prostokąta (na tym rysunku boki pionowe) odpowiadają kwartylom. Kreska wewnątrz prostokąta pokazuje medianę. „Wąsy" umieszcza się (w zasadzie) w miejscu minimum i maksimum z próbki. Wykres pudełkowy pozwala na graficzne porównanie kilku próbek. W tym przypadku na jednym obrazku widnieje kilka pudełek, a ich „grubość" może być związana z licznościami poszczególnych próbek.

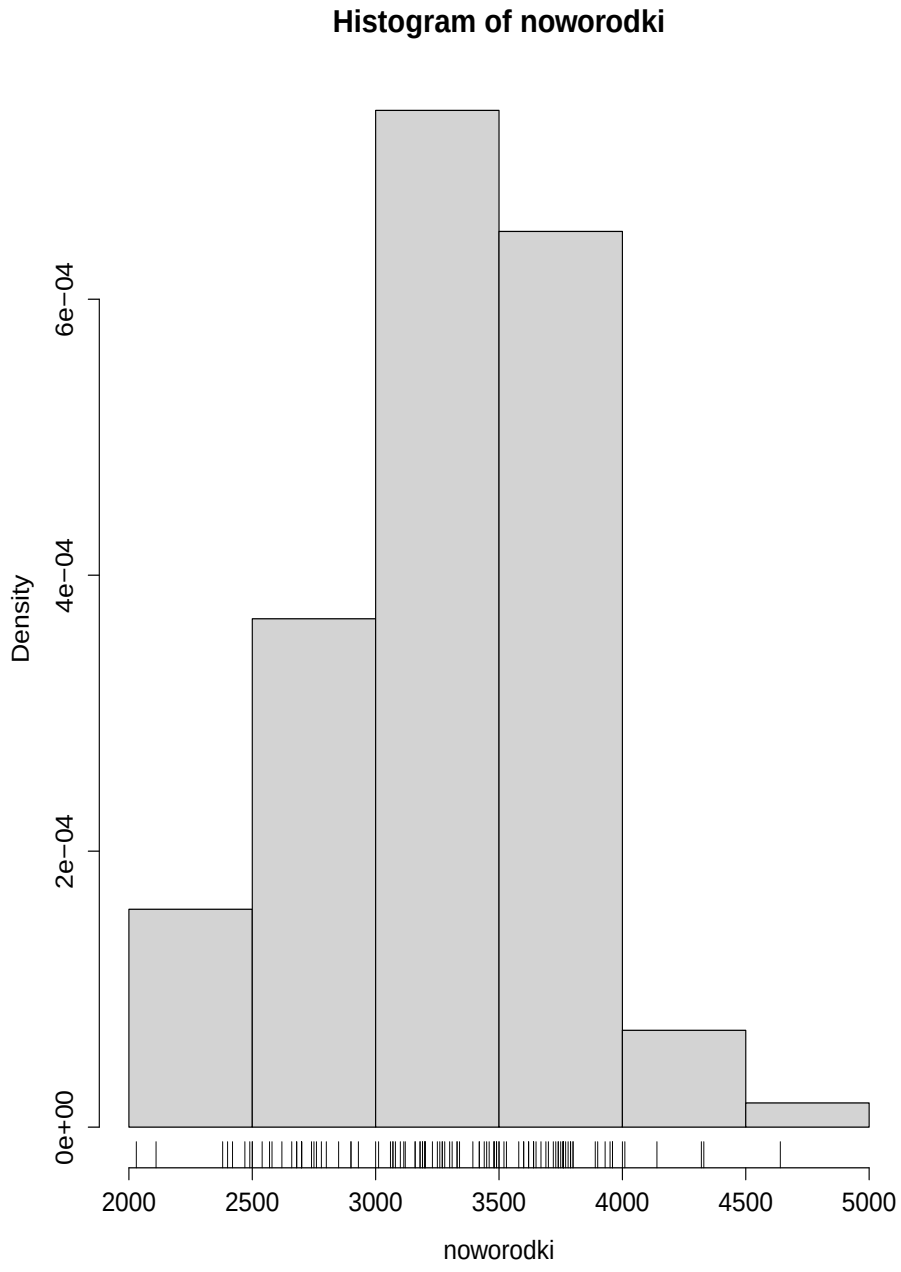
Na zakończenie naszych wstępnych rozważań wspomnimy o jeszcze jednym graficznym sposobie podsumowania danych. Na Rysunku 1.4 przedstawiony jest histogram danych z Przykładu 1.1.3. Wydaje się, że szczegółowe objaśnienia są zbędne, bo budowa histogramu jest

[2] Określenie kwantyla próbkowego w pakietach statystycznych nieco różni się od naszego, ale nie ma to zasadniczego znaczenia, szczególnie jeśli próbka jest duża. W naszym przykładzie R podaje następujące wartości kwartyli: $Q_1 = \hat{\xi}_{1/4} = 2947.5$ i $Q_3 = \hat{\xi}_{3/4} = 3697.5$.



Rysunek 1.3: Wykres pudełkowy. Dane z Przykładu 1.1.3.

dość oczywista i dobrze znana czytelnikom prasy i telewidzom. Zwróćmy tylko uwagę na to, że skala osi pionowej zastała tak dobrana, aby pole pod histogramem było równe 1, podobnie jak pole pod wykresem gęstości prawdopodobieństwa. W istocie, histogram jest w pewnym sensie empirycznym odpowiednikiem gęstości.



Rysunek 1.4: Histogram danych z Przykładu 1.1.3.

1.3 Zadania

1.1. Obliczyć $\mathbb{E}\hat{F}(x)$, $\mathrm{Var}\hat{F}(x)$.

1.2. Pokazać, że ciąg zmiennych losowych $\sqrt{n}(\hat{F}_n(x) - F(x))$ jest zbieżny do rozkładu normalnego. Zidentyfikować parametry tego rozkładu.

1.3. Podać granicę $\lim_{n\to\infty}\mathbb{P}(\hat{F}_n(x) \leqslant F(x))$ przy założeniu, że $0 < F(x) < 1$. Dokładnie uzasadnić odpowiedź.

1.4. Wyprowadzić alternatywny wzór na wariancję próbkową:

$$\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2.$$

1.5. Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu normalnego $\mathrm{N}(\mu, \sigma^2)$. Podać rozkład średniej próbkowej $\bar{X} = \sum X_i/n$.

1.6. Obliczyć dystrybuantę i gęstość rozkładu zmiennej losowej $U_{n:n} = \max(U_1, \ldots, U_n)$, gdzie $U_1, \ldots, U_n$ jest próbką z rozkładu jednostajnego $\mathrm{U}(0, 1)$.

1.7. (Ciąg dalszy). Obliczyć $\mathbb{E}U_{n:n}$, gdzie $U_{n:n}$ oznacza ostatnią statystykę pozycyjną (maksimum z próbki) z rozkładu jednostajnego $\mathrm{U}(0, 1)$.

1.8. (Ciąg dalszy). Obliczyć $\mathrm{Var}U_{n:n}$, gdzie $U_{n:n}$ oznacza maksimum z próbki z rozkładu jednostajnego $\mathrm{U}(0, 1)$.

1.9. (Ciąg dalszy). Zbadać zbieżność według rozkładu ciągu zmiennych losowych $n(1-U_{n:n})$, gdzie $U_{n:n}$ oznacza ostatnią statystykę pozycyjną (maksimum z próbki) z rozkładu jednostajnego $\mathrm{U}(0, 1)$.

