Statystyka

Wojciech Niemiro [1]

Wersja Robocza: 16 stycznia 2011

[1]Wydział Matematyki i Informatyki, Uniwersytet Mikołaja Kopernika, Toruń oraz Instytut Matematyki i Mechaniki, Uniwersytet Warszawski, wniem@mat.uni.torun.pl, wniem@mimuw.edu.pl

Spis treści

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 Ω 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 $X_1, \dots X_n \sim_{\text{iid}} \mathrm{N}(\mu, \sigma^2)$. Mówi się także, że $X_1, X_2, \dots, X_n$ jest próbką z rozkładu fikcyjnej zmiennej losowej $X \sim F$.

1.1.2 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.*

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

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

nazywamy ***dystrybuantą empiryczną.***

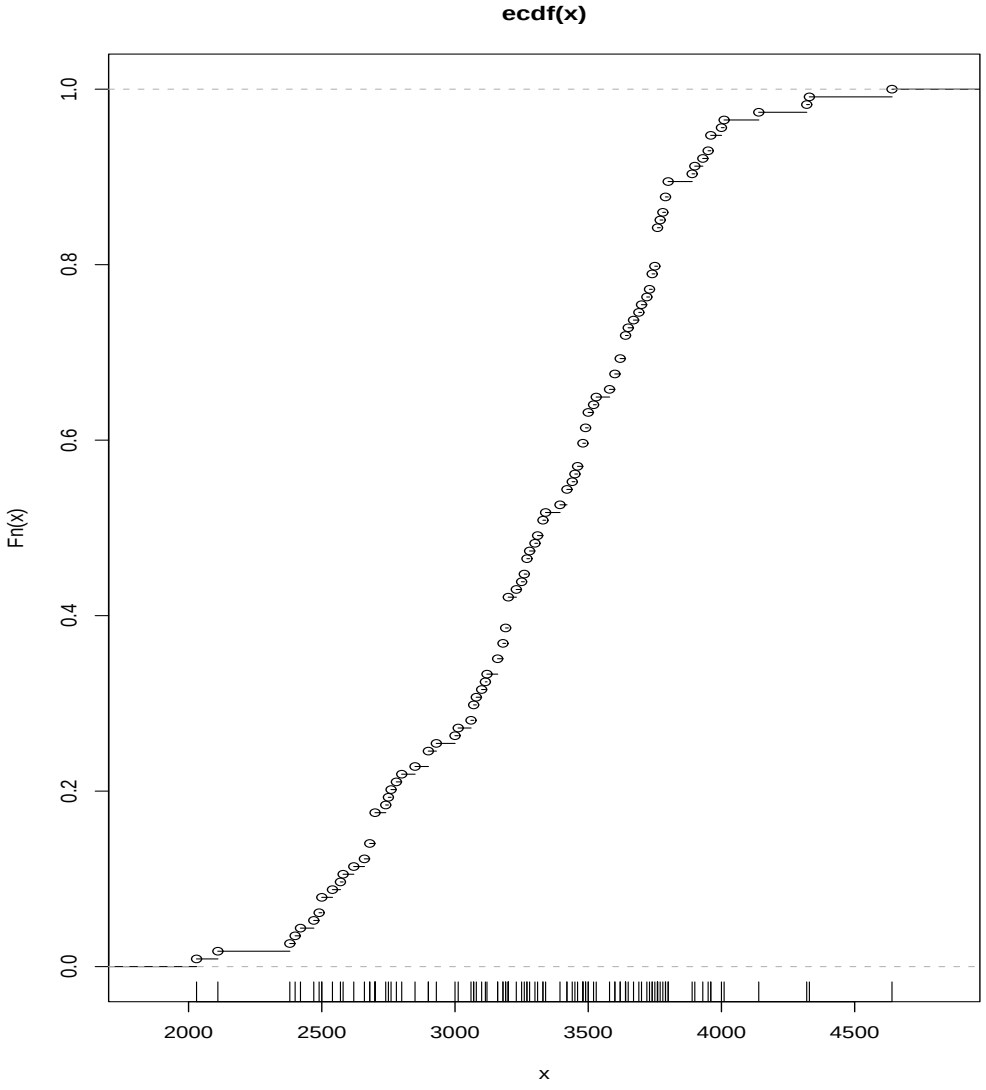
Gdy chcemy podkreślić, że próbka ma rozmiar n, to piszemy $\hat{F}_n$ zamiast $\hat{F}$. Traktujemy $\hat{F}$ jako „empiryczny odpowiednik" nieznanej dystrybuanty F.

1.1.4 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						

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

Dane traktujemy jako próbkę z rozkładu prawdopodobieństwa zmiennej losowej $X =$ „waga noworodka losowo wybranego z populacji". Na podstawie tej próbki możemy narysować dystrybuantę empiryczną $\hat{F}$.



Rysunek został wykonany w R przez zastosowanie instrukcji `plot.ecdf(noworodki)`; `rug(noworodki)`, gdzie `noworodki` jest wektorem zawierającym 114 liczb.

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 $\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i(\omega) \leqslant x)$. *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:

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

1.1.5 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 1.1.3, to znaczy, zależy od *początkowych* zmiennych $X_1, \ldots, X_n$. Rozpatrujemy teraz *ciąg* dystrybuant empirycznych $\hat{F}_1, \hat{F}_2, \ldots, \hat{F}_n, \ldots$.

1.1.6 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 $\mathbb{I}(X_1 \leqslant x), \ldots, \mathbb{I}(X_n \leqslant x), \ldots$ są niezależne i mają jednakowy rozkład prawdopodobieństwa: $\mathbb{I}(X_n \leqslant x)$ przyjmuje wartość 1 z prawdopodobieństwem $F(x)$ lub wartość 0 z prawdopodobieństwem $1-F(x)$. Oczywiście, $\mathbb{E}\mathbb{I}(X_n \leqslant x) = F(x)$. Z Mocnego Prawa Wielkich (MPWL) Liczb 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 $F(x)$. □

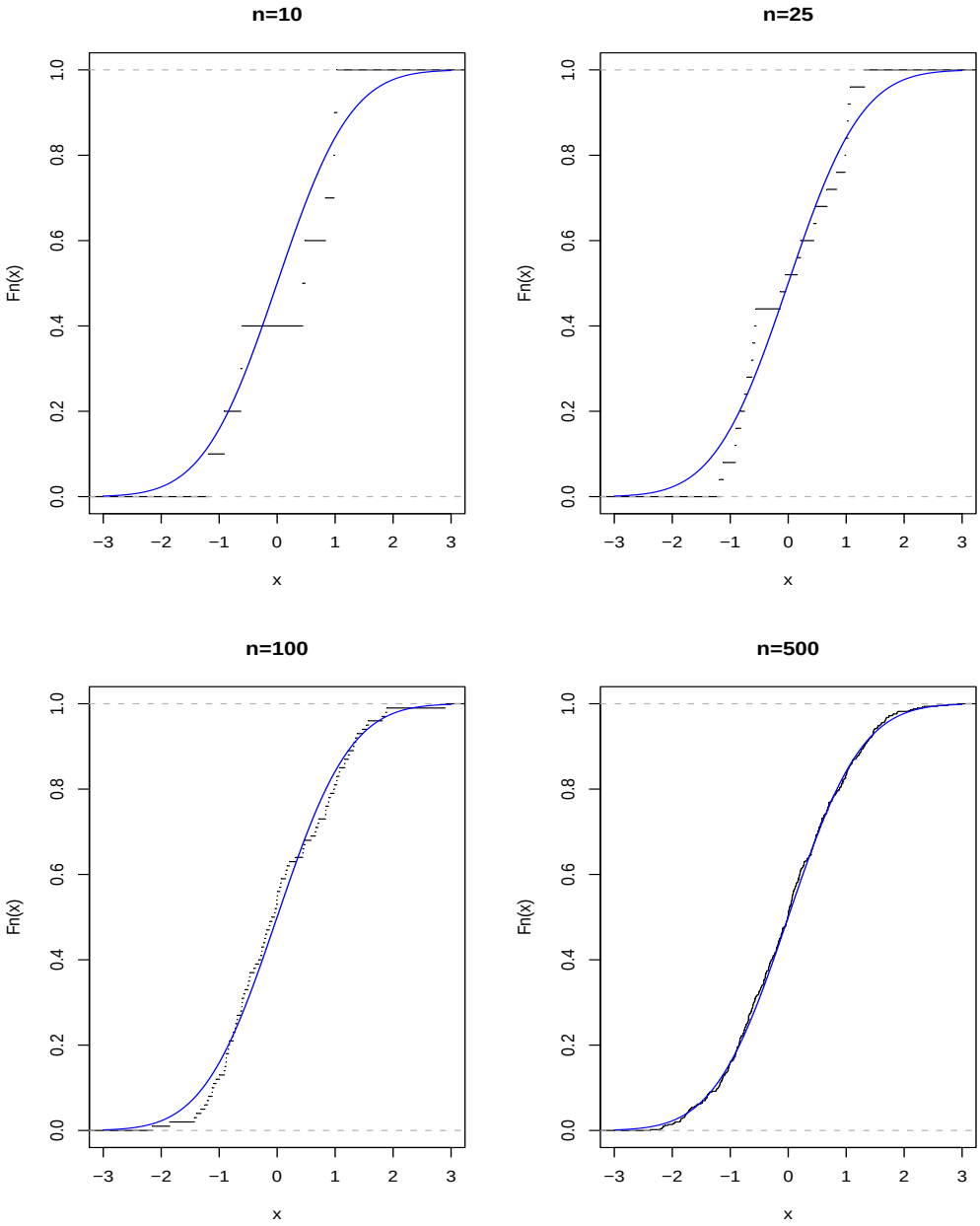
Istnieje mocniejsza wersja poprzedniego stwierdzenia, którą przytoczymy bez dowodu. Można pokazać, że zbieżność $\hat{F} \to F$ zachodzi *jednostajnie* z prawdopodobieństwem 1.

1.1.7 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 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). Przy okazji wspomnijmy, że próbkę ze standardowego rozkładu normalnego można w R wygenerować (symulować) jedną prostą instrukcją:

`X <- rnorm(n)` lub bardziej jawnie `X <- rnorm(n,mean=0,sd=1)`.

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

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

Określone w ten sposób odwzorowanie $\hat{P} : \mathcal{B} \times \Omega \to \mathbb{R}$ ($\mathcal{B}$ 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, $\hat{P}(]-\infty, x]) = \hat{F}(x)$.

1.1.8 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

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

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**

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

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 $\hat{m}_2 = \hat{a}_2 - \hat{a}_1^2$.

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

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

Oczywiście, 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

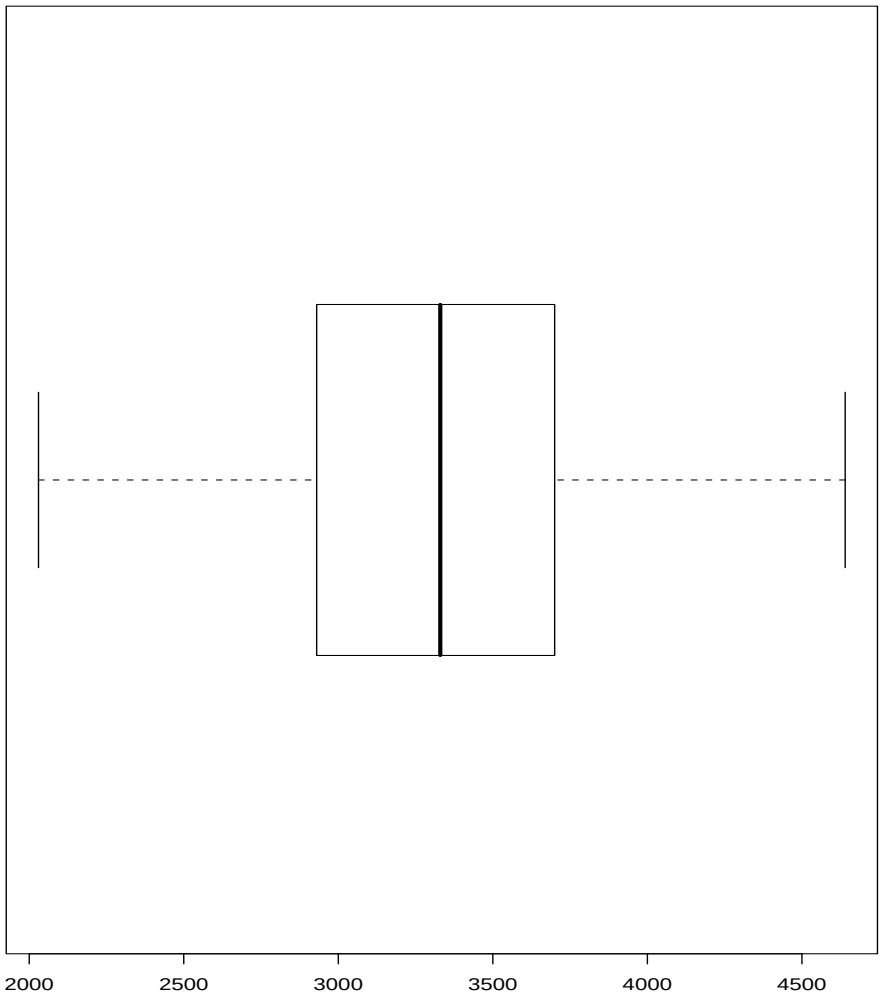
1.2.1 PRZYKŁAD (Waga noworodków, kontynuacja). Dla naszej „niemowlęcej" próbki z Przykładu 1.1.4 mamy

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

Jak już zauważyliśmy poprzednio, $\hat{\mathrm{med}} = 3330$. Kwartyle próbkowe są równe $Q_1 = \hat{\xi}_{1/4} = 2947.5$ [2] i $Q_3 = \hat{\xi}_{3/4} = 3697.5$.

Medianę, kwartyle, minimum i maksimum świetnie widać na tak zwanym „wykresie pudełkowym", instrukcja `boxplot(noworodki,horizontal=TRUE)`:

[2] Zgodnie z naszą definicją $\hat{\xi}_{1/4} = X_{29:114} = 2930$; określenie kwantyla w R nieco różni się od naszego, ale nie ma to zasadniczego znaczenia.



A tak wyląda „podsumowanie" naszych danych:

```
summary(noworodki)
 Min.  1st Qu.  Median  Mean  3rd Qu.  Max.
 2030    2948    3330   3302    3698   4640
```

1.3 Estymatory gęstości

Na zakończenie naszych wstępnych rozważań zastanówmy się, co może być próbkowym odpowiednikiem *gęstości prawdopodobieństwa* zmiennej losowej X o rozkładzie *absolutnie ciągłym*. Bezpośredniego odpowiednika *nie ma*, bo rozkład empiryczny jest dyskretny i nie ma gęstości względem miary Lebesgue'a. Możemy (całkiem heurystycznie) przyjąć, że gęstość zmiennej losowej X jest w przybliżeniu stała na pewnych przedziałach. Przyjmijmy, że $a < X < b$ i podzielmy odcinek $[a, b]$ na k odcinków o końcach w punktach $a = c_0 < c_1 < \cdots < c_k = b$. Połóżmy

$$\hat{f}(x) = \frac{1}{n(c_j - c_{j-1)}} \sum_{i=1}^{n} \mathbb{I}(c_{j-1} < X_i \leq c_j), \qquad \text{dla } c_{j-1} < x \leq c_j.$$

Funkcja $\hat{f}$ jest pewnym przybliżeniem gęstości f rozkładu, z którego pochodzi próbka. Widać, że $\hat{F}(c_j) = \int_a^{c_j} \hat{f}(x)\mathrm{d}x$. Graficzne przedstawienie funkcji $\hat{f}$ jest znane pod nazwą **histogramu**.

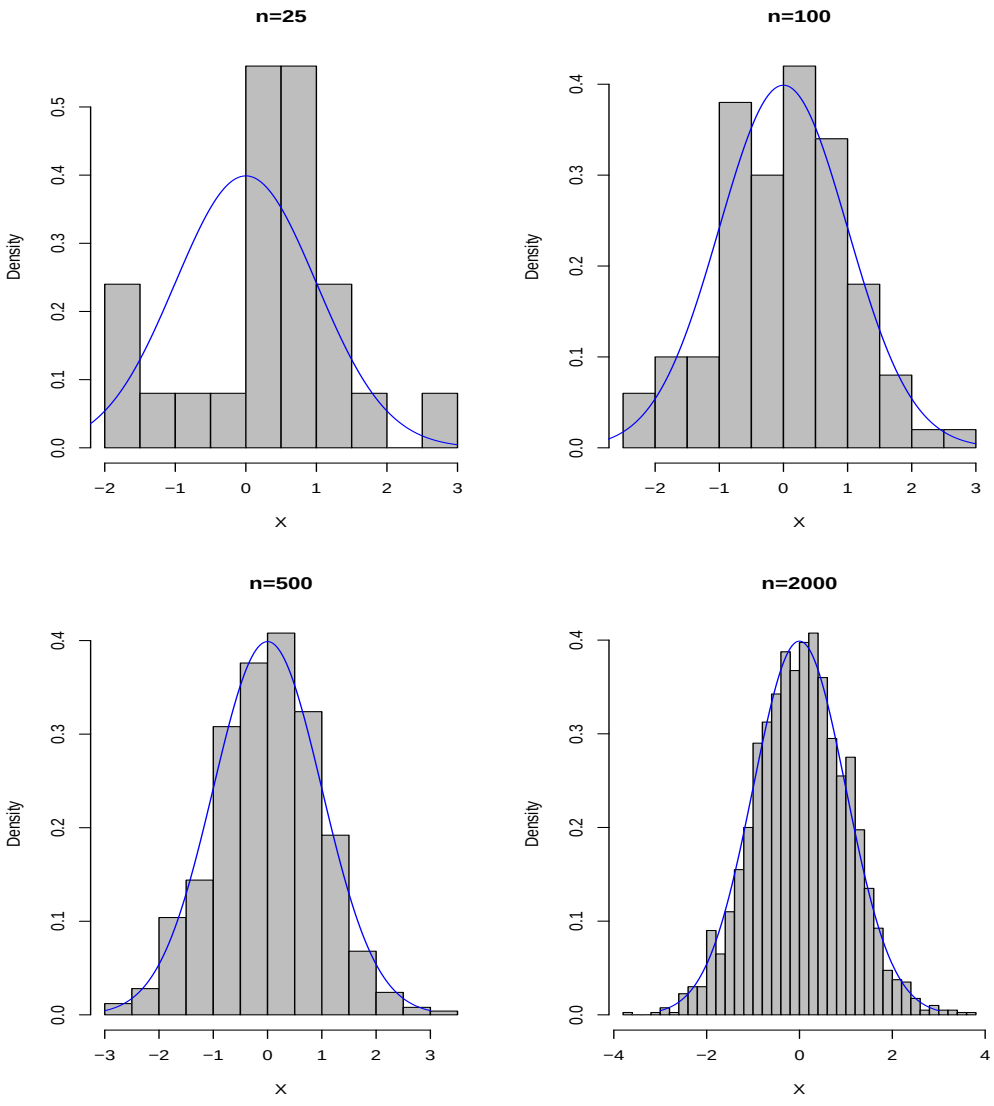
W dalszym ciągu załóżmy, że przedziały histogramu są jednakowej szerokości, $c_j - c_{j-1} = h$ dla wszystkich j. Sformułujemy odpowiednik Stwierdzenia 1.1.6 dla histogramów.

1.3.1 Stwierdzenie. *Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu o gęstości f, o rozmiarze rosnącym do nieskończoności, $n \to \infty$. Rozpatrzmy histogram $\hat{f}_n^{\text{hist}}$ zbudowany na podstawie próbki rozmiaru n i mający przedziały jednakowej szerokości h_n. Załóżmy, że $h_n \to 0$ i $nh_n \to \infty$. Dla każdego punktu $x \in \mathbb{R}$ takiego, że f jest ciągła w x mamy*

$$\hat{f}_n^{\text{hist}} \to_{\mathrm{P}} f(x), \quad (n \to \infty).$$

Dowód pominiemy, choć nie jest on trudny, porównaj Zadanie 11. Skomentujmy założenia. Warunek $h_n \to 0$ zapewnia, że prostokąty histogramu są coraz węższe i mogą aproksymować funkcję ciągłą f. Warunek $nh_n \to 0$ zapewnia, że coraz więcej punktów wpada do pojedynczego przedziału.

Zamiast dowodu, zilustrujemy Stwierdzenie 1.3.1 rysunkiem. Wywołanie R-owskiej funkcji ma postać `hist(X,breaks="FD",prob=TRUE)`, gdzie `X` jest wektorem danych. Parametr `"FD"` wskazuje, że szerokość przedziałów h jest dobierana zgodnie z algorytmem Friedmana - Diaconisa (domyślnie, przedziały są jednakowej szerokości). Parametr `prob=TRUE` powoduje takie wyskalowanie osi pionowej, że pole pod histogramem jest 1.



Często są używane **jądrowe estymatory** gęstości. Idea jest podobna jak dla histogramów, ale w ciekawy sposób zmodyfikowana. Niech k będzie ustaloną gęstością prawdopodobieństwa, to znaczy $k(x) \geqslant 0$ i $\int_{-\infty}^{\infty} k(x)\mathrm{d}x = 1$. Powiedzmy, że k jest funkcją parzystą z maksimum w zerze. Często używa się gęstości normalnej: $k(x) = (2\pi)^{-1/2}\exp(-x^2/2)$, ale wybór jądra jest dość dowolny i nie ma znaczenia zasadniczego. Nazwijmy k jądrem. Zauważmy, że dla dowolnego $h > 0$ funkcja $k(x/h)/h$ jest też gęstością. Dla małych h ta funkcja jest coraz bardziej „skupiona wokół zera". Niech teraz

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right).$$

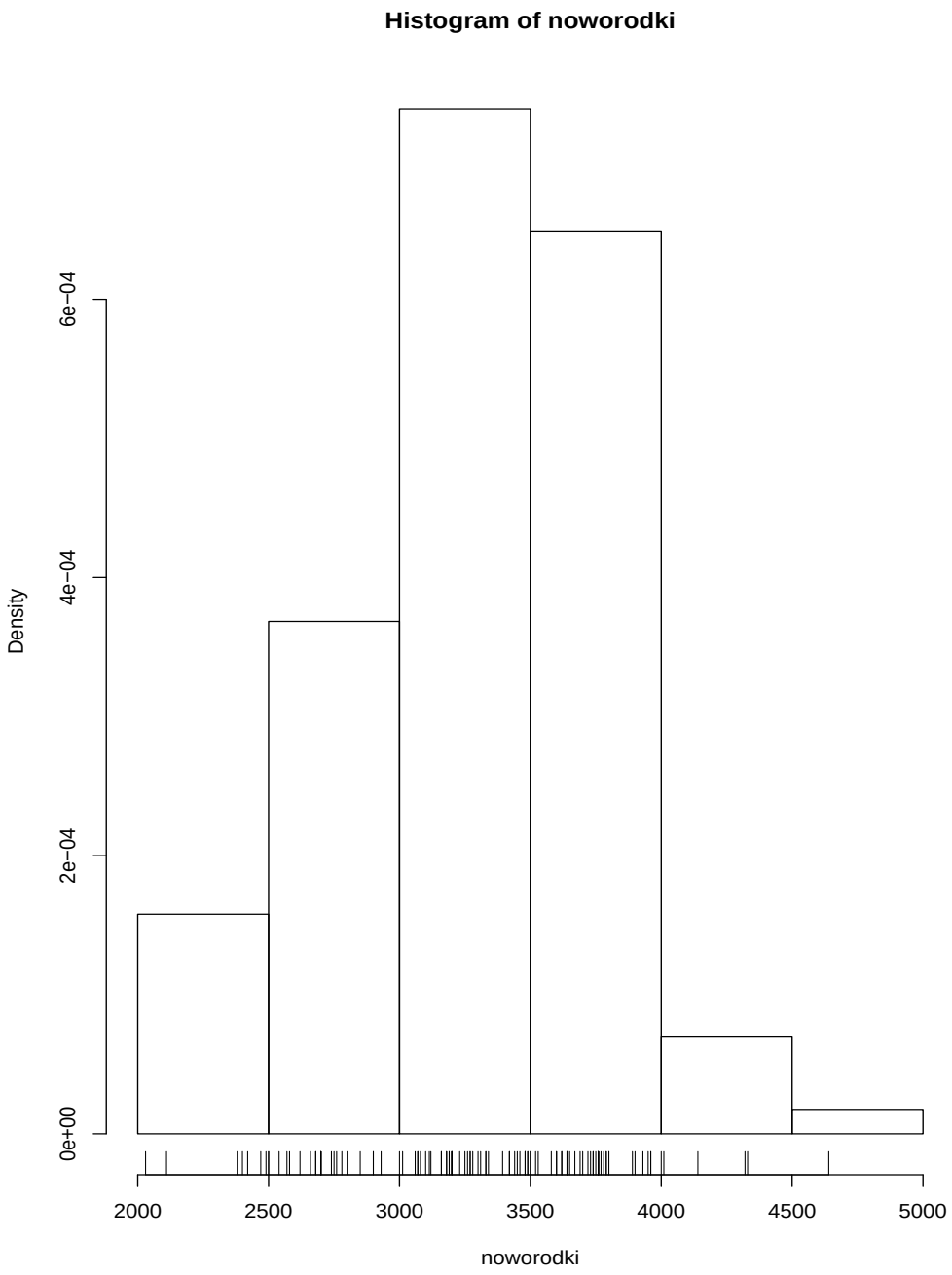
Analogia z histogramem stanie się widoczna, jeśli wybierzemy „prostokątne" jądro $k(x) = \mathbb{I}(-1/2 \leqslant x \leqslant 1/2)$. Alternatywnie możemy interpretować $\hat{f}_n$ jako mieszankę n gęstości, z których każda aproksymuje rozkład jednopunktowy, skupiony w pojedynczym punkcie próbki. Estymator jądrowy jest ciągłą funkcją x, jeśli tylko jądro jest ciągłe. To jest ich zaleta w porównaniu z histogramami. Niemniej, do pewnego stopnia własności tych estymatorów są podobne.

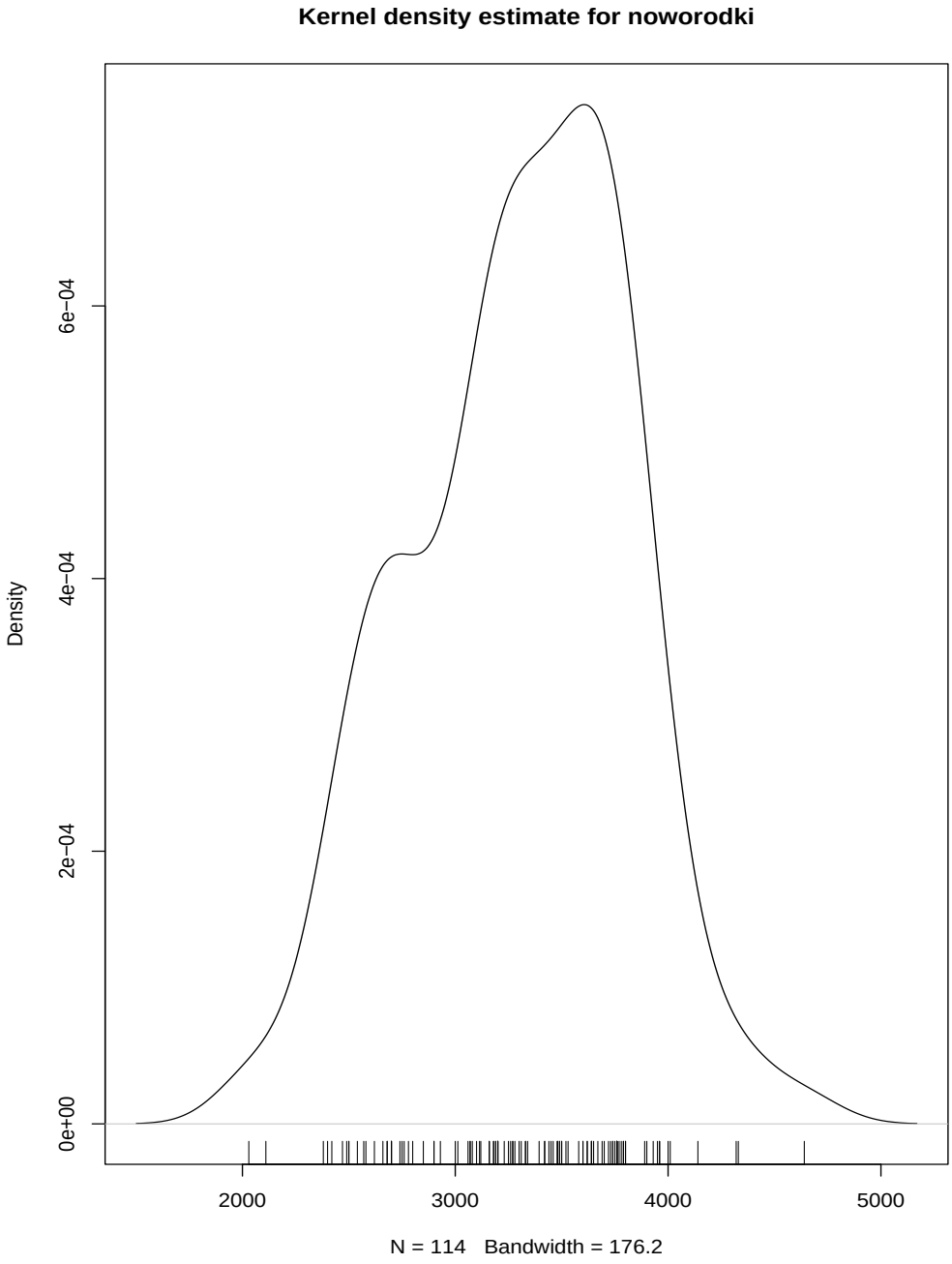
1.3.2 Stwierdzenie. *Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu o gęstości f, o rozmiarze rosnącym do nieskończoności, $n \to \infty$. Rozpatrzmy estymator jądrowy $\hat{f}_n^{\mathrm{kern}}$ zbudowany na podstawie próbki rozmiaru n i mający szerokość pasma h_n. Załóżmy, że $h_n \to 0$ i $nh_n \to \infty$. Dla każdego punktu $x \in \mathbb{R}$ takiego, że f jest ciągła w x mamy*

$$\hat{f}_n^{\mathrm{kern}} \to_{\mathrm{P}} f(x), \quad (n \to \infty).$$

Dowód w szczególnym przypadku jądra prostokątnego jest naszkicowany w Zadaniu 11. Ogólny przypadek nie jest wiele trudniejszy.

Poniżej widac wynik instrukcji `hist(noworodki,prob=TRUE)`. Liczba przedziałów histogramu została ustalona automatycznie, zgodnie z domyślnymi ustawieniami funkcji `hist`. Dla porównania, obejrzyjmy rezultaty estymacji gęstości metodą jądrowa (funkcja `density`). Wybór jądra (`gaussian`, czyli gęstość rozkładu normalnego) oraz szerokości pasma (`bandwidth`) są „domyślne".


Histogram of noworodki
Density
0e+00
2e−04
4e−04
6e−04
2000
2500
3000
3500
4000
4500
5000
noworodki


Kernel density estimate for noworodki
Density
0e+00
2e−04
4e−04
6e−04
2000
3000
4000
5000
N = 114 Bandwidth = 176.2

1.4 Zadania

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

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

3. Pokazać, że zmienna losowa $X_{k:n}$ ma dystrybuantę

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

4. Jeśli zmienne losowe X_i mają gęstość $f(x) = (\mathrm{d}/\mathrm{d}x)F(x)$, to k-ta statystyka pozycyjna ma gęstość

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(X_{k:n} \leqslant x) = n\binom{n-1}{k-1} f(x) F(x)^{k-1} (1 - F(x))^{n-k}.$$

5. Obliczyć $\mathbb{E}U_{k:n}$, gdzie $U_{k:n}$ oznacza statystykę pozycyjną z rozkładu jednostajnego $\mathrm{U}(0,1)$.

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

7. Pokazać, że wektor statystyk pozycyjnych $(U_{1:n}, \ldots, U_{n:n})$ z rozkładu jednostajnego $\mathrm{U}(0,1)$ ma łączny rozkład jednostajny na sympleksie $\{u : 0 \leqslant u_1 \leqslant \cdots \leqslant u_n \leqslant 1\}$.

8. Załóżmy (dla uproszczenia, to nie jest istotne), 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.

9. (Ciąg dalszy). Pokazać, że $X_{k:n} = F^{-1}(U_{k:n})$.

10. Udowodnić następujące stwierdzenie z rachunku prawdopodobieństwa: Jeśli $\mathbb{E}X_n \to a$ i $\mathrm{Var}X_n \to 0$ to $X_n \to_{\mathrm{P}} a$ (a jest liczbą).

11. Rozważmy estymator jądrowy z jądrem „prostokątnym" danym wzorem $k(x) = \mathbb{I}(-1/2 \leqslant x \leqslant 1/2)$. Obliczyć $\mathbb{E}\hat{f}_n^{\mathrm{kern}}(x)$ i $\mathrm{Var}\hat{f}_n^{\mathrm{kern}}(x)$. Zbadać granice wartości oczekiwanych i wariancji przy $n \to \infty$, przy założeniach Stwierdzenia 1.3.2.

    Udowodnić Stwierdzenie 1.3.2 w przypadku prostokątnego jądra.

Rozdział 2

Modele statystyczne

2.1 Przestrzenie statystyczne

Zakładamy, że obserwujemy pewne zmienne losowe $X_1, \ldots, X_n$ o *nieznanym* (łącznym) rozkładzie prawdopodobieństwa. Ciąg $X = (X_1, \ldots, X_n)$ możemy traktować jako pojedynczą wielowymiarową obserwację. W Przykładzie 1.1.8 założyliśmy, że proces kontroli jakości stanowi schemat Bernoulliego. Niech θ będzie prawdopodobieństwem „sukcesu" (tego, że wyrob jest prawidłowy). W statystyce mówi się, że θ jest nieznanym parametrem rozkładu prawdopodobieństwa. Musimy uwzględnić wiele teoretycznie możliwych wartości $\theta \in [0, 1]$. Rozpatrujemy *wiele* różnych schematów Bernoulliego i nie wiemy który z nich opisuje nasze doświadczenie. Model statystyczny precyzuje *rodzinę* wszystkich branych pod uwagę rozkładów prawdopodobieństwa $\{\mathbb{P}_\theta; \theta \in \Theta\}$. Zakładamy, że nieznany rozkład $\mathbb{P}$, który „rządzi" zachowaniem obserwacji X, należy do rozpatrywanej przez nas rodziny. Wiemy więc, że jest to rozkład $\mathbb{P}_{\theta_0}$ dla pewnego $\theta_0 \in \Theta$, tylko nie umiemy wskazać θ_0.