1.10. Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu wykładniczego $\mathrm{Ex}(\lambda)$. Obliczyć rozkład prawdopodobieństwa $X_{1:n}$, pierwszej statystyki pozycyjnej (minimum z próbki). Podać dystrybuantę, gęstość, nazwę tego rozkładu.

1.11. Rozważmy próbkę $X_1, \ldots, X_n$ z rozkładu o dystrybuancie F. Pokazać, że zmienna losowa $X_{k:n}$ (k-ta statystyka pozycyjna) ma dystrybuantę

$$\mathbb{P}(X_{k:n} \leqslant x) = \sum_{i=k}^{n} \binom{n}{i} F(x)^i (1 - F(x))^{n-i}.$$

1.12. Załóżmy, że dystrybuanta F jest funkcją ciągłą i ściśle rosnącą, a zatem istnieje funkcja odwrotna $F^{-1} :]0, 1[\to \mathbb{R}$. Pokazać, że jeśli $U \sim \mathrm{U}(0, 1)$ to zmienna losowa $X = F^{-1}(U)$ ma dystrybuantę F.

1.13. (Ciąg dalszy). Niech $U_{k:n}$ oznacza statystykę pozycyjną z rozkładu $\mathrm{U}(0, 1)$. Pokazać, że $X_{k:n} = F^{-1}(U_{k:n})$ ma rozkład taki jak statystyka pozycyjna z rozkładu o dystrybuancie F.

Rozdział 2

Modele statystyczne

2.1 Przestrzenie statystyczne

Zaczniemy od formalnej definicji, której sens postaramy się w dalszym ciągu wyjaśnić i zilustrować przykładami.

2.1.1 DEFINICJA. ***Przestrzeń statystyczna*** *jest to trójka* $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta; \theta \in \Theta\})$*, gdzie* $\mathcal{X}$ *jest zbiorem, wyposażonym w* σ*-ciało* $\mathcal{F}$ *podzbiorów, zaś* $\{\mathbb{P}_\theta; \theta \in \Theta\}$ *jest rodziną rozkładów prawdopodobieństwa na przestrzeni* $(\mathcal{X}, \mathcal{F})$*. Zbiór* $\mathcal{X}$ *nazywamy przestrzenią obserwacji zaś* Θ *nazywamy przestrzenią parametrów.*

Widoczny jest związek z definicją znaną z rachunku prawdopodobieństwa. Dla każdego ustalonego $\theta \in \Theta$, trójka $(\mathcal{X}, \mathcal{F}, \mathbb{P}_\theta)$ jest przestrzenią *probabilistyczną*. Najważniejszą nowością w Definicji 2.1.1 jest to, że rozważamy rodzinę rozkładów prawdopodobieństwa, $\{\mathbb{P}_\theta; \theta \in \Theta\}$. Jak już powiedzieliśmy w poprzednim rozdziale, w statystyce matematycznej traktujemy dane jako wynik doświadczenia losowego, ale nie wiemy, jaki rozkład „rządzi" badanym zjawiskiem. Wobec tego rozpatrujemy rodzinę wszystkich branych pod uwagę rozkładów prawdopodobieństwa. Zakładamy, że „prawdziwy" rozkład należy do tej rodziny, czyli jest to rozkład $\mathbb{P}_{\theta_0}$ dla pewnego $\theta_0 \in \Theta$, tylko nie umiemy wskazać θ_0.

2.1.2 Uwaga (Kanoniczna przestrzeń próbkowa). Powiedzmy, że wynikiem obserwacji są zmienne losowe $X_1, \ldots, X_n$. Niech Ω będzie zbiorem wszystkich możliwych wyników doświadczenia losowego, a więc w naszym przypadku zbiorem ciągów $\omega = (x_1, \ldots, x_n)$. Możemy przyjąć, że zmienne losowe X_i są funkcjami określonymi na przestrzeni próbkowej Ω wzorem $X_i(\omega) = x_i$. Wektor $X = (X_1, \ldots, X_n)$ możemy traktować jako pojedynczą, wielowymiarową obserwację i napisać $X(\omega) = \omega$. Przy tej umowie, milcząco przyjętej w Definicji 2.1.1, rozkład prawdopodobieństwa na przestrzeni $\Omega = \mathcal{X}$ jest tym samym, co rozkład prawdopodobieństwa obserwacji: $\mathbb{P}_\theta(B) = \mathbb{P}_\theta(X \in B)$, dla $B \in \mathcal{F}$. Jest to, co należy podkreślić, *łączny* rozkład wszystkich obserwowanych zmiennych losowych. Szczególny wybór przestrzeni Ω nie ma zasadniczego znaczenia, jest po prostu wygodny.

2.1.3 Uwaga (Ciągłe i dyskretne przestrzenie obserwacji). Skupimy uwagę na dwóch typach przestrzeni statystycznych, które najczęściej pojawiają się w zastosowaniach. Mówimy o modelu ciągłym, jeśli $\mathcal{X}$ jest borelowskim podzbiorem przestrzeni $\mathbb{R}^n$, wyposażonym w σ-ciało $\mathcal{B}$ zbiorów borelowskich i n-wymiarową miarę Lebesgue'a. Model nazywamy dyskretnym, jeśli przestrzeń $\mathcal{X}$ jest skończona lub przeliczalna, wyposażona w σ-ciało $2^{\mathcal{X}}$ wszystkich podzbiorów i miarę liczącą.

Rozkład prawdopodobieństwa obserwacji X najczęściej opisujemy przez gęstość f_θ na przestrzeni $\mathcal{X}$, zależną od parametru $\theta \in \Theta$. W zależności od kontekstu, posługujemy się gęstością względem odpowiedniej miary. W skrócie piszemy $X \sim f_\theta$. Jeśli zmienna X ma skończony lub przeliczalny zbiór wartości $\mathcal{X}$, to

$$f_\theta(x) = \mathbb{P}_\theta(X = x).$$

(jest to gęstość względem miary liczącej). Dla jednowymiarowej zmiennej losowej X o absolutnie ciągłym rozkładzie, f_θ jest „gęstością w zwykłym sensie", czyli względem miary Lebesgue'a. Mamy wówczas dla dowolnego przedziału $[a, b]$,

$$\mathbb{P}_\theta(a \leqslant X \leqslant b) = \int_a^b f_\theta(x)\mathrm{d}x.$$

Jeśli $X = (X_1, \ldots, X_n)$ to rozumiemy, że f_θ jest *łączną* gęstością prawdopodobieństwa na przestrzeni $\mathcal{X} = \mathbb{R}^n$. Dla dowolnego zbioru borelowskiego $B \subseteq \mathbb{R}^n$,

$$\mathbb{P}_\theta(X \in B) = \int_B f_\theta(x)\mathrm{d}x = \int \cdots \int_B f_\theta(x_1, \ldots, x_n)\mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

W szczególnym przypadku, gdy zmienne $X_1, \ldots, X_n$ są niezależne i mają jednakowy rozkład, pozwolimy sobie na odrobinę nieścisłości, oznaczając tym samym symbolem f_θ jednowymiarową gęstość pojedynczej obserwacji i n-wymiarową gęstość całej próbki: $f_\theta(x_1, \ldots, x_n) = f_\theta(x_1) \cdots f_\theta(x_n)$.

Jeśli $T : \mathcal{X} \to \mathbb{R}$, to wartość średnią (oczekiwaną) zmiennej losowej $T(X)$ obliczamy zgodnie ze wzorem

$$\mathbb{E}_\theta T(X) = \begin{cases} \int_{\mathcal{X}} T(x) f_\theta(x)\mathrm{d}x & \text{w przypadku ciągłym;} \\ \\ \sum_{x \in \mathcal{X}} T(x) f_\theta(x) & \text{w przypadku dyskretnym.} \end{cases}$$

Jeśli $\mathcal{X} \subseteq \mathbb{R}^n$, to całka $\int_{\mathcal{X}}$ jest n-wymiarowa, $\mathrm{d}x = \mathrm{d}x_1 \cdots \mathrm{d}x_n$. Podobnie, będziemy używać symboli Var_θ, Cov_θ i podobnych.

Jeśli rodzina rozkładów prawdopodobieństwa $\{\mathbb{P}_\theta; \theta \in \Theta\}$ jest zdefiniowana przez podanie rodziny gęstości $\{f_\theta; \theta \in \Theta\}$ względem pewnej (wspólnej dla wszystkich rozkładów) miary, to mówimy, że przestrzeń statystyczna jest *zdominowana*. Nasze rozważania będą niemal wyłącznie ograniczone do takich przestrzeni.

Przejdziemy teraz do przykładów, które wyjaśnią sens (nieco abstrakcyjnej) Definicji 2.1.1.

2.1.4 Przykład (Statystyczna kontrola jakości, kontynuacja)**.** Powróćmy do Przykładu 1.1.7. Przestrzenią obserwacji jest $\mathcal{X} = \{0,1\}^n$. Obserwacje $X_1, \ldots, X_n$ są zmiennymi losowymi o łącznym rozkładzie prawdopodobieństwa

$$\mathbb{P}_p(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\Sigma x_i}(1-p)^{n-\Sigma x_i},$$

gdzie $x_i \in \{0,1\}$ dla $i = 1, \ldots, n$ i $\sum x_i$ oznacza $\sum_{i=1}^n x_i$. Nieznanym parametrem jest prawdopodobieństwo „sukcesu", $\theta = p$. Przestrzenią parametrów jest $\Theta = [0,1]$. ♢

2.1.5 Przykład (Badanie reprezentacyjne)**.** Powiedzmy, że populacja składa się z r jednostek. Przedmiotem badania jest nieznana liczba m jednostek „wyróżnionych". Na przykład może to być liczba „euroentuzjastów" w populacji wyborców albo liczba palących w populacji studentów. Interesują nas własności całej populacji, ale pełne badanie jest niemożliwe lub zbyt kosztowne. Wybieramy losowo n jednostek spośród r i obserwujemy, ile jednostek wyróżnionych znalazło się wśród wylosowanych. Załóżmy, że stosujemy schemat losowania *bez zwracania* [1]. Najlepiej wyobrazić sobie losowe wybranie n kul z urny zawierającej r kul, w tym m czerwonych i $r-m$ białych. Liczby r i n są znane. Liczba X kul białych wśród wylosowanych jest obserwacją. Zmienną losowa X ma tak zwany *hipergeometryczny* rozkład prawdopodobieństwa:

$$\mathbb{P}_m(X = x) = \binom{m}{x}\binom{r-m}{n-x} \Big/ \binom{r}{n},$$

zależny od parametru $\theta = m$ ze zbioru $\Theta = \{0, 1, \ldots, r\}$. Przestrzenią obserwacji jest zbiór $\mathcal{X} = \{0, 1, \ldots, n\}$. ♢

Parametr θ jest „etykietką" identyfikującą rozkład prawdopodobieństwa. Nie zawsze θ jest liczbą, może wektorem lub nawet funkcją.

2.1.6 Przykład (Model nieparametryczny)**.** Zgodnie z Definicją 1.1.1, ciąg obserwowanych zmiennych losowych $X_1, \ldots, X_n$ stanowi *próbkę* z rozkładu o dystrybuancie F, jeśli

$$\mathbb{P}_F(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n) = F(x_1) \cdots F(x_n).$$

Symbol $\mathbb{P}_F$ przypomina, że dystrybuanta F jest nieznana i odgrywa rolę „nieskończenie wymiarowego parametru". Przestrzenią parametrów jest zbiór wszystkich dystrybuant. Przestrzenią obserwacji jest $\mathcal{X} = \mathbb{R}^n$ [2]. ♢

[1] Próbka wylosowana w ten sposób nie jest próbką w sensie Definicji 1.1.1.

[2] Jest to jedyny w tym skrypcie przykład przestrzeni statystycznej, która nie jest zdominowana.

2.1.7 Przykład (Wypadki)**.** Liczba wypadków drogowych w ciągu tygodnia ma, w dobrym przybliżeniu, rozkład Poissona. Niech $X_1, \ldots, X_n$ oznaczają liczby wypadków w kolejnych tygodniach. Jeśli nic specjalnie się nie zmienia (pogoda jest podobna i nie zaczyna się właśnie okres wakacyjny) to można przyjąć, że każda ze zmiennych X_i ma jednakowy rozkład. Mamy wtedy próbkę z rozkładu Poissona, czyli

$$f_\theta(x_1, \ldots, x_n) = \mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n) = \mathrm{e}^{-\theta n} \frac{\theta^{\Sigma x_i}}{x_1! \cdots x_n!}.$$

Przestrzenią obserwacji jest $\mathcal{X} = \{0, 1, 2, \ldots\}^n$, a przestrzenią parametrów $\Theta =]0, \infty[$. Wiemy, że $\mathbb{E}_\theta X_i = \theta$ i $\mathrm{Var}_\theta X_i = \theta$. ♢

2.1.8 Przykład (Czas życia żarówek)**.** Rozpatrzmy jeszcze jeden przykład z dziedziny statystycznej kontroli jakości. Producent bada partię n żarówek. Interesuje go czas życia, to jest liczba godzin do przepalenia się żarówki. Załóżmy, że czasy życia $X_1, \ldots, X_n$ badanych żarówek stanowią próbkę z rozkładu wykładniczego $\mathrm{Ex}(\theta)$, czyli

$$f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} (\theta \mathrm{e}^{-\theta x_i}) = \theta^n \mathrm{e}^{-\theta \Sigma x_i}.$$

Jest to typowe i dość realistyczne założenie. Mamy tutaj $\mathcal{X} = [0, \infty[^n$ i $\Theta =]0, \infty[$. Zauważmy, że $\mathbb{E}_\theta X_i = 1/\theta$ i $\mathrm{Var}_\theta X_i = 1/\theta^2$. ♢

2.1.9 Przykład (Pomiar z błędem losowym)**.** Powtarzamy niezależnie n razy pomiar pewnej wielkości fizycznej μ. Wyniki poszczególnych pomiarów $X_1, \ldots, X_n$ są zmiennymi losowymi bo przyrząd pomiarowy jest niedoskonały. Najczęściej zakłada się, że każdy z pomiarów ma jednakowy rozkład normalny $\mathrm{N}(\mu, \sigma^2)$. Mamy zatem

$$f_{\mu,\sigma}(x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp\left[-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2 \right].$$

Tutaj rolę parametru θ gra para liczb (μ, σ), gdzie $-\infty < \mu < \infty$ i $\sigma > 0$. Przestrzenią parametrów jest $\Theta = \mathbb{R} \times]0, \infty[$. Oczywiście, przestrzenią obserwacji jest $\mathcal{X} = \mathbb{R}^n$. Wiemy, że $\mathbb{E}_{\mu,\sigma} X_i = \mu$ i $\mathrm{Var}_{\mu,\sigma} X_i = \sigma^2$. ♢

2.2 Statystyki i rozkłady próbkowe

Rozpatrujemy, jak zwykle, przetrzeń statystyczną $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta; \theta \in \Theta\})$. Niech $(\mathcal{T}, \mathcal{A})$ będzie przestrzenią mierzalną (znaczy to, że zbiór $\mathcal{T}$ jest wyposażony w σ-ciało podzbiorów $\mathcal{A}$; zazwyczaj będzie to podzbiór przestrzeni $\mathbb{R}^d$ z σ-ciałem borelowskim).