2.1.1 DEFINICJA. ***Model statystyczny*** *określamy przez podanie rodziny* $\{\mathbb{P}_\theta; \theta \in \Theta\}$ *rozkładów prawdopodobieństwa na przestrzeni próbkowej* Ω *oraz zmiennej losowej* $X : \Omega \to \mathcal{X}$*, którą traktujemy jako obserwację. Zbiór* $\mathcal{X}$ *nazywamy przestrzenią obserwacji zaś* Θ *nazywamy przestrzenią parametrów.*

2.1.2 Uwaga (Mierzalność). Zakładamy, że przestrzeń Ω jest wyposażona w σ-ciało $\mathcal{F}$ zdarzeń losowych. Każdy z rozkładów prawdopodobieństwa $\mathbb{P}_\theta$ „żyje" na $\mathcal{F}$. Zmienna losowa X jest odwzorowaniem mierzalnym i przestrzeń $\mathcal{X}$ też musi być wyposażona w σ-ciało. Pojęcia teorii miary traktujemy marginesowo.

2.1.3 Uwaga (Kanoniczna przestrzeń próbkowa). Za przestrzeń Ω wybieramy zwykle zbiór wszystkich możliwych wyników doświadczenia losowego, a więc zbiór wartości wektora złożonego z interesujących nas zmiennych losowych. Jeśli uwzględniamy tylko obserwowane zmienne losowe, to możemy przyjąć, że $\Omega = \mathcal{X}$ i $X(\omega) = \omega$. Przy tej umowie, rozkład prawdopodobieństwa na przestrzeni próbkowej jest po prostu rozkładem prawdopodobieństwa obserwacji: $\mathbb{P}_\theta(B) = \mathbb{P}_\theta(X \in B)$, dla $B \subseteq \mathcal{X}$. Standardowo nazywa się $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta; \theta \in \Theta\}$ *przestrzenią statystyczną.* Jeśli chcemy uwzględnić również *nieobserwowane zmienne losowe* (na przykład dotyczące przyszłości) to nie można już przyjąć, że $\Omega = \mathcal{X}$. Dlatego przyjęliśmy nieco ogólniejszą Definicję 2.1.1.

2.1.4 Uwaga (Ciągłe i dyskretne przestrzenie obserwacji). Ograniczymy rozważania do następujących dwóch typów modeli. Mówimy o modelu ciągłym, jeśli $\mathcal{X}$ jest częścią przestrzeni $\mathbb{R}^n$ wyposażnoną w σ-ciało zbiorów borelowskich i n-wymiarową miarę Lebesgue'a. Model nazywamy dyskretnym, jeśli przestrzeń $\mathcal{X}$ jest skończona lub przeliczalna, wyposażona w σ-ciało 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 $h(X)$ 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.

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

2.1.5 PRZYKŁAD (Statystyczna kontrola jakości, kontynuacja). Powróćmy do Przykładu 1.1.8. Przestrzenią obserwacji jest $\mathcal{X} = \{0, 1\}^n$. Obserwacje $X_1, \ldots, X_n$ są zmiennymi losowymi o łącznym rozkładzie prawdopodobieństwa

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

gdzie $x_i \in \{0, 1\}$ dla $i = 1, \ldots, n$ i $\sum x_i$ oznacza $\sum_{i=1}^n x_i$. Przestrzenią parametrów jest $\Theta = [0, 1]$.

2.1.6 PRZYKŁAD (Badanie reprezentacyjne). Powiedzmy, że populacja składa się 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.

2.1.7 PRZYKŁAD (Model nieparametryczny). Zgodnie z Definicją 1.1.1, ciąg obserwacji $X_1,\ldots,X_n$ 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 $\mathcal{X} = \mathbb{R}^n$.

2.1.8 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 $\mathrm{Var}_\theta X_i = \theta$.

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

2.1.9 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 $\mathrm{Var}_\theta X_i = 1/\theta^2$.

2.1.10 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 ich rozkłady

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

Obie przestrzenie $\mathcal{X}$ i $\mathcal{T}$ muszą być wyposażona w σ-ciała. Zwykle są to borelowskie podzbiory przestrzeni euklidesowych.

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 2.1.5 (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:

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

W skrócie napiszemy $S \sim \mathrm{Bin}(n,\theta)$. Zmienna losowa $(S - n\theta)/\sqrt{n\theta(1-\theta)}$ *nie jest* statystyką. Jej rozkład prawdopodobieństwa jest w przybliżeniu normalny $\mathrm{N}(0,1)$, jeśli n jest duże a $\theta(1-\theta)$ nie jest zbyt małe.

W Przykładzie 2.1.8 (Wypadki) sumaryczna liczba wypadków $S = \sum_{i=1}^{n} X_i$ jest statystyką i ma rozkład $\mathrm{Poiss}(n\theta)$.

W Przykładzie 2.1.9 (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 2.1.10 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.

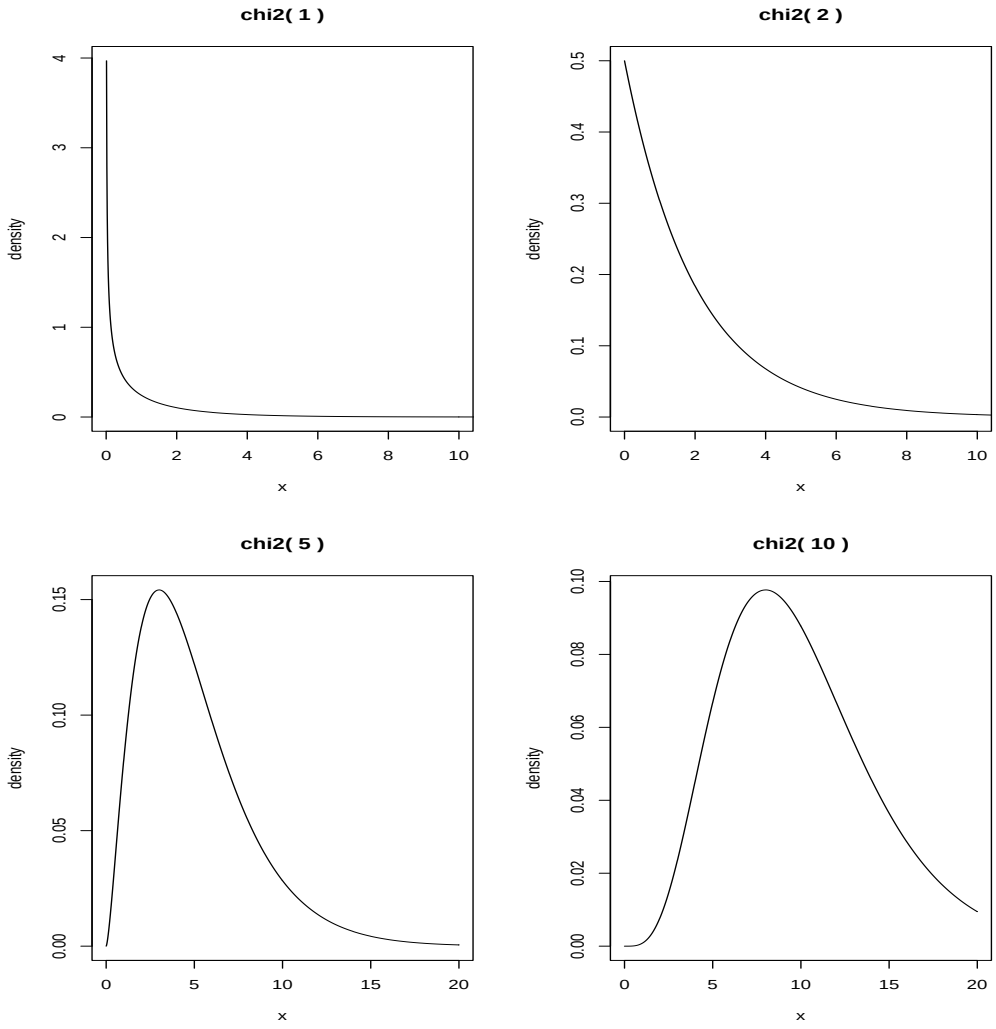
Rozkład chi-kwadrat z k stopniami swobody jest to, z definicji, rozkład zmiennej losowej

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

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

2.2.3 Uwaga. Rozkłady chi-kwadrat są szczególnej postaci rozkładami Gamma, mianowicie $\chi^2(k) = \text{Gamma}(k/2, 1/2)$ (Zadanie 4). Jeśli $Y \sim \chi^2(k)$ to $\mathbb{E}Y = k$ i $\text{Var}Y = 2k$.

Wykresy gęstości kilku rozkładów χ^2 są pokazane poniżej.



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

$$\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n);$$

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1).$$

Pominiemy dowód, bo w Rozdziale 8 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.

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

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, $Z \sim \mathrm{N}(0,1)$ i $Y \sim \chi^2(k)$. Będziemy pisali symbolicznie $T \sim \mathrm{t}(k)$.

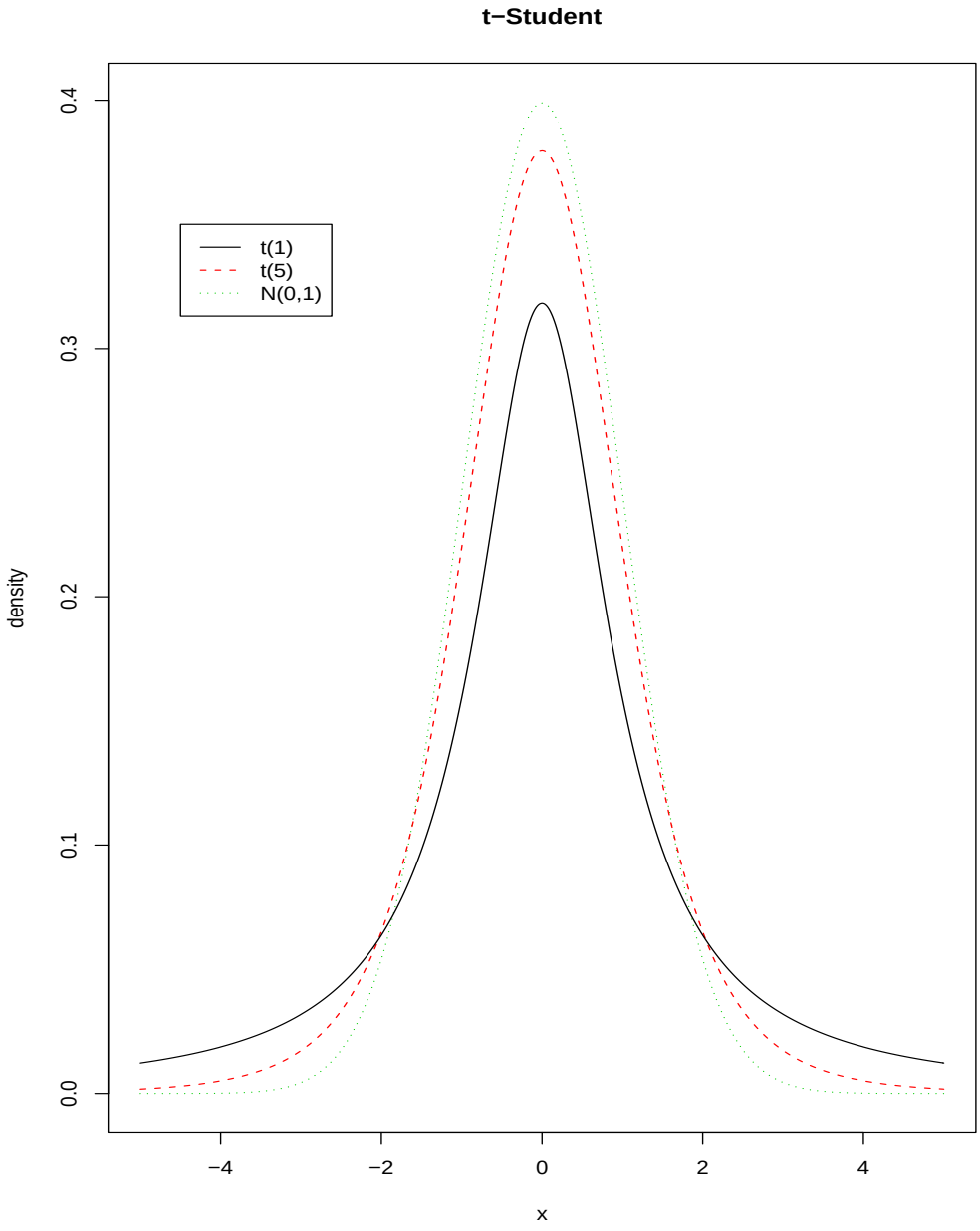
2.2.6 Wniosek. *W modelu normalnym, zmienna losowa $\sqrt{n}(\bar{X} - \mu)/S$ ma rozkład* $\mathrm{t}(n-1)$.

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 $R \sim \mathrm{F}(k, m)$.

Dwa rozkłady t oraz rozkład normalny są pokazane na poniższym rysunku.



2.2.7 PRZYKŁAD (Model dwóch próbek). Załóżmy, że obserwujemy niezależne zmienne losowe $X_1, \ldots, X_n$ i $Y_1, \ldots, Y_m$, przy tym $X_i \sim \mathrm{N}(\mu_X, \sigma_X^2)$ i $Y_j \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ 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

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

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 $\mathrm{F}(n-1, m-1)$.

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

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

2.3 Dostateczność

Rozważmy model statystyczny. Załóżmy, że rozkłady prawdopodobieństwa należące do rodziny $\{\mathbb{P}_\theta : \theta \in \Theta\}$ mają gęstości f_θ na przestrzeni obserwacji $\mathcal{X}$. Są to, jak zwykle, albo gęstości względem miary Lebesgue'a, albo „gęstości dyskretne" $f_\theta(x) = \mathbb{P}_\theta(X = x)$.

2.3.1 DEFINICJA (Własność faktoryzacji). *Statystykę $T = T(X)$ nazywamy* ***dostateczną****, jeśli gęstości obserwacji można przedstawić w postaci*

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

Sens tej definicji wyjaśni w pewnym stopniu następujące stwierdzenie.

2.3.2 Stwierdzenie. *Statystyka $T = T(X)$ jest dostateczna wtedy i tylko wtedy gdy warunkowy rozkład prawdopodobieństwa obserwacji X przy danej wartości statystyki $T = t$ nie zależy od parametru θ.*

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

$$\mathbb{P}_\theta(X \in B | T(X) = t) \qquad (*)$$

nie zależy od θ, dla dowolnego (borelowskiego) zbioru $B \subseteq \mathcal{X}$ i (prawie) każdego t. Ścisłe sformułowanie Stwierdzenia 2.3.2 wymaga znajomości ogólnego pojęcia warunkowego rozkładu prawdopodobieństwa. Zwróćmy uwagę, że elementarne podejście poprzez gęstości warunkowe tutaj się bezpośrednio nie stosuje, bo rozkład X przy danym $T(X) = t$ jest zazwyczaj skupiony na „podprzestrzeni o niższym wymiarze", patrz Zadanie 12.

Jeśli $\mathcal{X}$ jest przestrzenią dyskretną, to możemy się posłużyć elementarną definicją prawdopodobieństwa warunkowego. Warunek (*) redukuje się do tego, że

$$\mathbb{P}_\theta(X = x | T(X) = t) \qquad (**)$$

nie zależy od θ dla dowolnego x.

Dowód. Żeby uniknąć trudności technicznych, udowodnimy Stwierdzenie 2.3.2 tylko w 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 możemy przyjąć $h(x) = \mathbb{P}_\theta(X = x | T(X) = t)$ i $g_\theta(t) = \sum_{x':T(x')=t} f_\theta(x')$. □

2.3.4 Uwaga. Warunek sformułowany w Stwierdzeniu 2.3.2 jest zazwyczaj przyjmowany za *definicję* statystyki dostatecznej. Sens tego warunku wyjaśni następujące „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.

2.3.5 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

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

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.

2.3.6 PRZYKŁAD (Statystyki dostateczne w poprzednich przykładach). W Przykładzie 2.1.5 (Schemat Bernoulliego), liczba sukcesów $S = \sum_{i=1}^{n} X_i$ jest statystyką dostateczną.

W Przykładzie 2.1.8 (model Poissona) suma obserwacji $S = \sum_{i=1}^{n} X_i$ jest statystyką dostateczną.

W Przykładzie 2.1.9 (model wykładniczy) średnia $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$ jest statystyką dostateczną.

W Przykładzie 2.1.10 (model normalny z nieznanymi μ i σ) $(\bar{X}, S^2)$ jest dwuwymiarową statystyką dostateczną.

2.3.7 Uwaga. To co powiedzieliśmy w Uwadze 2.3.4 jest słuszne pod warunkiem, że jesteśmy pewni poprawności modelu. W Przykładzie 2.1.5 (statystyczna kontrola jakości), jeśli proces obserwowania sztuk dobrych/wadliwych jest rzeczywiście schematem Bernoulliego, to wystarczy zapisać sobie liczbę sztuk prawidłowych w próbce i zapomnieć o ich kolejności pojawiania się. Ale wyobraźmy sobie, że w trakcie obserwacji akurat maszyna się zepsuła i prawdopodobieństwo wyrobu wadliwego zmieniło się w pewnym momencie. Obserwacje mogą wyglądać tak:

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

Kolejność zer i jedynek jest przydatna do weryfikacji adekwatności modelu Bernoulliego.

2.4 Zadania

1. Rozpatrzmy proces statystycznej kontroli jakości przyjmując te same założenia co w Przykładzie 2.1.5 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. Obliczyć rozkład prawdopodobieństwa zmiennej losowej Z^2, jeśli $Z \sim \mathrm{N}(0,1)$ (obliczyć bezpośrednio dystrybuantę i gęstość rozkładu $\chi^2(1)$).

3. 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 $\chi^2(2)$).