2.2.1 DEFINICJA. *Mierzalną funkcję $T : \mathcal{X} \to \mathcal{T}$ określoną na przestrzeni obserwacji $\mathcal{X}$ nazywamy* ***statystyką*** *o wartościach w przestrzeni $\mathcal{T}$.*

W Definicji 2.2.1 istotne jest to, że statystyka jest wielkością obliczoną na podstawie danych i nie zależy od nieznanego parametru θ. Będziemy w skrócie pisać $T = T(X)$. Skupiamy uwagę na przypadkach, kiedy przestrzeń $\mathcal{T}$ ma wymiar znacznie mniejszy niż $\mathcal{X}$: staramy się obliczyć taką statystykę $T(X)$ która ma „streścić dane X".

2.2.2 Przykład (Statystyki i inne zmienne losowe)**.** W Przykładzie 2.1.4 (Statystyczna kontrola jakości), $S = \sum_{i=1}^{n} X_i$, a więc liczba prawidłowych wyrobów w próbce *jest* statystyką. Oczywiście, $S : \{0,1\}^n \to \{0,1,\ldots,n\}$. Statystyka S ma dwumianowy rozkład prawdopodobieństwa:

$$\mathbb{P}_p(S = s) = \binom{n}{s} p^s (1-p)^{n-s}.$$

W skrócie napiszemy $S \sim \mathrm{Bin}(n,p)$. Zmienna losowa $(S - np)/\sqrt{np(1-p)}$ *nie jest* statystyką, bo zależy od nieznanego parametru p. Ma w przybliżeniu normalny rozkład prawdopodobieństwa $\mathrm{N}(0,1)$, jeśli n jest duże a $p(1-p)$ nie jest zbyt małe.

W Przykładzie 2.1.7 (Wypadki) sumaryczna liczba wypadków $S = \sum_{i=1}^{n} X_i$ jest statystyką i ma rozkład $\mathrm{Poiss}(n\theta)$.

W Przykładzie 2.1.8 (Żarówki) średnia $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$ jest statystyką i ma rozkład $\mathrm{Gamma}(n, n\theta)$. ◊

Model normalny, wprowadzony w Przykładzie 2.1.9 zasługuje na więcej miejsca. Załóżmy, że $X_1,\ldots,X_n$ *jest próbką z rozkładu* $\mathrm{N}(\mu,\sigma^2)$. Ważną rolę w dalszych rozważaniach odgrywać będą statystyki:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad S = \sqrt{S^2}.$$

Zauważmy, że S^2 różni się od wariancji z próbki $\tilde{S}^2$, o której mówiliśmy w poprzednim rozdziale: mnożnik $1/n$ zastąpiliśmy przez $1/(n-1)$. Rozkład prawdopodobieństwa średniej z próbki jest w modelu normalnym niezwykle prosty: $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$. Zajmiemy się teraz rozkładem statystyki S^2.

Rozkład chi-kwadrat z k stopniami swobody jest to, z definicji, rozkład zmiennej losowej

$$Y = \sum_{i=1}^{k} Z_i^2,$$

gdzie $Z_1,\ldots,Z_k$ są niezależnymi zmiennymi losowymi o rozkładzie $\mathrm{N}(0,1)$. Będziemy pisali symbolicznie $Y \sim \chi^2(k)$.

Uwaga. Rozkłady chi-kwadrat są szczególnej postaci rozkładami Gamma, mianowicie $\chi^2(k) = \text{Gamma}(k/2, 1/2)$ (Zadanie 2.5). Jeśli $Y \sim \chi^2(k)$ to $\mathbb{E}Y = k$ i $\text{Var}Y = 2k$.

Wykresy gęstości kilku rozkładów χ^2 są pokazane na Rysunku 2.1.

2.2.3 Stwierdzenie (Twierdzenie Fishera)**.** *W modelu normalnym, $\bar{X}$ i S^2 są niezależnymi zmiennymi losowymi,*

$$\bar{X} \sim \text{N}(\mu, \sigma^2/n);$$
$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1).$$

Pominiemy dowód, bo w Rozdziale 9 udowodnimy twierdzenie znacznie ogólniejsze. *Niezależność* zmiennych losowych $\bar{X}$ i S^2 *nie jest* oczywista. Zauważmy też, że pojawia się rozkład chi-kwadrat z $n-1$ *stopniami swobody*, chociaż $(n-1)S^2$ jest sumą n kwadratów zmiennych normalnych.

2.2.4 Wniosek. $\mathbb{E}_{\mu,\sigma} S^2 = \sigma^2$ *i* $\text{Var}_{\mu,\sigma} S^2 = 2\sigma^4/(n-1)$.

Rozkład t **Studenta** z k stopniami swobody jest to, z definicji, rozkład zmiennej losowej

$$T = \frac{Z}{\sqrt{Y/k}},$$

gdzie Z i Y są niezależnymi zmiennymi losowymi, $Z \sim \text{N}(0,1)$ i $Y \sim \chi^2(k)$. Będziemy pisali symbolicznie $T \sim \text{t}(k)$. Dwa rozkłady t oraz rozkład normalny są pokazane na Rysunku 2.2.

2.2.5 Wniosek. *W modelu normalnym, zmienna losowa $\sqrt{n}(\bar{X} - \mu)/S$ ma rozkład* $\text{t}(n-1)$.

Rozkład F **Snedecora** z k i m stopniami swobody jest to, z definicji, rozkład zmiennej losowej

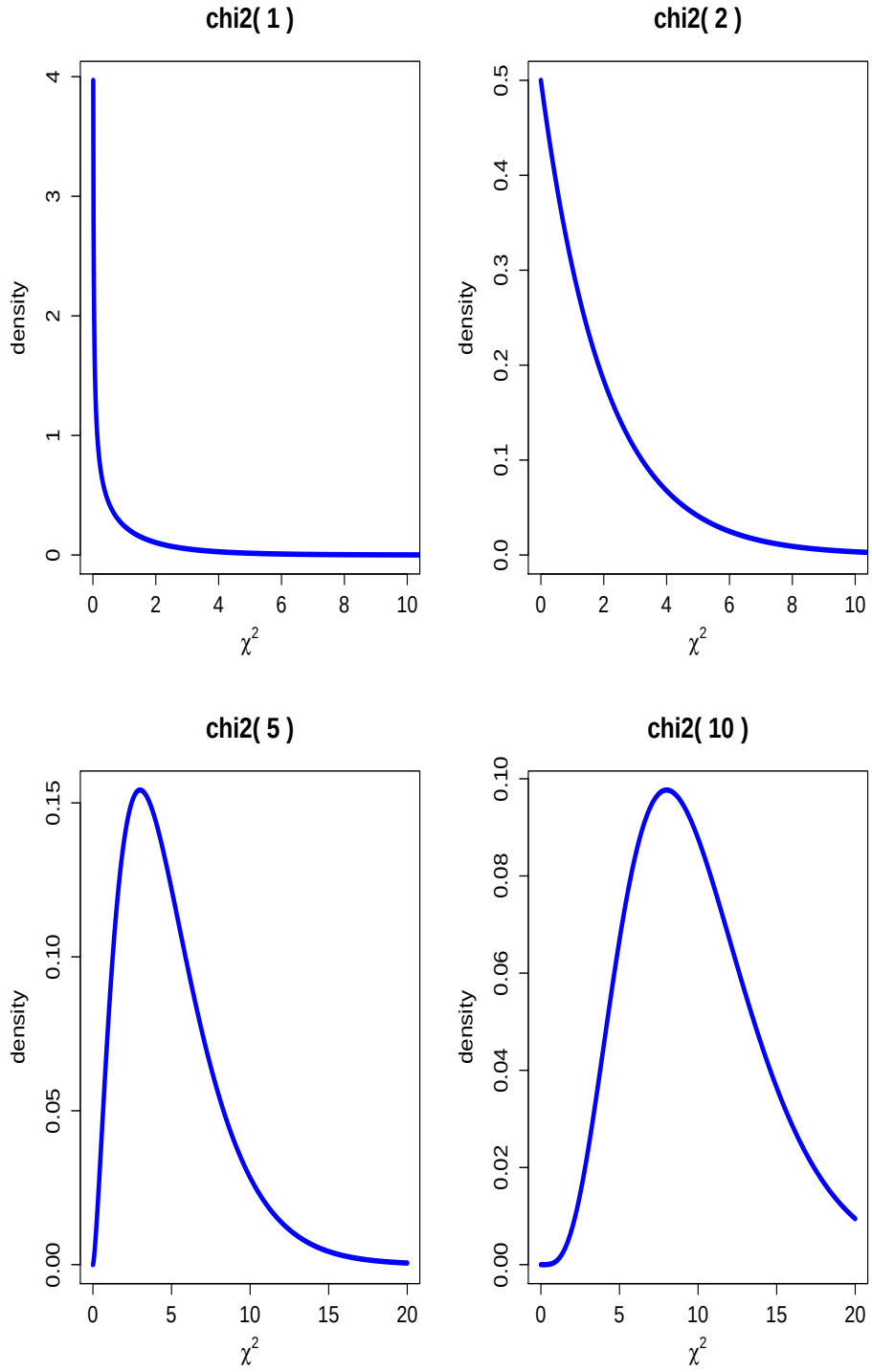
$$R = \frac{Y/k}{U/m},$$

gdzie Y i U są niezależnymi zmiennymi losowymi, $Y \sim \chi^2(k)$ i $U \sim \chi^2(m)$. Będziemy pisali symbolicznie $R \sim \text{F}(k, m)$.

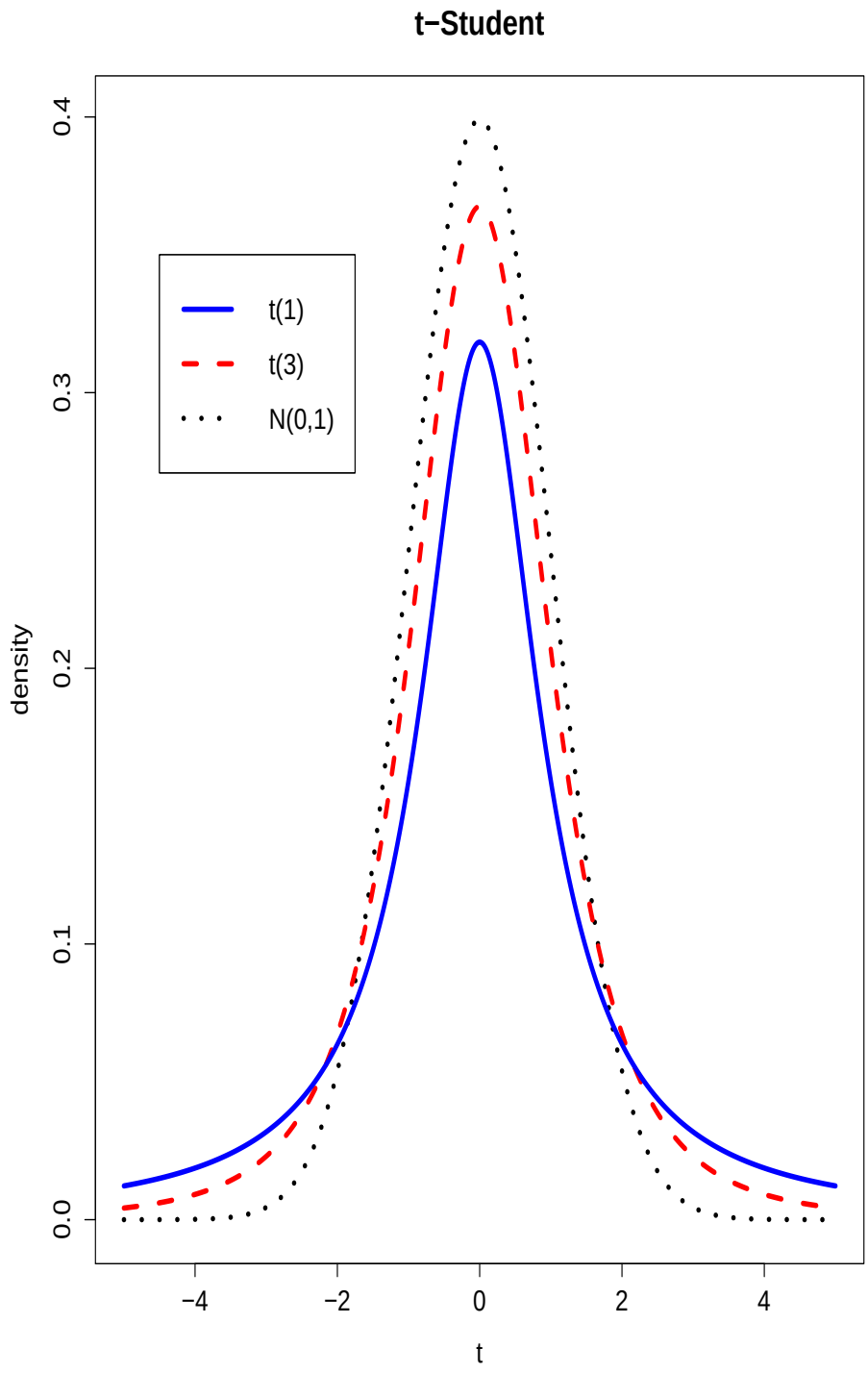
2.2.6 Przykład (Model dwóch próbek)**.** Załóżmy, że obserwujemy niezależne zmienne losowe $X_1, \ldots, X_n$ i $Y_1, \ldots, Y_m$, przy tym $X_i \sim \text{N}(\mu_X, \sigma_X^2)$ i $Y_j \sim \text{N}(\mu_Y, \sigma_Y^2)$ dla $i = 1, \ldots, n$ i $j = 1, \ldots, m$. Statystyki $\bar{X}$ i S_X^2 są określone tak jak poprzednio, dla próbki $X_1, \ldots, X_n$. Podobnie określamy statystyki $\bar{Y}$ i S_Y^2, dla próbki $Y_1, \ldots, Y_m$. Z tego, co powiedzieliśmy wcześniej wynika, że

$$\frac{S_X^2 \sigma_Y^2}{S_Y^2 \sigma_X^2} \sim \text{F}(n-1, m-1).$$

Zauważmy, że zmienna losowa $S_X^2 \sigma_Y^2 / (S_Y^2 \sigma_X^2)$ *nie jest statystyką*, bo zależy nie tylko od obserwacji, ale i od nieznanych paramerów σ_X i σ_Y. Jeśli założymy, że $\sigma_X^2 = \sigma_Y^2$ to *statystyka* S_X^2/S_Y^2 ma rozkład $\text{F}(n-1, m-1)$.



Rysunek 2.1: Rozkłady χ^2 dla różnej liczby stopni swobody.