4. Korzystając z Zadania 2 oraz z własności rozkładów gamma, udowodnić Uwagę 2.2.3: gęstość zmiennej losowej $Y \sim \chi^2(k)$ ma postać

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

5. Pokazać, że gęstość zmiennej losowej $T \sim \mathrm{t}(k)$ jest postaci

$$f_T(t) = \frac{\Gamma(k/2+1/2)}{\Gamma(k/2)} \frac{1}{\sqrt{\pi k}} \left(1 + \frac{t^2}{k}\right)^{-(k+1)/2}, \quad (-\infty < t < \infty).$$

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

7. Pokazać, że gęstość zmiennej losowej $R \sim \mathrm{F}(k,m)$ jest postaci

$$f_R(r) = \frac{k^{k/2} m^{m/2} \Gamma(k/2+1/2)}{\Gamma(k/2)\Gamma(m/2)} \frac{r^{k/2-1}}{(kr+m)^{(k+m)/2}}, \quad (r > 0).$$

8. Udowodnić wzór dotyczący rozkładu t-Studenta na końcu Przykładu 2.2.7.

9. Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu Gamma(α, λ). Znaleźć, dwuwymiarową statystykę dostateczną.

10. 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.

11. 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.

12. Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu wykładniczego Ex(θ). Obliczyć łączny rozkład zmiennych losowych $R_1, \ldots, R_{n-1}, S$, gdzie $S = \sum_{i=1}^{n} X_i$, zaś $R_i = X_i/S$. Znaleźć rozkład warunkowy $R_1, \ldots, R_{n-1}$ przy danym $S = s$. Zinterpretować fakt, że statystyka S jest dostateczna.

Część II

Estymacja

Rozdział 3

Estymacja punktowa

W statystyce matematycznej zakładamy, że rozkład prawdopodobieństwa opisujący doświadczenie należy do rodziny $\{\mathbb{P}_\theta : \theta \in \Theta\}$, ale nie znamy parametru θ. **Estymatorem** parametru θ nazywamy dowolną statystykę $T = T(X)$ o wartościach w zbiorze Θ. Interpretujemy T jako przybliżenie θ. Zwykle oznaczamy estymator tą samą literką, co wielkość estymowaną, dodając „daszek": $T = \hat{\theta}$. Zaczniemy od przeglądu prostych metod konstrukcji estymatorów.

3.1 Metody heurystyczne

Podstawianie częstości jest sposobem estymacji, który natychmiast się narzuca i jest zrozumiały dla każdego.

3.1.1 PRZYKŁAD (Prawdopodobieństwo napotkania wadliwego wyrobu). Rozpatrzmy jeszcze raz model statystycznej kontroli jakości z Przykładów 1.1.8 i 2.1.5. Przypomnijmy, że parametr θ oznacza prawdopodobieństwo pojawienia się sztuki prawidłowej. Oczywistym estymatorem θ jest

$$\hat{\theta} = \bar{X} = S/n,$$

gdzie $S = \sum X_i$, czyli frakcja prawidłowych sztuk w próbce losowej. Zastąpiliśmy nieznane prawdopodobieństwo przez *prawdopodobieństwo empiryczne*. Na tym właśnie polega metoda podstawiania częstości.

Metoda podstawiania częstości dopuszcza dużą dowolność. Pokazuje to następny przykład.

3.1.2 PRZYKŁAD (Model Hardy'ego – Weinberga). W populacji mamy osobników o trzech genotypach: 1, 2 i 3. Z rozważań genetycznych wynika, że te trzy typy powinny występować w proporcji $\theta^2 : 2\theta(1-\theta) : (1-\theta)^2$. Wybieramy losowo n osobników, wśród których liczby poszczególnych genotypów są, odpowiednio, N_1, N_2 i N_3. Mamy do czynienia z rozkładem wielomianowym: $(N_1, N_2, N_3) \sim \mathrm{Mult}(n, p_1, p_2, p_3)$:

$$\mathbb{P}_\theta(N_1 = n_1, N_2 = n_2, N_3 = n_3) = \frac{n!}{n_1! n_2! n_3!} p_1^{n_1} p_2^{n_2} p_3^{n_3}.$$

Prawdopodobieństwa p_i nie są tu dowolnymi liczbami dodatnimi sumującymi się do jedynki. Wiemy, że

$$p_1 = \theta^2, \qquad p_2 = 2\theta(1-\theta), \qquad p_3 = (1-\theta)^2.$$

Zadanie polega na estymacji θ. Nasuwają się dwa rozwiązania. Z jednej strony mamy $\theta = \sqrt{p_1}$ i za estymator możemy przyjąć

$$\hat{\theta} = \sqrt{\hat{p}_1} = \sqrt{\frac{N_1}{n}}.$$

Z drugiej strony, $\theta = 1 - \sqrt{p_3}$ i estymator

$$\tilde{\theta} = 1 - \sqrt{\hat{p}_3} = 1 - \sqrt{\frac{N_3}{n}}$$

wydaje się równie rozsądny. Nie wiadomo, który estymator wybrać.

Zagadka. Może istnieje *trzeci* estymator, lepszy od dwóch podanych wyżej? Trzeba jeszcze sprecyzować, co to znaczy „lepszy".

Metoda momentów jest równie prosta. Przyrównujemy momenty rozkładu teoretycznego, zależące od nieznanego parametru, do ich odpowiedników empirycznych. Z powstałych w ten sposób równań wyliczamy parametr. Najczęściej używamy momentów centralnych i równania wyglądają tak:

$$\mu(\theta) = \hat{\mu}; \qquad m_k(\theta) = \hat{m}_k,$$

gdzie $\mu(\theta) = \int x f_\theta(x) \mathrm{d}x$, $m_k(\theta) = \int (x - \mu(\theta))^k f_\theta(x) \mathrm{d}x$, $\hat{\mu} = \bar{X}$ i $\hat{m}_k = (1/n)\sum (X_i - \bar{X})^k$. Układamy tyle równań, ile jest niewiadomych parametrów (współrzędnych wektora θ).

3.1.3 PRZYKŁAD (Rozkład wykładniczy). Niech $X_1, \ldots, X_n$ będzie próbą z rozkładu $\mathrm{Ex}(\theta)$. Wiemy, że $\mu(\theta) = 1/\theta$. Mamy jeden nieznany parametr, więc wystarczy jedno równanie: $1/\theta = \bar{X}$. Estymator otrzymany metodą momentów jest równy $\hat{\theta} = 1/\bar{X}$.

3.1.4 PRZYKŁAD (Rozkład gamma). Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu $\mathrm{Gamma}(\alpha, \lambda)$. Mamy dwa nieznane parametry, więc wykorzystamy dwa momenty: wartość oczekiwaną $\mu(\alpha, \lambda) = \alpha/\lambda$ i wariancję $\sigma^2(\alpha, \lambda) = \alpha/\lambda^2$. Dostajemy układ równań

$$\frac{\alpha}{\lambda} = \bar{X}, \qquad \frac{\alpha}{\lambda^2} = \tilde{S}^2,$$

gdzie $\tilde{S}^2$ jest wariancją z próbki. Rozwiązanie jest następujące:

$$\hat{\lambda} = \frac{\bar{X}}{\tilde{S}^2}, \qquad \hat{\alpha} = \frac{\bar{X}^2}{\tilde{S}^2}.$$

Metoda kwantyli. Jeśli momenty są trudne do obliczenia lub prowadzą do zbyt zawiłych równań, zamiast momentów można użyć kwantyli. Przyrównujemy kwantyle teoretyczne do empirycznych. Otrzymujemy równania postaci $\xi_q(\theta) = \hat{\xi}_q$ lub, równoważnie, $F_\theta(\hat{\xi}_q) = q$. Wybieramy tyle różnych q, ile mamy niewiadomych i rozwiązujemy układ równań względem współrzędnych wektora θ.

3.1.5 PRZYKŁAD (Rozkład Weibulla). Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu o dystrybuancie

$$F_\theta(x) = 1 - \exp\left[-cx^b\right], \quad (x > 0),$$

gdzie $c > 0$ i $b > 0$ są nieznanymi parametrami. To się nazywa rozkład *Weibulla*. Momenty są trudne do obliczenia. Ponieważ mamy dwa nieznane parametry, ułożymy równania dla dwóch kwantyli. Zdecydujmy się na *kwartyle*, to znaczy kwantyle rzędu 1/4 i 3/4. Wybór jest, oczywiście, całkiem arbitralny. Estymatory parametrów c i b wyznaczamy rozwiązując układ równań

$$\begin{cases} 1 - \exp\left[-c \cdot \hat{\xi}_{1/4}^b\right] = 0.25; \\ 1 - \exp\left[-c \cdot \hat{\xi}_{3/4}^b\right] = 0.75. \end{cases}$$

Proste przekształcenia dają $-c \cdot \hat{\xi}_{1/4}^{\hat{b}} = \log 0.75$ i $-c \cdot \hat{\xi}_{3/4}^{\hat{b}} = \log 0.25$. Ostatecznie,

$$\hat{b} = \log 3/(\log \hat{\xi}_{3/4} - \log \hat{\xi}_{1/4}), \qquad \hat{c} = \exp(\log 4 - \hat{b} \log \hat{\xi}_{1/4}).$$

Przejdziemy do opisu metody, która jest zazwyczaj lepsza niż dotychczas przedstawione.

3.2 Wiarogodność

Idea jest taka: wybieramy taki parametr θ, dla którego otrzymane wyniki doświadczenia losowego są „najbardziej prawdopodobne". Sformułujemy to dokładniej. Niech $f_\theta(x)$ będzie łączną gęstością obserwacji.

3.2.1 DEFINICJA. ***Wiarogodność*** *jest to funkcja* $\mathcal{L} : \Theta \to \mathbb{R}$ *dana wzorem*

$$\mathcal{L}(\theta) = f_\theta(x).$$

Wiarogodność jest właściwie tym samym, co gęstość prawdopodobieństwa, ale rozważana *jako funkcja parametru* θ, przy ustalonych wartościach obserwacji $x = X(\omega)$.

3.2.2 DEFINICJA. *Mówimy, że* $\hat{\theta} = \hat{\theta}(X)$ *jest* ***estymatorem największej wiarogodności*** *parametru* θ*, jeśli*

$$f_{\hat{\theta}(x)}(x) = \sup_{\theta \in \Theta} f_\theta(x),$$

dla dowolnego x*. Symbolicznie będziemy pisać* $\hat{\theta} = ENW(\theta)$*.*

W skrócie, definicję ENW(θ) zapiszemy w postaci

$$\mathcal{L}(\hat{\theta}) = \sup_{\theta \in \Theta} \mathcal{L}(\theta),$$

musimy jednak pamiętać, że zarówno $\mathcal{L}$, jak i $\hat{\theta}$ zależą od obserwacji. Pominiemy dyskusję na temat istnienia i jednoznaczności ENW.

3.2.3 PRZYKŁAD (Liczenie ryb w jeziorze). W jeziorze pływa pewna liczba, powiedzmy r, ryb. Żeby oszacować nieznaną liczbę r bez osuszania jeziora, musimy przywołać na pomoc statystykę matematyczną. Najpierw odławiamy m ryb, *znaczymy* je i wpuszczamy z powotem do jeziora. Czekamy, aż znaczone ryby „wymieszają się" z pozostałymi. Następnie wyławiamy n ryb. Stwierdzamy, że jest wśród nich k znaczonych (złapanych powtórnie). W tym doświadczeniu, r jest nieznanym parametrem, m i n są znanymi liczbami, zaś k jest znaną wartością *zmiennej losowej* K, o rozkładzie hipergeometrycznym

$$\mathcal{L}(r) = \mathbb{P}_r(K = k) = \binom{m}{k}\binom{r-m}{n-k} \Big/ \binom{r}{n}.$$

Wyznaczymy ENW(r). W tym celu trzeba znaleźć maksimum

$$\mathcal{L}(r) = \max_r .$$

Przyjmijmy wygodne oznaczenie $(n)_k = n(n-1)\cdots(n-k+1)$. Ponieważ

$$\frac{\mathcal{L}(r+1)}{\mathcal{L}(r)} = \frac{(r+1-m)_{n-k}}{(r+1)_n} \cdot \frac{(r)_n}{(r-m)_{n-k}} = \frac{r+1-m}{r-m-n+k+1} \cdot \frac{r-n+1}{r+1},$$

więc $L(r+1) > L(r)$ wtedy i tylko wtedy gdy $r < mn/k$. Stąd widać, że $\mathcal{L}(r)$ osiąga maksymalną wartość dla najmniejszej liczby całkowitej przekraczającej mn/k. Innymi słowy,

$$\hat{r} = \mathrm{ENW}(r) = \left[\frac{mn}{k}\right] + 1.$$

Wynik jest zgodny ze zdrowym rozsądkiem.

Prosty chwyt często ułatwia wyznaczenie ENW. Funkcja $\mathcal{L}(\theta)$ osiąga maksimum w tym samym punkcie, co jej *logarytm*. Niech

$$\ell(\theta) = \log \mathcal{L}(\theta).$$

3.2.4 PRZYKŁAD (Rozkład wykładniczy). Jeśli $X_1, \ldots, X_n$ jest próbką z rozkładu $\mathrm{Ex}(\theta)$, to wiarogodność jest dana wzorem

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \left(\theta \mathrm{e}^{-\theta x_i}\right),$$

więc

$$\ell(\theta) = n \log \theta - \theta \sum x_i.$$

Wystarczy teraz przyrównać pochodną do zera:

$$\ell'(\theta) = \frac{n}{\theta} - \sum x_i = 0.$$

Łatwo się przekonać, że $\ell(\theta)$ osiąga maksimum w punkcie

$$\hat{\theta} = \frac{n}{\sum x_i} = \frac{1}{\bar{x}}.$$

Zatem $ENW(\theta) = 1/\bar{X}$. Metoda największej wiarogodności doprowadziła do tego samego estymatora, co metoda momentów.

3.2.5 PRZYKŁAD (Rozkład Poissona). Dla próbki z rozkładu $\mathrm{Poiss}(\theta)$ mamy

$$\ell(\theta) = -\theta n + \log(\theta) \sum x_i - \sum \log(x_i!),$$

więc $\ell'(\theta) = -n + \sum x_i/\theta = 0$ dla $\theta = \sum x_i/n$. Otrzymujemy znów dobrze znany estymator: $ENW(\theta) = \bar{X}$.

Sposób obliczenia ENW w Przykładach 3.2.4 i 3.2.5 jest bardzo typowy. Jeśli mamy *próbkę z rozkładu o gęstości* f_θ i zbiór parametrów $\Theta \subseteq \mathbb{R}$ jest przedziałem, to

$$\mathcal{L}(\theta) = f_\theta(x_1, \ldots, x_n) = f_\theta(x_1) \cdots f_\theta(x_n),$$

$$\ell(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i).$$

ENW wyznaczamy zazwyczaj rozwiązując równanie

$$\ell'(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f_\theta(x_i) = 0.$$

Przypadek *wielu* nieznanych parametrów wymaga tylko oczywistej modyfikacji. Jeśli mamy zbudować estymator wektora $\theta = (\theta_1, \ldots, \theta_k) \in \Theta \subseteq \mathbb{R}^k$, to obliczamy pochodne *cząstkowe* logarytmu wiarogodności. Rozwiązujemy układ k równań z k niewiadomymi:

$$\frac{\partial}{\partial \theta_j} \ell(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} \log f_\theta(x_i) = 0, \qquad (j = 1, \ldots, k).$$

3.2.6 PRZYKŁAD (Rozkład normalny). Rozważmy próbkę $X_1, \ldots, X_n$ z rozkładu $\mathrm{N}(\mu, \sigma^2)$, z nieznanymi parametrami μ i $\sigma > 0$. Mamy

$$\log f_{\mu,\sigma}(x) = -\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}(x-\mu)^2.$$

Logarytm wiarogodności dla całej próbki jest równy

$$\ell(\mu, \sigma) = \sum_{i=1}^{n} \log f_{\mu,\sigma}(x_i) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2,$$

czyli

$$\ell(\mu, \sigma) = \mathrm{const} - n\log\sigma - \frac{1}{2\sigma^2}\left(\sum x_i^2 - 2\mu\sum x_i + n\mu^2\right).$$

Układ równań przybiera postać

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\left(\sum x_i^2 - 2\mu\sum x_i + n\mu^2\right) = 0;$$

$$\frac{\partial l}{\partial \mu} = -\frac{1}{\sigma^2}\sum x_i - \frac{n\mu}{\sigma^2} = 0.$$

Z drugiego równania łatwo wyliczamy $\mu = \sum x_i/n$. Podstawiając do pierwszego równania otrzymujemy ENW:

$$\hat{\mu} = \bar{x};$$

$$\hat{\sigma}^2 = \frac{1}{n}\left(\sum x_i^2 - \bar{x}^2\right) = \frac{1}{n}\sum (x_i - \bar{x})^2.$$

Dobrze znane estymatory $\bar{X}$ i $\tilde{S}^2$ okazują się być $\mathrm{ENW}(\mu)$ i $\mathrm{ENW}(\sigma^2)$, odpowiednio.

3.2.7 PRZYKŁAD (Rozkład Laplace'a). Mówimy, że zmienna losowa X ma rozkład **Laplace'a** z parametrami μ i λ, jeśli ma gęstość daną wzorem

$$f_{\mu,\lambda}(x) = \frac{\lambda}{2}\mathrm{e}^{-\lambda|x-\mu|}.$$

Inaczej, taki rozkład nazywa się rozkładem *podwójnie wykładniczym*. Symbolicznie, piszemy $X \sim \mathrm{Lapl}(\mu, \lambda)$. Interesujący jest $\mathrm{ENW}(\mu)$ dla próbki z rozkładu Laplace'a. Mamy

$$\ell(\mu,\lambda) = \sum_{i=1}^{n} \log f_{\mu,\lambda}(x_i) = -n\log 2 + n\log\lambda - \frac{\lambda}{2}\sum_{i=1}^{n}|x_i-\mu|.$$

Popatrzmy na $\ell(\mu,\lambda)$ jak na funkcję μ, przy ustalonym λ. Funkcja ℓ jest *kawałkami liniowa* i wklęsła. Jej wykres jest łamaną, której wierzchołki mają współrzędne $\mu = x_i$. Pochodna (poza punktami nieróżniczkowalności) jest równa $\partial\ell/\partial\mu = (\lambda/2)\sum \mathrm{sign}(\mu - x_i)$. Widać, że funkcja jest rosnąca na lewo od *mediany próbkowej* i malejąca na prawo od niej. Zatem

$$\mathrm{ENW}(\mu) = \hat{\mathrm{med}}.$$

Dla parzystych n mediana próbkowa, a więc i $\mathrm{ENW}(\mu)$, nie są wyznaczone jednoznacznie. Najlepiej naszkicować wykres $\ell(\mu)$ dla, powiedzmy, $n = 4$ i $n = 5$. Wtedy powyższe rozważania staną się jasne.