Rysunek 2.2: Rozkłady t Studenta i rozkład normalny.

Podobnie, jeśli $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ to

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{(k-1)S_X^2 + (m-1)S_Y^2}} \sqrt{\frac{km}{k+m}(k+m-2)} \sim \mathrm{t}(k+m-2).$$

$\Diamond$

2.3 Dostateczność

Rozważmy przestrzeń statystyczną $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta : \theta \in \Theta\})$ i statystykę $T = T(X)$ o wartościach w przestrzeni $(\mathcal{T}, \mathcal{A})$.

2.3.1 DEFINICJA. *Statystykę $T = T(X)$ nazywamy **dostateczną**, jeśli warunkowy rozkład prawdopodobieństwa obserwacji X przy danej wartości statystyki $T = t$ nie zależy od parametru θ, dla każdego $t \in \mathcal{T}$.*

Uwaga. W pewnym uproszczeniu, statystyka jest dostateczna, jeśli prawdopodobieństwo warunkowe

$$(*) \qquad \mathbb{P}_\theta(X \in B | T(X) = t) \quad \text{nie zależy od } \theta,$$

dla dowolnego zbioru $B \in \mathcal{F}$ i (prawie) każdego t. Niestety, ścisłe sformułowanie Definicji 2.3.1 wymaga znajomości ogólnego pojęcia warunkowego rozkładu prawdopodobieństwa i teorii miary. Zwróćmy uwagę, że określenie warunkowego rozkładu poprzez gęstość tutaj się bezpośrednio nie stosuje, bo rozkład X przy danym $T(X) = t$ jest zazwyczaj skupiony na „podprzestrzeni o niższym wymiarze", patrz Zadanie 2.15. Jeśli jednak $\mathcal{X}$ jest przestrzenią dyskretną, to możemy się posłużyć elementarną definicją prawdopodobieństwa warunkowego. W tym przypadku warunek $(*)$ redukuje się do tego, że

$$(**) \qquad \mathbb{P}_\theta(X = x | T(X) = t) \quad \text{nie zależy od } \theta,$$

dla dowolnych t i x (to prawdopodobieństwo jest niezerowe tylko jeśli $T(x) = t$).

Sens Definicji 2.3.1 wyjaśni „doświadczenie myślowe". Wyobraźmy sobie, że statystyk zaobserwował $X = x$, obliczył i zapisał $T(x) = t$, po czym... zgubił dane, czyli stracił x. Może jednak wylosować „sztuczne dane" X' z rozkładu warunkowego obserwacji przy danym $T = t$, ponieważ ten rozkład nie wymaga znajomości θ. Skoro sztuczne dane X' mają ten sam rozkład prawdopodobieństwa co prawdziwe dane X, więc nasz statystyk *nic nie stracił* zapisując t i zapominając x. Stąd właśnie nazwa: statystyka dostateczna zawiera całość informacji o parametrze zawartych w obserwacji.

Załóżmy teraz, że przestrzeń statystyczną $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta : \theta \in \Theta\})$ jest zdominowana (Uwaga 2.1.3), to znaczy rozkłady $\mathbb{P}_\theta$ mają gęstości f_θ. Zwykle są to albo gęstości względem miary Lebesgue'a, albo „gęstości dyskretne" $f_\theta(x) = \mathbb{P}_\theta(X = x)$.

2.3.2 TWIERDZENIE (Kryterum faktoryzacji)**.** *Statystyka $T = T(X)$ jest dostateczna wtedy i tylko wtedy gdy gęstości obserwacji można przedstawić w postaci*

$$f_\theta(x) = g_\theta(T(x))h(x).$$

Dowód. Żeby uniknąć trudności technicznych, ograniczymy się tylko do przypadku dyskretnej przestrzeni $\mathcal{X}$. Jeśli $T(x) = t$ to

$$\mathbb{P}_\theta(X = x|T(X) = t) = \frac{f_\theta(x)}{\sum_{x':T(x')=t} f_\theta(x')}$$

i oczywiście $\mathbb{P}_\theta(X = x|T(X) = t) = 0$ jeśli $T(x) \neq t$. Jeżeli spełniony jest warunek faktoryzacji to natychmiast otrzymujemy, w przypadku $T(x) = t$,

$$\mathbb{P}_\theta(X = x|T(X) = t) = \frac{g_\theta(t)h(x)}{\sum_{x':T(x')=t} g_\theta(t)h(x')} = \frac{h(x)}{\sum_{x':T(x')=t} h(x')}.$$

Odwrotnie, jeśli $\mathbb{P}_\theta(X = x|T(X) = t)$ nie zależy od θ to własność faktoryzacji zachodzi dla $h(x) = \mathbb{P}_\theta(X = x|T(X) = t)$ i $g_\theta(t) = \sum_{x':T(x')=t} f_\theta(x')$. □

2.3.3 Przykład (Ile jest kul w urnie?)**.** Kule w urnie są ponumerowane: $U = \{1, 2, \ldots, r\}$ ale r jest nieznane. Pobieramy próbkę n kul, bez zwracania. Niech S oznacza losowy zbiór numerów a $\max(S)$ – największy spośród nich. Prawdopodobieństwo wylosowania zbioru $s \subset U$ jest równe

$$\mathbb{P}_r(S = s) = \frac{\mathbb{1}(r \geqslant \max(s))}{\binom{r}{n}} = \begin{cases} 1/\binom{r}{n} & \text{jeśli } r \geqslant \max(s), \\ 0 & \text{jeśli } r < \max(s). \end{cases}$$

Stąd widać, że $\max(S)$ jest statystyką dostateczną. W czasie II wojny światowej alianci notowali seryjne numery zdobytych czołgów niemieckich w celu oszacowania liczby produkowanych przez nieprzyjaciela czołgów. Rozważany schemat urnowy jest uproszczonym modelem takiej sytuacji. ◊

2.3.4 Przykład (Statystyki dostateczne w poprzednich przykładach)**.** W Przykładzie 2.1.4 (Schemat Bernoulliego), liczba sukcesów $S = \sum_{i=1}^n X_i$ jest statystyką dostateczną.

W Przykładzie 2.1.7 (model Poissona) suma obserwacji $S = \sum_{i=1}^n X_i$ jest statystyką dostateczną.

W Przykładzie 2.1.8 (model wykładniczy) średnia $\bar{X} = (1/n)\sum_{i=1}^n X_i$ jest statystyką dostateczną.

W Przykładzie 2.1.9 (model normalny z nieznanymi μ i σ) $(\bar{X}, S^2)$ jest dwuwymiarową statystyką dostateczną. ◊

2.4 Rodziny wykładnicze

Tak jak poprzednio, rozważamy model statystyczny, a więc rodzinę rozkładów prawdopodobieństwa na przestrzeni obserwacji $\mathcal{X}$.

2.4.1 DEFINICJA. *Rodzina rozkładów prawdopodobieństwa* $\{\mathbb{P}_\theta : \theta \in \Theta\}$ *jest* ***rodziną wykładniczą*** *jeśli rozkłady* $\mathbb{P}_\theta$ *mają, względem pewnej miary na* $\mathcal{X}$*, gęstości* f_θ *postaci:*

$$f_\theta(x) = \exp\Big(\sum_{j=1}^{k} T_j(x)\psi_j(\theta) + \psi_0(\theta)\Big)h(x), \quad (\theta \in \Theta).$$

Podkreślmy, że w tej definicji wymagamy, żeby istniały gęstości względem jednej miary wspólnej dla wszystkich θ. W większości zastosowań spotykamy, jak zwykle, albo gęstości względem miary Lebesgue'a, albo „gęstości dyskretne" $f_\theta(x) = \mathbb{P}_\theta(X = x)$. Bez straty ogólności można zakładać, że funkcje $T_1(x), \ldots, T_k(x)$ są liniowo niezależne. To założenie będze w dalszym ciągu obowiązywać. Zauważmy prostą konsekwencję Definicji 2.4.1. Zbiór $\{x : f_\theta > 0\}$, który nazywamy nośnikiem rozkładu $\mathbb{P}_\theta$, [3] jest taki sam dla wszystkich θ.

2.4.2 Przykład. Rodzina rozkładów jednostajnych $\{\mathrm{U}(0,\theta) : \theta > 0\}$ *nie jest* rodziną wykładniczą. Ponieważ

$$f_\theta(x) = \frac{1}{\theta}\mathbb{1}(0 \leqslant x \leqslant \theta),$$

więc nośnikiem rozkładu $\mathrm{U}(0,\theta)$ jest przedział $[0,\theta]$, który oczywiście zależy od θ. ◊

2.4.3 Przykład. Rodzina rozkładów wykładniczych $\{\mathrm{Ex}(\theta) : \theta > 0\}$ jest rodziną wykładniczą, bo gęstości możemy napisać w następującej postaci:

$$f_\theta(x) = \theta e^{-\theta x} = \exp(-\theta x + \log\theta).$$

Nośnikiem każdego rozkładu wykładniczego jest ten sam przedział $[0,\infty[$. ◊

2.4.4 Przykład. Rodzina rozkładów $\{\mathrm{Poiss}(\theta) : \theta > 0\}$ jest rodziną wykładniczą, bo

$$f_\theta(x) = e^{-\theta}\frac{\theta^x}{x!} = \exp\Big(-\theta + x\log\theta\Big)\frac{1}{x!}.$$

Oczywiście, każdy rozkład Poissona ma nośnik $\{0,1,2,\ldots\}$. ◊

2.4.5 Przykład. Rodzina przesuniętych rozkładów Cauchy'ego o gęstościach

$$f_\theta(x) = \frac{1}{\pi(1+(x-\theta)^2)},$$

$\theta \in]-\infty,\infty[$, nie jest rodziną wykładniczą, bo funkcja $\log f_\theta(x) = -\log\pi - \log(1+(x-\theta)^2)$ nie da się przedstawić w postaci sumy iloczynów $\sum_{j=1}^{k} T_j(x)\psi_j(\theta) + \psi_0(\theta)$. ◊

[3] Pozwalamy tu sobie na drobne uproszczenie, bo gęstość rozkładu prawdopodobieństwa jest wyznaczona jednoznacznie tylko prawie wszędzie.

2.4.6 Przykład. Rodzina rozkładów $\{\text{Gamma}(\alpha,\lambda) : \alpha > 0, \lambda > 0\}$ jest rodziną wykładniczą.

$$f_{\alpha,\lambda}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\lambda x} = \exp\left(-\lambda x + (\alpha-1)\log x + \log\frac{\lambda^\alpha}{\Gamma(\alpha)} \right)$$

Oczywiście, wspólnym nośnikiem wszystkich rozkładów Gamma jest przedział $]0,\infty[$. ◊

Inne przykłady rodzin wykładniczych to między innymi rodzina rozkładów normalnych $\{\text{N}(\mu,\sigma) : -\infty < \mu < \infty, \lambda > 0\}$, rozkładów $\{\text{Beta}(\alpha,\beta) : \alpha,\beta > 0\}$, rodzina rozkładów dwumianowych $\{\text{Bin}(n,\theta) : 0 < \theta < 1\}$, ujemnych dwumianowych i wiele innych. Przejdźmy do omówienia kilku ciekawych własnosci rodzin wykładniczych.

2.4.7 Stwierdzenie. *Jeżeli $X_1,\ldots,X_n \sim_{\text{iid}} f_\theta$ jest próbką z rozkładu należącego do rodziny wykładniczej, to k-wymiarowy wektor*

$$\left(\sum_{i=1}^n T_1(X_i), \ldots, \sum_{i=1}^n T_k(X_i) \right)$$

jest statystyką dostateczną.

Dowód. Jeżeli f_θ ma postać taką jak w Definicji 2.4.1, to łaczna gęstość wektora obserwacji jest następująca:

$$\begin{aligned} f_\theta(x_1,\ldots,x_n) &= \prod_{i=1}^n \exp\left(\sum_{j=1}^k T_j(x_i)\psi_j(\theta) + \psi_0(\theta) \right) h(x_i) \\ &= \exp\left(\sum_{j=1}^k \sum_{i=1}^n T_j(x_i)\psi_j(\theta) + n\psi_0(\theta) \right) \prod_{i=1}^n h(x_i). \end{aligned}$$

Wystarczy teraz skorzystać z kryterium faktoryzacji (Twierdzenie 2.3.2). □

Zwróćmy uwagę, że dla wymiar statystyki dostatecznej w powyższym stwierdzeniu jest równy k, niezależnie od rozmiaru próbki n. Dla próbki z rodziny wykładniczej możliwa jest bardzo radykalna redukcja danych bez straty informacji. Zauważmy jeszcze, że k w Definicji 2.4.1 wydaje się być związane z wymiarem przestrzeni parametrów. W Przykładach 2.4.3 i 2.4.4 mieliśmy jednoparametrowe rodziny wykładnicze, w Przykładzie 2.4.6 – dwuparametrową rodzinę. Staje się to bardziej przejrzyste, jeśli posłużymy się tak zwaną naturalną parametryzacją rodzin wykładniczych. Przyjmijmy wektor

$$\psi = (\psi_1,\ldots,\psi_k) = (\psi_1(\theta),\ldots,\psi_k(\theta))$$

za nowy parametr, który identyfikuje rozkłady prawdopodobieństwa rozpatrywanej rodziny. Nieco nadużywając oznaczeń możemy napisać

$$f_\psi(x) = \exp\left(\sum_{j=1}^k T_j(x)\psi_j - b(\psi) \right) h(x), \tag{2.4.8}$$

gdzie

$$b(\psi) = \log \int_{\mathcal{X}} \exp\left(\sum_{j=1}^{k} T_j(x)\psi_j\right) h(x)\mathrm{d}x.$$

Jeśli istnieje wzajemnie jednoznaczna odpowiedniość pomiędzy „starym parametrem" $\theta \in \Theta$ i „nowym parametrem" ψ, to wybór jednej lub drugiej parametryzacji jest tylko kwestią wygody.