Po wyznaczeniu $\mathrm{ENW}(\mu)$, nietrudno znaleźć $\mathrm{ENW}(\lambda)$. Przyrównujemy do zera $\partial\ell/\partial\lambda$ i natychmiast dostajemy

$$\mathrm{ENW}(\lambda) = \frac{2n}{\sum|\hat{\mathrm{med}} - x_i|}.$$

3.2.8 PRZYKŁAD (Rozkład jednostajny). Jeśli $X_1, \ldots, X_n$ jest próbką z rozkładu $\mathrm{U}(0,\theta)$, z nieznanym parametrem $\theta > 0$, to

$$l(\theta) = \prod_i \left(\frac{1}{\theta}\mathbb{I}(x_i \geqslant \theta)\right) = \begin{cases} 1/\theta^n & \text{dla } \theta \geqslant \max(x_1,\ldots,x_n); \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

Wystarczy popatrzeć chwilę na ten wzór, żeby stwierdzić, że

$$\mathrm{ENW}(\theta) = \max(X_1, \ldots, X_n).$$

Metoda największej wiarogodności jest bardzo giętka i daje się zastosować w różnych „nietypowych" sytuacjach.

3.2.9 PRZYKŁAD (Dane cenzurowane). Przypuśćmy, że interesuje nas rozkład zmiennych losowych $X_1, \ldots, X_n$ ale obserwujemy tylko $Y_1, \ldots, Y_n$, gdzie

$$Y_i = \min(X_i, a),$$

dla pewnej znanej liczby a. Jest to najprostszy model „cenzurowania". Można sobie wyobrazić, że X_i to „czasy życia" elementów zaś a jest „horyzontem obserwacji". Dla ustalenia uwagi załóżmy, że $X_1, \ldots, X_n$ jest próbką z rozkładu $\mathrm{Ex}(\theta)$. Pojedyncza zmienna losowa Y_i ma rozkład, który nie jest ani absolutnie ciągły ani dyskretny. Niemniej, ten rozkład ma gęstość względem sumy miary Lebesgue'a i miary o masie 1 skupionej w a:

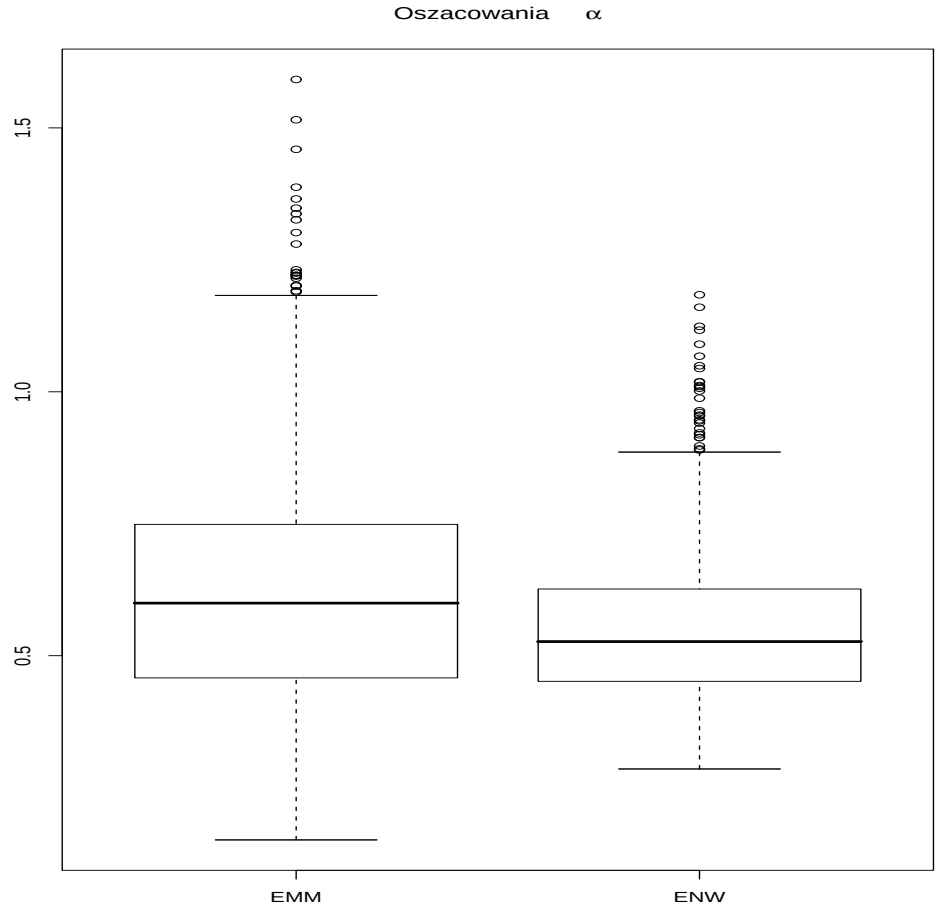
$$f_\theta(y) = \begin{cases} \theta \mathrm{e}^{-\theta y} & \text{dla } y < a, \\ \mathrm{e}^{-\theta y} & \text{dla } y = a. \end{cases}$$

Istotnie, w punkcie $y < a$ gęstość zmiennej Y_i jest równa gęstości zmiennej X_i. Dla $y = a$ mamy $f_\theta(a) = \mathbb{P}_\theta(Y_i = a) = \mathbb{P}_\theta(X_i \geq a) = \mathrm{e}^{-\theta a}$. Wiarogodność jest dana wzorem

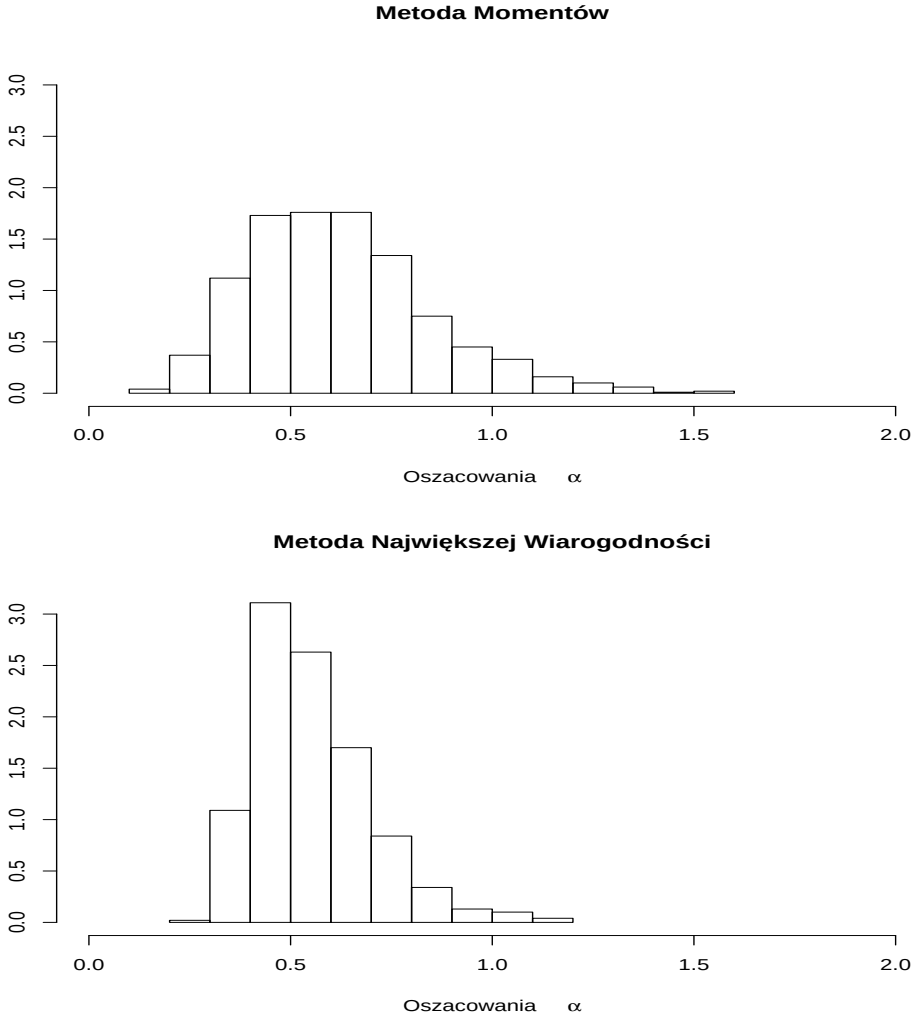
$$\mathcal{L}(\theta) = \theta^m \exp\left(-\theta \sum_{i=1}^{n} y_i\right),$$

gdzie $s = \sum_{i=1}^{n} y_i$ jest sumą wszystkich obserwacji zaś $m = \sum \mathbb{I}(y_i < a)$ oznacza liczbę obserwacji „nieocenzurowanych". Stąd otrzymujemy wzór na ENW: $\hat{\theta} = m/s$.

3.2.10 PRZYKŁAD. Rozważmy próbkę $X_1, \ldots, X_n$ z rozkładu $\mathrm{Gamma}(\alpha, \lambda)$. Estymatora największej wiarogodności parametru α nie da się wyrazić prostym wzorem ale można łatwo obliczyć numerycznie w R przy użyciu funkcji `nlm`. Porównajmy ENW z estymatorem otrzymanym metodą momentów w Przykładzie 3.1.4. Próbki rozmiaru n=25 były generowane z rozkładu $\mathrm{Gamma}(\alpha, \lambda)$ z $\alpha = 0.5$. Estymowane były oba parametry, ale przedstawimy tylko wyniki dla α. Doświadczenie było powtórzone $m = 1000$ razy. Poniższy rysunek przedstawia wykresy pudełkowe otrzymanych oszacowań parametru α.



Następny rysunek przedstawia histogramy otrzymanych oszacowań.



Widać z rysunków, że estymatory ENW są „średnio bliższe" estymowanej wielkości $\alpha = 0.5$[1]. Pojęcia wprowadzone w następnym podrozdziale uściślają i kwantyfikują pojęcie „średniego błędu" estymacji.

[1] W doświadczeniach symulacyjnych „udajemy", że nie znamy α i obliczamy estymatory, a później porównujemy z prawdziwą wartością. To typowa metodologia.

3.3 Błąd średniokwadratowy

Zajmiemy się bardziej systematycznie teorią estymacji. Zasadnicze pytania, na które chcemy odpowiedzieć dotyczą dokładności estymatorów. Jaki błąd popełniamy szacując nieznany parametr? Co to znaczy, że estymator jest „dobry"? Który z konkurujących ze sobą estymatorów uznać za „lepszy"? Czy można znaleźć estymator „najlepszy"?

Tak jak zwykle, rozważamy model statystyczny, a więc losową obserwację X i rodzinę $\{\mathbb{P}_\theta; \theta \in \Theta\}$ rozkładów prawdopodobieństwa. Zadanie estymacji sformułujemy dokładniej niż w poprzednim podrozdziale. Przypuśćmy, że chcemy estymować parametr θ lub liczbę $g(\theta)$, gdzie $g : \Theta \to \mathbb{R}$ jest znaną funkcją (ale argument θ jest nieznany)[2]. Estymatorem $g(\theta)$ może być dowolna *statystyka* $\hat{g} : \mathcal{X} \to \mathbb{R}$. Staramy się znaleźć taki estymator, dla którego mały jest „błąd przybliżenia". Chcemy, żeby różnica

$$\hat{g}(X) - g(\theta),$$

miała możliwie małą wartość bezwzględną dla każdego θ. Są dwie trudności. Po pierwsze, błąd jest zmienną losową. Po drugie, zależy od parametru θ. Jest sposób na ominięcie pierwszej trudności. Skoro wielkość błędu zależy od przypadku, to możemy żądać, żeby błąd był „średnio" możliwie mały. Trudność druga pozostaje: uśredniony błąd zależy od parametru θ, który jest przecież nieznany.

3.3.1 DEFINICJA. *Niech $\hat{g}(X)$ będzie estymatorem $g(\theta)$. Funkcję*

$$R(\theta) = \mathbb{E}_\theta(\hat{g}(X) - g(\theta))^2, \quad (\theta \in \Theta),$$

nazywamy ***błędem średniokwadratowym*** *(BŚK) tego estymatora.*

3.3.2 Uwaga. BŚK ma angielski skrót MSE (*mean square error*) i nazywa się także *funkcją ryzyka przy kwadratowej funkcji straty.* Ogólniejszą definicję ryzyka i inne funkcje straty spotkamy później. Na razie rozważamy tylko BŚK.

[2] Na przykład zależy nam na dobrym przybliżeniu θ^2 albo e^θ, albo jednej ze współrzędnych w sytuacji, gdy θ jest wektorem.

3.3.3 DEFINICJA. *Jeśli statystyka $\hat{g}(X)$ jest estymatorem $g(\theta)$ to*

$$b(\theta) = \mathbb{E}_\theta(\hat{g}(X) - g(\theta)), \quad (\theta \in \Theta),$$

nazywamy ***obciążeniem*** *tego estymatora. Mówimy, że estymator jest* ***nieobciążony****, jeśli $b(\theta) \equiv 0$, to znaczy dla każdego $\theta \in \Theta$,*

$$\mathbb{E}_\theta \hat{g}(X) = g(\theta).$$

3.3.4 Stwierdzenie. *Jeśli $R(\theta)$ jest BŚK estymatora $\hat{g}$ i $b(\theta)$ jest jego obciążeniem to*

$$R(\theta) = \mathrm{Var}_\theta \hat{g}(X) + b(\theta)^2.$$

Dowód. BŚK jest, z definicji, równy

$$\begin{aligned}
\mathbb{E}_\theta(\hat{g}(X) - g(\theta))^2 &= \mathbb{E}_\theta(\hat{g}(X) - \mathbb{E}_\theta\hat{g}(X) + \mathbb{E}_\theta\hat{g}(X) - g(\theta))^2 \\
&= \mathbb{E}_\theta(\hat{g}(X) - \mathbb{E}_\theta\hat{g}(X))^2 + \mathbb{E}_\theta(\mathbb{E}_\theta\hat{g}(X) - g(\theta))^2 \\
&\qquad + 2\mathbb{E}_\theta(\hat{g}(X) - \mathbb{E}_\theta\hat{g}(X))(\mathbb{E}_\theta\hat{g}(X) - g(\theta)) \\
&= \mathbb{E}_\theta(\hat{g}(X) - \mathbb{E}_\theta\hat{g}(X))^2 + \mathbb{E}_\theta(\mathbb{E}_\theta\hat{g}(X) - g(\theta))^2,
\end{aligned}$$

bo wartość oczekiwana iloczynu mieszanego jest zerem. □

Rozkład BŚK na dwa składniki: wariancję i kwadrat obciążenia wskazuje, mówiąc nieprecyzyjnie, na dwa odmienne źródła błędu estymacji.

3.3.5 PRZYKŁAD (Nieporównywalne estymatory). Rozważmy ciąg obserwacji $X_1, \ldots, X_n$ będący próbką z rozkładu $\mathrm{N}(\mu, 1)$. Wiemy, że wariancja jest równa 1 i jedynym nieznanym parametrem jest μ. Spróbujmy porównać „bardzo rozsądny" estymator ze „zgadywaniem w ciemno". Niech

$$\hat{\mu}_1 = \bar{X}, \qquad \hat{\mu}_2 \equiv 5.$$

Mamy, jak łatwo widzieć, $R_1(\mu) \equiv 1/n$ i $R_2(\mu) = (\mu - 5)^2$. Estymator $\hat{\mu}_1$ jest nieobciążony, zaś $\hat{\mu}_2$ ma zerową wariancję. Nie można twierdzić, że jeden z tych estymatorów ma jednostajnie mniejszy BŚK. Jednostajnie, to znaczy dla wszystkich wartości nieznanego parametru μ. Wykresy funkcji R_1 i R_2 się „krzyżują".

Przykład jest wyjątkowo jaskrawy ale nieporównywalność estymatorów jest raczej powszechnym zjawiskiem w statystyce.

3.3.6 PRZYKŁAD (Dwa estymatory wariancji rozkładu normalnego). Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu $\mathrm{N}(\mu, \sigma^2)$. Rozważmy dwa estymatory wariancji:

$$\hat{\sigma}_1^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2;$$
$$\hat{\sigma}_2^2 = \tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Pokazaliśmy już wcześniej, przy okazji twierdzenia Fishera, że $\mathbb{E}_{\mu,\sigma}S^2 = \sigma^2$ i $\mathrm{Var}_{\mu,\sigma}S^2 = (2/(n-1))\sigma^4$. Estymator $\hat{\sigma}_1^2$ ma więc obciążenie $b_1(\theta) = 0$ i ryzyko tego estymatora jest równe

$$R_1 = \frac{2}{n-1}\sigma^4.$$

Ponieważ $\tilde{S}^2 = S^2(n-1)/n$, więc wnioskujemy, że $\mathbb{E}_{\mu,\sigma}\tilde{S}^2 = \sigma^2(n-1)/n$ i $\mathrm{Var}_{\mu,\sigma}\tilde{S}^2 = ((n-1)/n)^2(2/(n-1))\sigma^4 = 2((n-1)/n^2)\sigma^4$. Estymator $\hat{\sigma}_2^2$ ma więc obciążenie $b_2(\theta) = -(1/n)\sigma^2$ i ryzyko

$$R_2 = \frac{2(n-1)}{n^2}\sigma^4 + \frac{1}{n^2}\sigma^4 = \frac{2n-1}{n^2}\sigma^4.$$

Natychmiast można sprawdzić, że zawsze mamy $R_2 < R_1$. Co prawda, $\tilde{S}^2$ ma *ujemne obciążenie*, czyli systematycznie zaniża wielkość wariancji σ^2, lecz z drugiej strony ma *mniejszą wariancję* niż S^2, to znaczy mniejszy „rozrzut losowy".

Chociaż estymator $\tilde{S}^2$ ma jednostajnie mniejszy BŚK niż S^2, przeważnie używa się nieobciążonego estymatora wariancji S^2. Własność nieobciążoności jest uznawana za istotną zaletę. Rzecz jasna, dla estymatora nieobciążonego, BŚK pokrywa się z wariancją:

$$R(\theta) = \mathrm{Var}_\theta \hat{g}(X).$$

Poniższe twierdzenie pokazuje jak można „poprawiać" estymatory nieobciążone.

3.3.7 TWIERDZENIE (Rao-Blackwella). *Jeśli $\hat{g}(X)$ jest nieobciążonym estymatorem $g(\theta)$ i $T = T(X)$ jest statystyką dostateczną, to warunkowa wartość oczekiwana*

$$g^*(T) = \mathbb{E}_\theta(\hat{g}(X)|T)$$

jest nieobciążonym estymatorem $g(\theta)$ i nierówność $\mathrm{Var}_\theta(g^*(T)) \leqslant \mathrm{Var}_\theta(\hat{g}(X))$ zachodzi dla każdego θ.

Dowód. Najpierw zauważmy, że $g^*(T)$ jest *statystyką*. Wynika to ze Stwierdzenia 2.3.2: ponieważ rozkład warunkowy X przy danym $T = t$ nie zależy od θ, to wartość oczekiwana definiująca $g^*(T)$ też nie zależy od θ. Nieobciążoność wynika ze wzoru na prawdopodobieństwo całkowite: $\mathbb{E}_\theta\mathbb{E}_\theta(\hat{g}(X)|T) = \mathbb{E}_\theta\hat{g}(X) = g(\theta)$. Ze znanego wzoru na „dekompozycję wariancji" wynika, że $\mathrm{Var}_\theta\hat{g}(X) = \mathrm{Var}_\theta\mathbb{E}_\theta(\hat{g}(X)|T) + \mathbb{E}_\theta\mathrm{Var}_\theta(\hat{g}(X)|T) \geqslant \mathrm{Var}_\theta\mathbb{E}_\theta(\hat{g}(X)|T)$. □

3.3.8 PRZYKŁAD (Rozkład wykładniczy; ENMW($\mathrm{e}^{-\theta a}$)). Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu $\mathrm{Ex}(\theta)$. Interesuje nas estymacja funkcji $g(\theta) = \mathrm{e}^{-\theta a} = \mathbb{E}_\theta(X_1 > a)$, dla danego $a > 0$. Estymator $\hat{g}(X_1, \ldots, X_n) = \mathbb{I}(X_1 > a)$ jest marny, ale nieobciążony. Poprawimy go korzystając z faktu, że $S = \sum_{i=1}^n X_i$ jest statystyką dostateczną. Zadanie 12 pokazuje, że rozkład warunkowy $(X_1, \ldots, X_{n-1})$ przy danym $S = s$ jest jednostajny na sympleksie $\{(x_1, \ldots, x_{n-1}) : x_i \geqslant 0, \sum_{i=1}^n x_i \leqslant s\}$. Wobec tego $g^*(s) = \mathbb{E}_\theta(\hat{g}|S = s) = \mathbb{P}_\theta(X_1 > a|S = s) = (1 - a/s)^{n-1}$ dla $s \geqslant a$ i zero dla $s < a$. Ostatecznie,

$$g^*(s) = \begin{cases} \left(1 - \frac{a}{S}\right)^{n-1} & \text{dla } s \geqslant a; \\ 0 & \text{dla } s < a. \end{cases}$$

Przykład 3.3.5 dobitnie pokazał, że poszukiwanie najlepszych estymatorów jest zadaniem beznadziejnym. Jeśli ograniczymy się tylko do estymatorów nieobciążonych, sytuacja zmienia się zasadniczo. W wielu ciekawych zadaniach istnieje estymator *najlepszy wśród nieobciążonych.*

3.3.9 DEFINICJA. *Statystyka $g^*(X)$ jest* ***estymatorem nieobciążonym o minimalnej wariancji*** *(ENMW) wielkości $g(\theta)$, jeśli*

(i) $g^*(X)$ jest estymatorem nieobciążonym $g(\theta)$;

(ii) dla każdego nieobciążonego estymatora $\hat{g}(X)$ spełniona jest nierówność $\mathrm{Var}_\theta g^*(X) \leqslant \mathrm{Var}_\theta\hat{g}(X)$ dla każdego $\theta \in \Theta$.

Estymator skonstruowany w Przykładzie 3.3.8 w rzeczywistości jest ENMW. Pozostawmy ten fakt bez dowodu. Twierdzenie Rao-Blackwella często pozwala skonstruować dobre estymatory nieobciążone, ale nie daje gwarancji, że są one ENMW.

3.4 Informacja Fishera i nierówność Craméra-Rao

3.4.1 DEFINICJA (Informacja Fishera). *Niech X będzie zmienną losową o gęstości f_θ, zależnej od jednowymiarowego parametru $\theta \in \Theta \subset \mathbb{R}$. Funkcję*

$$I(\theta) = \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right)^2$$

nazywamy informacją Fishera zawartą w obserwacji X.

Jak zwykle, gęstość f_θ w powyższej definicji jest rozumiana w szerszym sensie, obejmującym również zmienne dyskretne. Mamy więc

$$\begin{aligned} I(\theta) &= \int \left(\frac{\partial}{\partial \theta} \log f_\theta(x) \right)^2 f_\theta(x) \mathrm{d}x \text{ dla zmiennej ciągłej;} \\ I(\theta) &= \sum_x \left(\frac{\partial}{\partial \theta} \log f_\theta(x) \right)^2 f_\theta(x) \text{ dla zmiennej dyskretnej.} \end{aligned} \tag{3.4.2}$$

Jeżeli mówimy o informacji zawartej w wektorze obserwacji $X = (X_1, \ldots, X_n)$ to f_θ jest gęstością łączną, zaś całkę należy rozumieć jako całkę wielokrotną.

3.4.3 PRZYKŁAD (Rozkłady Poissona). Jeśli

$$f_\theta(x) = \mathbb{P}_\theta(X = x) = \mathrm{e}^{-\theta} \frac{\theta^x}{x!}, \qquad (x = 0, 1, 2, \ldots)$$

to $\log f_\theta(x) = -\theta + x \log \theta - \log(x!)$ i $(\partial/\partial\theta) \log f_\theta(x) = -1 + x/\theta$. Mamy więc

$$\mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \log f_\theta(x) \right)^2 = \sum_{x=0}^{\infty} \left(\frac{x}{\theta} - 1 \right)^2 \mathrm{e}^{-\theta} \frac{\theta^x}{x!} = \frac{1}{\theta^2} \mathrm{Var}_\theta(X) = \frac{1}{\theta}.$$

Ostatecznie, $I(\theta) = 1/\theta$.

3.4.4 PRZYKŁAD (Rozkłady wykładnicze). Jeśli

$$f_\theta(x) = \theta \mathrm{e}^{-\theta x}, \qquad (x > 0)$$

to $\log f_\theta(x) = \log\theta - \theta x$ i $(\partial/\partial\theta)\log f_\theta(x) = 1/\theta - x$. Teraz

$$\mathbb{E}_\theta \left(\frac{\partial}{\partial\theta}\log f_\theta(x)\right)^2 = \int_0^\infty \left(\frac{1}{\theta} - x\right)^2 \theta \mathrm{e}^{-\theta x}\mathrm{d}x = \mathrm{Var}_\theta(X) = \frac{1}{\theta^2}.$$

Zatem $I(\theta) = 1/\theta^2$.

3.4.5 Uwaga (Warunki regularności). Rodzina gęstości musi być „dostatecznie regularna" aby pewne kroki rachunkowe w dalszych rozumowaniach były poprawne. Ścisłe sformułowanie potrzebnych założeń przekracza ramy naszego wykładu. W dużym uproszczeniu zakładamy co następuje.

(i) Informacja Fishera jest dobrze określona. Zakładamy, że Θ jest przedziałem otwartym, istnieje pochodna $(\partial/\partial\theta)\log f_\theta$, całka/suma we wzorze (3.4.2) jest bezwzględnie zbieżna i $0 < I(\theta) < \infty$

(ii) Wszystkie gęstości f_θ mają ten sam „nośnik", to znaczy zbiór $\{x \in \mathcal{X} : f_\theta(x) > 0\}$ nie zależy od θ.

(iii) Można „przenosić pochodną przed znak całki", czyli zamienić kolejność operacji różniczkowania $(\partial/\partial\theta)$ i całkowania $\int \cdots \mathrm{d}x$.

Sens mglistego sformułowania (iii) wyjaśni się częściowo w trakcie dowodów. W przypadku zmiennych dyskretnych, całkę należy zastąpić przez sumę.

3.4.6 PRZYKŁAD. Rodzina gęstości jednostajnych:

$$f_\theta(x) = \begin{cases} 1/\theta & \text{dla } 0 \leqslant x \leqslant \theta; \\ 0 & \text{w przeciwnym przypadku,} \end{cases}$$

dla $\theta \in \Theta =]0, \infty[$. Ta rodzina nie spełnia warunku (ii).

3.4.7 Stwierdzenie. *Jeśli spełnione są warunki regularności* (3.4.5) *to*

$$\begin{aligned}
&(i) \quad \mathbb{E}_\theta \frac{\partial}{\partial\theta} \log f_\theta(x) = 0, \\
&(ii) \quad I(\theta) = \mathrm{Var}_\theta \left(\frac{\partial}{\partial\theta} \log f_\theta(X) \right) \\
&(iii) \quad I(\theta) = -\mathbb{E}_\theta \left(\frac{\partial^2}{\partial\theta^2} \log f_\theta(X) \right),
\end{aligned}$$

Dowód. Posłużymy się uproszczoną notacją, która powinna być zrozumiała. Pominiemy argumenty θ i x, pisząc $f = f_\theta(x)$. "Prim" oznacza wszędzie pochodną $(\partial/\partial\theta)$, zaś całkujemy względem x. Teza (i) wynika z następującego rachunku:

$$\begin{aligned}
\mathbb{E}_\theta \frac{\partial}{\partial\theta} \log f_\theta(X) &= \int (\log f)' f \\
&= \int \frac{f'}{f} f = \int (f') = \left(\int f \right)' = 1' = 0.
\end{aligned}$$

Pisząc $\int (f') = \left(\int f\right)'$, zamieniliśmy kolejność operacji $(\partial/\partial\theta)$ i $\int \cdots \mathrm{d}x$. Punkt (ii) jest natychmiastową konsekwencją (i). Dowód (iii) jest podobny:

$$\begin{aligned}
\mathbb{E}_\theta \frac{\partial^2}{\partial\theta^2} \log f_\theta(X) &= \int (\log f)'' f \\
&= \int \left(f' \frac{1}{f} \right)' f = \int \frac{f''}{f} f - \int f' \frac{f'}{f^2} f = 0 - \int \left(\frac{f'}{f} \right)^2 f \\
&= -\int \Big((\log f)' \Big)^2 f = -\mathbb{E}_\theta \left(\frac{\partial}{\partial\theta} \log f_\theta(X) \right)^2,
\end{aligned}$$

Zamiana kolejności operacji $(\partial/\partial\theta)$ i $\int \cdots \mathrm{d}x$ jest tu potrzebna aby uzasadnić równość $\int f'' = \left(\int f\right)'' = 1'' = 0$. □

3.4.8 Stwierdzenie. *Niech $X = (X_1, \ldots, X_n)$ będzie ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie. Jeżeli $I_n(\theta)$ jest informacją Fishera zawartą w n-wymiarowej obserwacji X, zaś $I_1(\theta)$ jest informacją zawartą w pojedynczej współrzędnej X_1, to*

$$I_n(\theta) = nI_1(\theta).$$

Dowód. Najłatwiej skorzystać ze Stwierdzenia 3.4.7 (ii):

$$\begin{aligned} I_n(\theta) &= -\mathrm{Var}_\theta \left(\frac{\partial^2}{\partial \theta^2} \log(f_\theta(X_1) \cdots f_\theta(X_n)) \right) \\ &= -\mathrm{Var}_\theta \left(\frac{\partial}{\partial \theta} \sum_{i=1}^{n} \log f_\theta(X_i) \right) \\ &= -\sum_{i=1}^{n} \mathrm{Var}_\theta \left(\frac{\partial}{\partial \theta} \log f_\theta(X_i) \right) = nI_1(\theta). \end{aligned}$$

Oczywiście, wynika to z niezależności obserwacji i z faktu, że gęstość każdej z nich jest jednakowa. □

3.4.9 TWIERDZENIE (Nierówność Cramèra-Rao). *Załóżmy, że model spełnia warunki regularności* (3.4.5) *i* $\hat{g}(X)$ *jest nieobciążonym estymatorem* $g(\theta)$. *Wtedy dla każdego* $\theta \in \Theta$,

$$\mathrm{Var}_\theta \, \hat{g}(X) \geqslant \frac{(g'(\theta))^2}{I(\theta)}.$$

W szczególności, jeśli $\hat{\theta}(X)$ *jest nieobciążonym estymatorem parametru* θ, *to*

$$\mathrm{Var}_\theta \hat{\theta}(X) \geqslant 1/I(\theta).$$

Dowód. Zastosujemy nierówność Schwarza do zmiennych losowych $\hat{g}(X)$ i $(\partial/\partial\theta) \log f_\theta(X)$:

$$\begin{aligned} \mathrm{Var}_\theta \, \hat{g}(X) \, I(\theta) &= \mathrm{Var}_\theta \, \hat{g}(X) \mathrm{Var}_\theta \left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right) \\ &\geqslant \left(\mathrm{Cov}_\theta \left(\hat{g}(X), \frac{\partial}{\partial \theta} \log f_\theta(X) \right) \right)^2 \\ &= \left(\mathbb{E}_\theta \left(\hat{g}(X) \frac{\partial}{\partial \theta} \log f_\theta(X) \right) \right)^2 \\ &= \left(\int \hat{g}(x) \left(\frac{\partial}{\partial \theta} \log f_\theta(x) \right) f_\theta(x) \mathrm{d}x \right)^2 = (*) \end{aligned}$$

Skorzystaliśmy kolejno: ze Stwierdzenia 3.4.7 (ii), nierówności Schwarza i Stwierdzenia 3.4.7 (i).

Dalej,

$$\begin{aligned}(*) &= \left(\int \hat{g}(x)\left(\frac{\partial}{\partial\theta}\log f_\theta(x)\right) f_\theta(x)\mathrm{d}x\right)^2 \\ &= \left(\int \hat{g}(x)\frac{(\partial/\partial\theta)f_\theta(x)}{f_\theta(x)} f_\theta(x)\mathrm{d}x\right)^2 \\ &= \left(\int \frac{\partial}{\partial\theta}\hat{g}(x) f_\theta(x)\mathrm{d}x\right)^2 = \left(\frac{\partial}{\partial\theta}\int \hat{g}(x) f_\theta(x)\mathrm{d}x\right)^2 \\ &= \left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta\hat{g}(X)\right)^2 = \left(\frac{\partial}{\partial\theta}g(\theta)\right)^2.\end{aligned}$$

Pierwsza równość wynika ze wzoru (3.4.2), druga z „przeniesienia pochodnej przed znak całki" na mocy warunku regularności (3.4.5) (iii). Wreszcie, ostatnia równość wynika z założenia o nieobciążoności estymatora. □

Nierówność Craméra-Rao stwierdza, że funkcja ryzyka każdego nieobciążonego estymatora nie może nigdzie być mniejsza od pewnej ustalonej funkcji. Istnieje granica jakości nieprzekraczalna dla wszystkich estymatorów *nieobciążonych*. Jeśli ryzyko jakiegoś estymatora nieobciążonego jest *równe* dolnemu oszacowaniu Cramèra-Rao, to jest oczywiste, że jest to estymator najlepszy wśród nieobciążonych, a więc ENMW.

3.4.10 Wniosek. *Jeśli $\hat{g}(X_1,\ldots,X_n)$ jest nieobciążonym estymatorem $g(\theta)$, obliczonym na podstawie niezależnej próbki losowej $X_1,\ldots,X_n$ to*

$$\mathrm{Var}_\theta\hat{g}(X_1,\ldots,X_n) \geqslant \frac{(g'(\theta))^2}{nI_1(\theta)},$$

pod warunkiem, że spełnione są założenia regularności (3.4.5).