2.4.9 Przykład. Rozkłady dwumianowe $\mathrm{Bin}(\theta, n)$ mają gęstości postaci

$$\begin{aligned} f_\theta(x) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} = \exp\left(x\log\theta + (n-x)\log(1-\theta)\right)\binom{n}{x} \\ &= \exp\left(x\log\frac{\theta}{1-\theta} + n\log(1-\theta)\right)\binom{n}{x}. \end{aligned}$$

Naturalnym parametrem jest

$$\psi = \log\frac{\theta}{1-\theta}$$

zaś $b(\psi) = n\log(1+e^\psi)$. Zauważmy, że $\theta/(1-\theta)$ jest tak zwanym „ilorazem szans": stosunkiem prawdopodobieństwa sukcesu do prawdopodobieństwa porażki. Funkcja Jeśli θ zmienia się w przedziale $]0,1[$ to ψ przebiega przedział $]\infty,\infty[$. Naturalną przestrzenią parametrów jest więc cała prosta rzeczywista. ◊

2.4.10 Uwaga. Mówimy, że rodzina wykładnicza jest **regularna**, jeśli przestrzeń naturalnych parametrów $\{\psi(\theta) : \theta \in \Theta\}$, traktowana jako podzbiór $\mathbb{R}^k$, ma niepuste wnętrze. Ważną własnością regularnych rodzin wykładniczych jest dopuszczalność „różniczkowania pod znakiem całki". Jeśli $U : \mathcal{X} \to \mathbb{R}$ jest statystyką to

$$\frac{\partial}{\partial\psi_j}\int_{\mathcal{X}} U(x)f_\psi(x)\mathrm{d}x = \int_{\mathcal{X}} U(x)\frac{\partial}{\partial\psi_j}f_\psi(x)\mathrm{d}x,$$

jeśli ψ jest punktem wewnętrznym naturalnej przestrzeni parametrów i całka po lewej stronie jest dobrze określona. Co więcej, podobna własność zachodzi dla pochodnych wyższych rzędów. Oczywiście, jeśli funkcje $\psi_j(\theta)$ są odpowiednio gładkie, to możemy bezpiecznie „różniczkować pod znakiem całki" również względem θ. W następnym rozdziale takie operacje rachunkowe będą odgrywały ważną rolę.

2.5 Zadania

2.1. Rozpatrzmy proces statystycznej kontroli jakości przyjmując te same założenia co w Przykładzie 2.1.4 z tą różnicą, że obserwujemy kolejne wyroby do momentu gdy natrafimy na k wybrakowanych, gdzie k jest ustaloną z góry liczbą. Zbudować model statystyczny.

2.2. Uogólnić rozważania z Przykładu 2.1.5 (badanie reprezentacyjne), uwzględniając więcej niż jeden rodzaj jednostek „wyróżnionych". Powiedzmy, że mamy w urnie m_1 kul czerwonych, m_2 zielonych i $r - m_1 - m_2$ białych, gdzie r jest znaną liczbą, a m_1 i m_2 są nieznane i są przedmiotem badania. Opisać dokładnie odpowiedni model statystyczny.

2.3. Obliczyć rozkład prawdopodobieństwa zmiennej losowej Z^2, jeśli $Z \sim \mathrm{N}(0,1)$ (obliczyć bezpośrednio dystrybuantę i gęstość rozkładu $\chi^2(1)$).

2.4. Obliczyć rozkład prawdopodobieństwa zmiennej losowej $Z_1^2 + Z_2^2$, jeżeli $Z_i \sim \mathrm{N}(0,1)$ są niezależne dla $i = 1, 2$ (obliczyć bezpośrednio dystrybuantę i gęstość rozkładu $\chi^2(2)$).

2.5. Korzystając z Zadania 2.3 oraz z własności rozkładów gamma, udowodnić Uwagę 2.2: gęstość zmiennej losowej $Y \sim \chi^2(k)$ ma postać

$$f_Y(y) = \frac{1}{2^{k/2}\Gamma(k/2)} y^{k/2-1} \mathrm{e}^{-y/2}, \quad (y > 0).$$

2.6. Udowodnić zbieżność rozkładów: $\mathrm{t}(k) \to_{\mathrm{d}} \mathrm{N}(0,1)$ dla $k \to \infty$.

2.7. Udowodnić wzór dotyczący rozkładu t-Studenta na końcu Przykładu 2.2.6.

2.8. Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu (Weibulla) o gęstości

$$f_\theta(x) = \begin{cases} 3\theta x^2 \mathrm{e}^{-\theta x^3} & \text{dla } x > 0; \\ 0 & \text{dla } x \leqslant 0, \end{cases}$$

gdzie $\theta > 0$ jest nieznanym parametrem. Znaleźć jednowymiarową statystykę dostateczną.

2.9. Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu Gamma(α, λ). Znaleźć dwuwymiarową statystykę dostateczną, zakładając że $\theta = (\alpha, \lambda)$ jest nieznanym parametrem.

2.10. Rozważamy rodzinę przesuniętych rozkładów wykładniczych o gęstości

$$f_\mu(x) = \begin{cases} \mathrm{e}^{-(x-\mu)} & \text{dla } x \geqslant \mu; \\ 0 & \text{dla } x < \mu. \end{cases}$$

Niech $X_1, \ldots, X_n$ będzie próbką losową z takiego rozkładu. Znaleźć jednowymiarową statystykę dostateczną dla parametru μ.

2.11. Rozważamy rodzinę przesuniętych rozkładów wykładniczych z parametrem skali o gęstości

$$f_{\mu,\lambda}(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda(x-\mu)} & \text{dla } x \geqslant \mu; \\ 0 & \text{dla } x < \mu. \end{cases}$$

Niech $X_1, \ldots, X_n$ będzie próbką losową z takiego rozkładu. Znaleźć dwuwymiarową statystykę dostateczną dla parametru (μ, λ).

2.12. Rozważamy rodzinę rozkładów na przestrzeni $\{0, 1, 2, \ldots\}$:

$$f_\theta(x) = \mathbb{P}_\theta(X = x) = \begin{cases} \theta & \text{dla } x = 0; \\ (1-\theta)/2^x & \text{dla } x \in \{1, 2, \ldots\}. \end{cases}$$

gdzie $\theta \in]0, 1[$ jest nieznanym parametrem. Niech $X_1, \ldots, X_n$ będzie próbką losową z wyżej podanego rozkładu. Znaleźć jednowymiarową statystykę dostateczną.

2.13. Niech $X_1, \ldots, X_n$ będzie schematem Bernoulliego z prawdopodobieństwem sukcesu θ. Obliczyć warunkowy rozkład prawdopodobieństwa zmiennych losowych $X_1, \ldots, X_n$ przy danym $S = s$, gdzie $S = \sum_{i=1}^n X_i$ jest liczbą sukcesów. Zinterpretować fakt, że statystyka S jest dostateczna.

2.14. Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu Poiss(θ). Obliczyć warunkowy rozkład prawdopodobieństwa zmiennych losowych $X_1, \ldots, X_n$ przy danym $S = s$, gdzie $S = \sum_{i=1}^n X_i$. Zinterpretować fakt, że statystyka S jest dostateczna.

2.15. Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu Ex(θ). Niech $S = \sum_{i=1}^n X_i$. Pokazać, że rozkład warunkowy $(X_1, \ldots, X_{n-1})$ przy danym $S = s$ jest jednostajny na sympleksie $\{(x_1, \ldots, x_{n-1}) : x_i \geqslant 0, \sum_{i=1}^{n-1} x_i \leqslant s\}$. Zinterpretować fakt, że statystyka S jest dostateczna.

2.16. Znaleźć rozkład zmiennej losowej

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$$

w modelu normalnym. Porównać z twierdzeniem Fishera (Stwierdzenie 2.2.3).

2.17. (Ciąg dalszy). Wyprowadzić tożsamość

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

Jaki jest rozkład prawdopodobieństwa pierwszego i drugiego składnika po prawej stronie?

2.18. Rozważmy jednoparametryczną wykładniczą rodzinę rozkładów z gęstościami danymi wzorem $f_\psi(x) = \exp\big(T(x)\psi - b(\psi)\big)h(x)$. Pokazać, że

$$\mathbb{E}_\psi T(X) = \frac{\partial b(\psi)}{\partial \psi}.$$

2.19. (Ciąg dalszy). Pokazać, że

$$\mathrm{Var}_\psi T(X) = \frac{\partial^2 b(\psi)}{\partial \psi^2}.$$

2.20. (Ciąg dalszy). Pokazać, że

$$\mathbb{E}_\psi \exp\big(rT(X)\big) = \exp\big(b(\psi + r) - b(\psi)\big).$$