Jakość nieobciążonego estymatora parametru θ można ocenić, porównując wariancję tego estymatora z dolnym ograniczeniem Cramèra-Rao. Wielkość $1/I_n(\theta)$ jest po prostu wygodnym *punktem odniesienia*. Co prawda wariancja ENMW często jest większa niż $1/I_n(\theta)$, ale może być bardzo trudna do obiczenia. Poza tym ENMW może w ogóle nie istnieć. Te same uwagi dotyczą nieco ogólniejszej sytuacji, kiedy estymujemy pewną funkcję parametru. Dlatego przyjmiemy następującą definicję:

Efektywność nieobciążonego estymatora $\hat{g}$ wielkości $g(\theta)$ określamy jako

$$\mathrm{ef}(\hat{g}) = \frac{(g'(\theta))^2}{\mathrm{Var}_\theta(\hat{g})I(\theta)}.$$

Użyteczne jest też pojęcie efektywności względnej dwóch estymatorów. Jeśli $\hat{g}_1$ i $\hat{g}_2$ są nieobciążone to

$$\mathrm{ef}(\hat{g}_1, \hat{g}_2) = \frac{\mathrm{Var}_\theta(\hat{g}_2)}{\mathrm{Var}_\theta(\hat{g}_1)}$$

nazywamy *efektywnością* estymatora $\hat{g}_1$ *względem* $\hat{g}_2$. Oczywiście, $\mathrm{ef}(\hat{g}_1, \hat{g}_2) = \mathrm{ef}(\hat{g}_1)/\mathrm{ef}(\hat{g}_2)$.

Jeśli spełniona jest nierówność Cramèra-Rao, to $\mathrm{ef}(\hat{g}) \leqslant 1$ dla każdego estymatora nieobciążonego. Jeśli $\hat{g} = \mathrm{ENMW}(g)$ to może się zdarzyć, że $\mathrm{ef}(\hat{g}) = 1$ ale też może być $\mathrm{ef}(\hat{g}) < 1$. Zwróćmy jeszcze uwagę na to, że $\mathrm{ef}(\hat{g})$ jest na ogół funkcją parametru θ (a nie liczbą).

3.4.11 PRZYKŁAD (Rozkład Poissona; ENMW(θ)). Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu Poiss(θ), gdzie $\theta > 0$ jest nieznanym parametrem. Średnia z próbki $\bar{X}$ jest ENMW(θ). Istotnie, wiemy, że $I_n(\theta) = nI(\theta) = n/\theta$. Estymator $\bar{X}$ jest nieobciążony i jego wariancja jest równa dolnemu ograniczeniu Cramèra-Rao, bo

$$\mathrm{Var}_\theta(\bar{X}) = \frac{\theta}{n} = \frac{1}{I_n(\theta)}.$$

Każdy inny estymator nieobciążony musi więc mieć wariancję nie mniejszą, niż wariancja $\bar{X}$.

3.4.12 PRZYKŁAD (Rozkład wykładniczy; ENMW($1/\theta$)). Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu Ex(θ), gdzie $\theta > 0$ jest nieznanym parametrem. Średnia z próbki $\bar{X}$ jest ENMW($1/\theta$), bo

$$\mathrm{Var}_\theta(\bar{X}) = \frac{1}{n\theta^2} = \frac{((1/\theta)')^2}{I_n(\theta)}.$$

Stosujemy tu Wniosek 3.4.10 do funkcji $g(\theta) = 1/\theta$. Wiemy, że $I_n(\theta) = nI_1(\theta) = n/\theta^2$. Średnia z próbki jest estymatorem efektywnym, to znaczy ma efektywność 1.

Twierdzenie 3.4.9 nie daje uniwersalnej metody sprawdzania, że estymator jest ENMW. Pokazuje to następny przykład.

3.4.13 PRZYKŁAD (Rozkład wykładniczy; ENMW(θ)). Tak jak poprzednio, niech $X_1, \ldots, X_n$ będzie próbką z rozkładu Ex(θ). Naturalnym estymatorem parametru θ jest $1/\bar{X}$, ale jest to estymator *obciążony*. W istocie, zmienna losowa $Y = \sum X_i$ ma rozkład Gamma(n, θ), więc

$$\mathbb{E}_\theta \frac{1}{Y} = \int_0^\infty \frac{1}{y} \frac{\theta^n}{(n-1)!} y^{n-1} \mathrm{e}^{-\theta y} \mathrm{d}y = \int_0^\infty \frac{\theta^n}{(n-1)!} y^{n-2} \mathrm{e}^{-\theta y} \mathrm{d}y = \frac{\theta}{n-1}.$$

Widzimy, że $\mathbb{E}_\theta(1/\bar{X}) = n/(n-1)$. Łatwo zmodyfikować nasz wyjściowy estymator tak, żeby "usunąć obciążenie":

$$\hat{\theta} = \frac{n-1}{Y} = \frac{n-1}{n\bar{X}}$$

jest estymatorem nieobciążonym. Co więcej, $\hat{\theta} = \mathrm{ENMW}(\theta)$, choć to trudniej pokazać. Zaakceptujmy ten fakt bez dowodu. Zbadajmy, jaka jest wariancja $\hat{\theta}$. Podobnie jak poprzednio, obliczamy

$$\mathbb{E}_\theta \frac{1}{Y^2} = \int_0^\infty \frac{1}{y^2} \frac{\theta^n}{(n-1)!} y^{n-1} \mathrm{e}^{-\theta y} \mathrm{d}y = \frac{\theta^2}{(n-1)(n-2)}.$$

Stąd $\mathbb{E}_\theta(\hat{\theta}^2) = \theta^2(n-1)/(n-2)$ i $\mathrm{Var}_\theta \hat{\theta} = \theta^2/(n-2)$. Dla rodziny rozkładów wykładniczych znamy nformację Fishera, $I_1(\theta) = n/\theta^2$. Możemy teraz porównać wariancję z ograniczeniem wynikającym z nierówności Cramèra-Rao.

$$\mathrm{Var}_\theta \hat{\theta} = \frac{\theta^2}{n-2} > \frac{\theta^2}{n} = \frac{1}{I_n(\theta)}.$$

Nierówność jest *ostra*. ENMW nie osiąga dolnego ograniczenia Cramèra-Rao. Mamy $\mathrm{ef}(\hat{\theta}) = n/(n-2) < 1$.

3.5 Zadania

1. Samoloty bombowe przedzierają się przez dwie linie obrony przeciwlotniczej. Każdy samolot, niezależnie od pozostałych, z prawdopodobieństwem θ może zostać strącony przez pierwszą linię obrony. Jeśli pokona pierwszą linię, z prawdopodobieństwem θ może zostać strącony przez drugą linię. Prawdopodobieństwo θ jest nieznane. Spośród $n = 100$ samolotów, $K_1 = 40$ zostało strąconych przez pierwszą linię, a dalszych $K_2 = 20$ zostało strąconych przez drugą linię.

   (a) Oblicz wiarogodność dla zaobserwowanych wartości K_1 i K_2, czyli $\mathcal{L} = \mathbb{P}_\theta(K_1 = 40, K_2 = 20)$.

   (b) Podaj *estymator największej wiarogodności* parametru θ.

2. W modelu Hardy'ego-Weinberga, Przykłady 3.1.2 i 4.2.4, pokazać, $\hat{\theta} = \mathrm{ENW}(\theta)$ jest ENMW.

   Wskazówka: Obliczyć $Var\hat{\theta}$ i porównać z dolnym ograniczeniem Cramér-Rao.

3. Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu $\mathrm{N}(\mu, \sigma^2)$, z nieznanymi parametrami μ i σ. Dobrać stałą c tak, żeby $\hat{\sigma} = cS$ był estymatorem nieobciążonym odchylenia standardowego σ. Jak zwykle, S^2 jest nieobciążonym estymatorem wariancji (Przykład 3.3.6), zaś $S = \sqrt{S^2}$.

4. Niech $X \sim \mathrm{Bin}(n, \theta)$ będzie liczbą sukcesów w schemacie Bernoulliego. Obliczyć i porównać MSE dwóch estymatorów: ENW $\hat{\theta} = X/n$ oraz $\tilde{\theta} = (X + 1)/(n + 2)$.

   Uwaga: W Części V okaże się, że $\tilde{\theta}$ jest estymatorem Bayesowskim przy jednostajnym rozkładzie *a priori*.

Rozdział 4

Asymptotyczne własności estymatorów

Zajmiemy się teraz sytuacją, kiedy rozmiar próbki $X_1, \ldots, X_n$ jest duży. Własności asymptotyczne są to, z matematycznego punktu widzenia, twierdzenia graniczne, w których n dąży do nieskończoności. W praktyce, te twierdzenia opisują *w przybliżeniu* zachowanie estymatorów dla „dostatecznie dużych" próbek. Niestety, teoria asymptotyczna (przynajmniej w najprostszej swojej postaci) nie dostarcza informacji o tym, *jak duża* powinna być próbka, żeby przybliżenie było *dostatecznie dobre*.

Wyobraźmy sobie (potencjalnie) nieskończony ciąg obserwacji $X_1, \ldots, X_n, \ldots$. Ograniczymy się do rozpatrzenia sytuacji, kiedy obserwacje są *niezależnymi zmiennymi losowymi o jednakowym rozkładzie* (zależnym, jak zwykle, od nieznanego parametru). Przez estymator rozumiemy funkcję zależącą od początkowych n obserwacji $X_1, \ldots, X_n$. Faktycznie, mówiąc o „estymatorze" będziemy mieli na myśli *ciąg* estymatorów $\hat{g}(X_1, \ldots, X_n)$, gdzie $n = 1, 2, \ldots$. Niekiedy będziemy pisali $\hat{g}_n = \hat{g}(X_1, \ldots, X_n)$, żeby uwidocznić rozmiar próbki. Na przykład,

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

i tym podobnie.

4.1 Zgodność

4.1.1 DEFINICJA. *Estymator* $\hat{g}(X_1, \ldots, X_n)$ *wielkości* $g(\theta)$ *jest* ***zgodny****, jeśli dla każdego* $\theta \in \Theta$*,*

$$\lim_{n\to\infty} \mathbb{P}_\theta \left(|\hat{g}(X_1, \ldots, X_n) - g(\theta)| \leqslant \varepsilon\right) = 1,$$

dla każdego $\varepsilon > 0$*. Estymator jest* **mocno zgodny***, jeśli*

$$\mathbb{P}_\theta \left(\lim_{n\to\infty} \hat{g}(X_1, \ldots, X_n) = g(\theta) \right) = 1. \qquad \square$$

Zgodność (mocna zgodność) znaczy tyle, że

$$\hat{g}(X_1, \ldots, X_n) \to g(\theta), \qquad (n \to \infty)$$

według prawdopodobieństwa (prawie na pewno). Innymi słowy, estymator jest zgodny, jeśli zmierza do estymowanej wielkości przy nieograniczonym powiększaniu próbki. W świetle twierdzenia Gliwienki-Cantelliego wydaje się jasne, że od każdego „rozsądnego" estymatora powinniśmy oczekiwać mocnej zgodności. W istocie, niemal wszystkie rozważane przez nas estymatory są mocno zgodne (wyjątkami są specjalnie dobrane „głupie" estymatory konstruowane dla zilustrowania pewnych osobliwości).

4.1.2 PRZYKŁAD (Momenty z próbki). Średnia z próbki, $\bar{X}_n$, jest mocno zgodnym estymatorem wartości oczekiwanej $\mu(\theta) = \mathbb{E}_\theta X_1$ dla każdej rodziny rozkładów prawdopodobieństwa takiej, że $\mu(\theta)$ jest dobrze określone. Zbieżność $\bar{X}_n \to_{\text{p.n.}} \mu(\theta)$ wynika po prostu z Mocnego Prawa Wielkich Liczb (MPWL). Podobnie, $\tilde{S}_n^2$ i S_n^2 są mocno zgodnymi estymatorami wariancji $\sigma^2(\theta) = \text{Var}_\theta X_1$. Wystarczy zauważyć, że $\tilde{S}_n^2 = \sum_1^n X_i^2/n - \bar{X}_n^2$ i $S_n^2 = \tilde{S}_n^2 n/(n-1)$ i znowu powołać się na MPWL.

4.1.3 PRZYKŁAD (Rozkłady Pareto). Niech $X_1, X_2 \ldots$, będzie próbką z rozkładu Pareto(α, λ). Z definicji, jest to rozkład o dystrybuancie $F_{\alpha,\lambda}(x) = 1 - \lambda^\alpha/(x-\lambda)^\alpha$ dla $x > 0$ i gęstości

$$f_{\alpha,\lambda}(x) = \begin{cases} \alpha\lambda^\alpha/(x-\lambda)^{\alpha+1} & \text{dla } x > 0; \\ 0 & \text{wpp.} \end{cases}$$

Jeśli $\alpha > 1$, to wartość oczekiwana rozkładu Pareto jest równa $\lambda/(\alpha-1)$. Jeśli $\alpha > 2$, to wariancja jest równa $\alpha\lambda^2/((\alpha-1)^2(\alpha-2))$. To łatwo można obliczyć metodą całkowania przez części. Dla $\alpha \leqslant 1$, wartość oczekiwana jest nieskończona. Dla $\alpha \leqslant 2$, wariancja nie istnieje.

Metoda momentów prowadzi do następujących estymatorów parametrów α i λ:

$$\hat{\alpha}_n = \frac{2\tilde{S}_n^2}{\tilde{S}_n^2 - \bar{X}_n^2}, \qquad \hat{\lambda}_n = (\hat{\alpha}_n - 1)\bar{X}_n.$$

Jeśli ograniczymy się do rozkładów Pareto mających skończoną wariancję, czyli rozpatrzymy rodzinę $\{\text{Pareto}(\alpha,\lambda), \alpha > 2, \lambda > 0\}$, to estymatory są mocno zgodne: $\hat{\alpha}_n \to_{\text{p.n.}} \alpha$ i $\hat{\lambda}_n \to_{\text{p.n.}} \lambda$. Wynika to z MPWL, podobnie jak w poprzednim przykładzie. Dla $\alpha \leqslant 2$ sytuacja się zmienia: wariancja rozkładu Pareto nie istnieje i nie należy oczekiwać, że metoda momentów da przyzwoite wyniki. W istocie, dla każdego $\alpha \leqslant 2$ mamy $\hat{\alpha}_n \to_{\text{p.n.}} 2$ przy $n \to \infty$. (łatwo to uzasadnić dla $1 < \alpha \leqslant 2$, bo wtedy $\tilde{S}_n^2 \to_{\text{p.n.}} \infty$ i $\bar{X}_n \to_{\text{p.n.}} \lambda/(\alpha-1)$). Podsumowując, estymatory otrzymane metodą momentów *nie są zgodne*, jeśli rozpatrujemy rodzinę *wszystkich* rozkładów Pareto $\{\text{Pareto}(\alpha,\lambda), \alpha > 0, \lambda > 0\}$.

Estymatory największej wiarogodności są *zazwyczaj* mocno zgodne. Dowody (mocnej) zgodności wykorzystują (mocne) PWL, ale szczegóły mogą być żmudne. Nie będziemy się przy tym zatrzymywali, bo zgodność (nawet w mocnym sensie) nie jest specjalnie satysfakcjonującą własnością estymatora. Jest zaledwie minimalnym żądaniem, które powinien spełniać każdy „przyzwoity" estymator.

4.2 Asymptotyczna normalność

4.2.1 DEFINICJA. *Mówimy, że estymator $\hat{g}(X_1,\ldots,X_n)$ wielkości $g(\theta)$ jest **asymptotycznie normalny**, jeśli dla każdego $\theta \in \Theta$ istnieje funkcja $\sigma^2(\theta)$, zwana asymptotyczną wariancją, taka że*

$$\sqrt{n}\,(\hat{g}(X_1,\ldots,X_n) - g(\theta)) \to_{\text{d}} \text{N}\big(0,\sigma^2(\theta)\big), \qquad (n\to\infty).$$

Estymator jest asymptotycznie normalny, jeśli

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(\frac{\sqrt{n}}{\sigma(\theta)}\big(\hat{g}(X_1,\ldots,X_n)-g(\theta)\big)\le a\right)=\Phi(a).$$

Mówiąc jeszcze inaczej, rozkład prawdopodobieństwa statystyki $\hat{g}(X_1,\ldots,X_n)$ jest dla dużych n zbliżony do rozkładu

$$\mathrm{N}\left(g(\theta),\frac{\sigma^2(\theta)}{n}\right).$$

Jeśli estymator jest asymptotycznie normalny, to jest zgodny, choć nie musi być *mocno* zgodny. Zazwyczaj dla estymatora asymptotycznie normalnego mamy

$$\begin{aligned}&\lim_{n\to\infty}\mathbb{E}_\theta\hat{g}(X_1,\ldots,X_n)=g(\theta),\\ &\lim_{n\to\infty} n\mathrm{Var}_\theta\hat{g}(X_1,\ldots,X_n)=\sigma^2(\theta),\end{aligned}\tag{4.2.2}$$

ale te relacje *nie wynikają* z Definicji 4.2.1. Chodzi o to, że zbieżność według rozkładu nie pociąga za sobą zbieżności wartości oczekiwanych ani wariancji. Kontrprzykłady mają jednak dość "patologiczny" charakter i nie będziemy ich przytaczać.

Dlaczego oczekujemy, że wariancja aproksymującego rozkładu normalnego jest *odwrotnie proporcjonalna* do rozmiaru próbki n? Najprostszą statystyką jest średnia z próbki. Centralne Twierdzenie Graniczne (CTG) mówi, że

$$\sqrt{n}(\bar{X}_n-\mu)\to_{\mathrm{d}}\mathrm{N}\left(0,\sigma^2\right),$$

gdzie μ i σ^2 oznaczają wartość oczekiwaną i wariancję pojedynczej obserwacji X_i. Asymptotyczna wariancja σ^2/n pokrywa się w tu po prostu z wariancją statystyki $\bar{X}$. Dla wielu innych statystyk, asymptotyczna normalność wraz z odpowiednią postacią asymptotycznej wariancji daje się też wywnioskować z CTG. Pomocny jest następujący lemat.

4.2.3 Lemat (Metoda delta)**.** *Jeżeli dla ciągu zmiennych losowych T_n mamy $\sqrt{n}\big(T_n-\mu\big)\to_{\mathrm{d}}\mathrm{N}(0,\sigma^2)$ przy $n\to\infty$ i $h:\mathbb{R}\to\mathbb{R}$ jest funkcją różniczkowalną w punkcie μ, to*

$$\sqrt{n}\Big(h(T_n)-h(\mu)\Big)\to_{\mathrm{d}}\mathrm{N}\Big(0,\sigma^2\big(h'(\mu)\big)^2\Big).$$

Dla uproszczenia odstąpiliśmy chwilowo od notacji, jawnie wskazującej na zależność rozkładów prawdopodobieństwa od nieznanego parametru. W zastosowaniach zarówno μ jak i σ^2 będą funkcjami parametru θ.

Dowód. Rozwińmy funkcję h zgodnie ze wzorem Taylora, uwzględniając tylko wyrazy rzędu zerowego i pierwszego. Resztę oznaczymy przez $r(t)$:

$$h(t) = h(\mu) + h'(\mu)(t-\mu) + r(t), \qquad \text{gdzie } \frac{r(t)}{t-\mu} \to 0, \quad (t \to \mu).$$

Wstawmy teraz zmienną losową T_n w miejsce t i pomnóżmy równanie przez $\sqrt{n}$. Dostajemy

$$\sqrt{n}\big(h(T_n) - h(\mu)\big) = h'(\mu)\sqrt{n}(T_n - \mu) + \sqrt{n}r(T_n). \qquad (*)$$

Bezpośrednio z założenia wynika, że

$$h'(\mu)\sqrt{n}(T_n - \mu) \to_{\mathrm{d}} \mathrm{N}\Big(0, \sigma^2\big(h'(\mu)\big)^2\Big). \qquad (**)$$

Pozostaje zająć się wyrazem związanym z resztą we wzorze $(*)$. Pokażemy, że

$$\sqrt{n}r(T_n) \to_{\mathrm{P}} 0, \quad (n \to \infty) \qquad (***)$$

Istotnie,

$$\sqrt{n}r(T_n) = \frac{r(T_n)}{T_n - \mu}\sqrt{n}(T_n - \mu).$$

Ponieważ $T_n \to_{\mathrm{P}} \mu$, więc $r(T_n)/(T_n - \mu) \to_{\mathrm{P}} 0$. Z kolei $\sqrt{n}(T_n - \mu) \to_{\mathrm{d}} \mathrm{N}(0, \sigma^2)$. Na mocy Lematu Słuckiego, iloczyn zmiennych losowych zbieżnych według prawdopodobieństwa do zera i zmiennych zbieżnych według rozkładu dąży według prawdopodobieństwa do zera. Ostatecznie, teza lematu wynika z $(*)$, $(**)$ i $(***)$. □

4.2.4 PRZYKŁAD (Model Hardy'ego-Weinberga). Wróćmy do Przykładu 3.1.2. Obserwujemy trójkę zmiennych losowych (N_1, N_2, N_3) o rozkładzie wielomianowym $\mathrm{Mult}(n, p_1, p_2, p_3)$, gdzie $p_1 = \theta^2$, $p_2 = 2\theta(1-\theta)$ i $p_3 = (1-\theta)^2$: Niech

$$\hat{\theta} = \sqrt{\frac{N_1}{n}}.$$

Asymptotyczna normalność estymatora $\hat{\theta}$ wynika z zastosowania metody delta do funkcji $h(x) = \sqrt{x}$, gdyż $\sqrt{n}(N_1/n - p_1) \to_{\mathrm{d}} \mathrm{N}(0, p_1(1-p_1))$ na mocy CTG. Ponieważ $(h'(p_1))^2 = 1/(4p_1)$, więc

$$\sqrt{n}(\hat{\theta} - \theta) \to_{\mathrm{d}} \mathrm{N}\Big(0, \frac{1}{4}(1-\theta^2)\Big).$$

4.3 Efektywność

4.3.1 TWIERDZENIE (Asymptotyczna normalność ENW). *Rozważmy rodzinę gęstości $\{f_\theta : \theta \in \Theta\}$ spełniającą warunki regularności (3.4.5). Niech $X_1, \ldots, X_n, \ldots$ będzie próbką z rozkładu o gęstości f_{θ_0}. Zakładamy, że $\Theta \subseteq \mathbb{R}$ i θ_0 jest punktem wewnętrznym zbioru Θ. Niech $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$ będzie ciągiem estymatorów takich, że $\sum_1^n (\partial/\partial\theta) \log f_{\hat{\theta}_n}(X_i) = 0$ i $\hat{\theta}_n \to_{\mathrm{P}} \theta_0$.*

Wtedy

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \to_{\mathrm{d}} \mathrm{N}\big(0, 1/I_1(\theta_0)\big), \qquad (n \to \infty).$$

Szkic dowodu. Jak zwykle, „prim" oznaczać będzie różniczkowanie względem parametru θ. Niech $\ell'(\theta) = \sum_1^n (\partial/\partial\theta) \log f_\theta(X_i)$. Napiszmy rozwinięcie tej funkcji wynikające z wzoru Taylora: $\ell'(\theta) = \ell'(\theta_0) + \ell''(\theta_0)(\theta - \theta_0) + r(\theta)$. Biorąc pod uwagę założenie (i), możemy napisać

$$0 = \ell'(\hat{\theta}_n) = \ell'(\theta_0) + \ell''(\theta_0)(\hat{\theta}_n - \theta_0) + r(\hat{\theta}_n).$$

Ponieważ $\hat{\theta}_n \to \theta_0$, więc reszta $r(\hat{\theta}_n)$ jest mała w porównaniu z pierwszymi wyrazami rozwinięcia. Zaniedbując resztę otrzymujemy *przybliżoną* równość $\hat{\theta}_n - \theta_0 \simeq -\ell'(\theta_0)/\ell''(\theta_0)$. Innymi słowy,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \simeq -\frac{\ell'(\theta_0)/\sqrt{n}}{\ell''(\theta_0)/n}. \qquad (*)$$

Zarówno $\ell'(\theta_0)$, jak i $\ell''(\theta_0)$ są sumami niezależnych zmiennych losowych o jednakowym rozkładzie. Zastosujmy CTG do licznika we wzorze $(*)$:

$$\ell'(\theta_0)/\sqrt{n} = \frac{\sum_1^n (\partial/\partial\theta) \log f_{\theta_0}(X_i)}{\sqrt{n}} \to_{\mathrm{d}} \mathrm{N}(0, I_1(\theta_0)),$$

bo zmienne losowe $(\partial/\partial\theta)\log f_{\theta_0}(X_i)$ są niezależne, mają wartość oczekiwaną zero i wariancję $I_1(\theta_0)$ na mocy Stwierdzenia 3.4.7 (i-ii). Z kolei PWL, zastosowane do mianownika we wzorze $(*)$, daje

$$\ell''(\theta_0)/n = \frac{1}{n}\sum_1^n (\partial^2/\partial\theta^2)\log f_{\theta_0}(X_i) \to_{\mathrm{P}} -I_1(\theta_0),$$

na mocy Stwierdzenia 3.4.7 (iii). Prawa strona $(*)$ zachowuje się w granicy tak, jak zmienna o rozkładzie $\mathrm{N}(0, I_1(\theta_0))$ podzielona przez stałą $I_1(\theta_0)$. Ma więc graniczny rozkład $\mathrm{N}(0, I_1(\theta_0)^{-1})$.

Idea dowodu jest prosta. Nieco żmudne jest tylko uzasadnienie tego, że rzeczywiście można zaniedbać resztę we wzorze Taylora. □

Twierdzenie 4.3.1 obejmuje estymatory największej wiarogodności takie, że $\ell'(\hat\theta_n) = 0$. Wyklucza to ENW(θ) w Przykładzie 3.2.8 (dla rodziny rozkładów jednostajnych). Zadanie 6 w Rozdziale 1 pokazuje, że ten estymator nie jest asymptotycznie normalny. Estymator w Przykładzie 3.2.7 (rozkłady Laplace'a) jest asymptotycznie normalny, mimo nieróżniczkowalności funkcji ℓ.

Jeśli rozważamy estymację wielkości $g(\theta)$ to przyjmujemy *z definicji*, że $g(\hat\theta) = \mathrm{ENW}(g(\theta))$, gdzie $\hat\theta = \mathrm{ENW}(\theta)$. Z Twierdzenia 4.3.1 i Lematu 4.2.3 (delta) otrzymujemy natychmiast asymptotyczną normalność tego estymatora:

4.3.2 Wniosek. *Przy założeniach Twierdzenia 4.3.1, jeśli g jest funkcją różniczkowalną, to*

$$\sqrt{n}\left(g(\hat\theta_n) - g(\theta)\right) \to_{\mathrm{d}} \mathrm{N}\left(0, \frac{g'(\theta)^2}{I_1(\theta)}\right), \qquad (n\to\infty).$$

4.3.3 TWIERDZENIE (Asymptotyczna efektywność ENW). *Rozważmy rodzinę gęstości $\{f_\theta : \theta\in\Theta\}$ spełniającą warunki regularności (3.4.5). Niech $X_1,\dots,X_n,\dots$ będzie próbką z rozkładu o gęstości f_θ. Niech $\hat g_n = \hat g(X_1,\dots,X_n)$ będzie ciągiem estymatorów różniczkowalnej funkcji $g(\theta)$. Jeżeli dla każdego θ,*

$$\sqrt{n}\,(\hat g_n - g(\theta)) \to_{\mathrm{d}} \mathrm{N}\big(0, \sigma^2(\theta)\big), \qquad (n\to\infty),$$

to nierówność

$$\sigma^2(\theta) \geqslant \frac{g'(\theta)^2}{I_1(\theta)}$$

zachodzi dla wszystkich θ z wyjątkiem, być może, zbioru o mierze Lebesgue'a zero.

Pominiemy dowód. Porównajmy Twierdzenie 4.3.3 i Wniosek 4.3.2. W pewnym uproszczeniu mówią one, że estymatory największej wiarogodności mają najmniejszą możliwą wariancję asymptotyczną. Mówi się, że są one asymptotycznie efektywne.

Niech $\hat{g} = \hat{g}_n = \hat{g}(X_1, \ldots, X_n)$ będzie estymatorem asymptotycznie normalnym, $\sqrt{n}(\hat{g}_n - g(\theta)) \to_{\mathrm{d}} \mathrm{N}(0, \sigma^2(\theta))$. Jego **asymptotyczną efektywność** określamy jako

$$\mathrm{as.ef}(\hat{g}) = \frac{(g'(\theta))^2}{\sigma^2(\theta) I_1(\theta)}.$$

Jest to oczywista modyfikacja definicji „zwykłej" efektywności: rolę wariancji estymatora nieobciążonego przejęła *asymptotyczna* wariancja estymatora asymptotycznie normalnego. Ponieważ zakładamy, że obserwacje $X_1, \ldots, X_n$ tworzą ciąg i.i.d. to zgodnie ze Stwierdzeniem 3.4.8 mamy $I_n(\theta) = nI_1(\theta)$. *Asymptotyczna efektywność względna* dwóch estymatorów jest określona analogicznie. Jeśli $\hat{g}_1 = \hat{g}_n^{(1)}$ i $\hat{g}_1 = \hat{g}_n^{(2)}$ są asymptotycznie normalne i mają asymptotyczne wariancje odpowiednio $\sigma_1^2(\theta)$ i $\sigma_2^2(\theta)$, to

$$\mathrm{as.ef}(\hat{g}_1, \hat{g}_2) = \frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}.$$

Oczywiście, $\mathrm{as.ef}(\hat{g}_1, \hat{g}_2) = \mathrm{as.ef}(\hat{g}_1)/\mathrm{as.ef}(\hat{g}_2)$. Pojęcie względnej efektywności asymptotycznej ma bardzo intuicyjną interpretację. Wyobraźmy sobie, że $\mathrm{as.ef}(\hat{g}_1, \hat{g}_2) = 3$. Wtedy estymatory $\hat{g}_n^{(1)} = \hat{g}_1(X_1, \ldots, X_n)$ i $\hat{g}_{3n}^{(2)} = \hat{g}_2(X_1, \ldots, X_{3n})$ mają w przybliżeniu ten sam rozkład $\mathrm{N}(g(\theta), \sigma_1^2(\theta)/n) = \mathrm{N}(g(\theta), \sigma_2^2(\theta)/(3n))$. To znaczy, że estymator $\hat{g}_2$ potrzebuje trzykrotnie więcej danych niż $\hat{g}_1$ żeby osiągnąć podobną dokładność. Krótko mówiąc, jest trzykrotnie mniej efektywny.

4.3.4 PRZYKŁAD (Estymacja prawdopodobieństwa braku szkód; rozkład Poissona). Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu Poiss(θ), gdzie $\theta > 0$. W naukach aktuarialnych, rozkład Poissona często opisuje liczby szkód w poszczególnych latach (dla pojedynczej polisy ubezpieczeniowej lub dla grupy polis). Interesującą wielkością jest $\mathrm{e}^{-\theta}$, ponieważ jest to prawdopodobieństwo, że w ciągu roku nie będzie żadnych szkód. Zajmiemy się zagadnieniem estymacji funkcji

$$g(\theta) = \mathrm{e}^{-\theta} = \mathbb{P}_\theta(X_1 = 0).$$

Rozpatrzmy dwa estymatory:

$$\hat{g}_1 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i = 0) = \frac{\text{liczba lat w których nie było szkód}}{\text{liczba lat}},$$
$$\hat{g}_2 = \mathrm{e}^{-\bar{X}} = \mathrm{ENW}(\mathrm{e}^{-\theta}).$$

Mamy

$$\sqrt{n}\big(\hat{g}_1 - g(\theta)\big) \to_{\mathrm{d}} \mathrm{N}(0, \sigma_1^2), \quad \text{gdzie } \sigma_1^2 = \mathrm{Var}_\theta \mathbb{I}(X_1 = 0) = \mathrm{e}^{-\theta}(1 - \mathrm{e}^{-\theta});$$
$$\sqrt{n}\big(\hat{g}_2 - g(\theta)\big) \to_{\mathrm{d}} \mathrm{N}(0, \sigma_2^2), \quad \text{gdzie } \sigma_2^2 = \big(g'(\theta)\big)^2 \mathrm{Var}_\theta X_1 = \mathrm{e}^{-2\theta}\theta.$$

Asymptotyczna normalność $\hat{g}_1$ wynika wprost z CTG dla schematu Bernoulliego (za sukces uważamy obserwację równą zeru). Do $\hat{g}_2$ możemy zastosować Lemat 4.2.3. Estymator największej wiarogodności $\hat{g}_2$ jest asymptotycznie efektywny, as.ef$(\hat{g}_2) = 1$. Względna efektywność asymptotyczna jest dana wzorem

$$\text{as.ef}(\hat{g}_1, \hat{g}_2) = \frac{\sigma_2^2}{\sigma_1^2} = \frac{\theta \mathrm{e}^{-2\theta}}{\mathrm{e}^{-\theta}(1 - \mathrm{e}^{-\theta})} = \frac{\theta}{\mathrm{e}^{\theta} - 1}$$

i oczywiście as.ef$(\hat{g}_1) =$ as.ef$(\hat{g}_1, \hat{g}_2) < 1$. Dla małych θ (czyli dla małej częstości szkód) efektywność obu estymatorów jest zbliżona: $\theta/(\mathrm{e}^{\theta} - 1) \to 1$ przy $\theta \to 0$. Dla dużych θ estymator $\hat{g}_2$ jest *wyraźnie* lepszy niż $\hat{g}_1$.

Nie należy pochopnie dyskwalifikować estymatorów mniej efektywnych. Mogą odznaczać się innymi zaletami. W rozpatrywanym modelu estymator $\hat{g}_2$ jest bardziej efektywny niż $\hat{g}_1$. Wyobraźmy sobie jednak, że wysokość roszczeń ma rozkład inny, niż rozkład Poissona. Estymator $\hat{g}_1$ będzie w dalszym ciągu *zgodnym* estymatorem $\mathbb{P}_\theta(X_1 = 0)$, bo nie wykorzystuje założenia o postaci gęstości. Natomiast $\hat{g}_2$ faktycznie estymuje wielkość $\exp(-\mathbb{E}_\theta X_1)$, która jest na ogół różna od $\mathbb{P}_\theta(X_1 = 0)$, jeśli rozkład nie jest rozkładem Poissona.

4.3.5 Stwierdzenie (Asymptotyczna normalność kwantyli próbkowych). *Niech $X_1, X_n, \ldots$ będzie próbką z rozkładu ciągłego o gęstości $f(x)$. Załóżmy, że ξ jest kwantylem rzędu p tego rozkładu, gęstość jest ciągła w ξ i $f(\xi) > 0$. Jeśli $\hat{\xi}_n$ jest kwantylem z próbki to*

$$\sqrt{n}\big(\hat{\xi}_n - \xi\big) \to_{\mathrm{d}} \mathrm{N}\left(0, \frac{p(1-p)}{f(\xi)^2}\right), \qquad (n \to \infty).$$

Przypomnijmy, że oznaczamy medianę symbolem med, a medianę z próbki symbolem $\hat{\mathrm{med}}$.

4.3.6 PRZYKŁAD (Średnia czy mediana z próbki?). Dla symetrycznych rozkładów prawdopodobieństwa wartość oczekiwana pokrywa się z medianą. Powstaje pytanie, czy wybrać średnią z próbki, czy medianę z próbki, do estymacji „środka rozkładu". Odpowiedź zależy od rodziny rozkładów, z którą mamy do czynienia.

(i) *Rozkłady normalne.* Rozważmy próbkę z rozkładu $\mathrm{N}(\mu, \sigma^2)$. Tutaj μ jest wartością oczekiwaną i jednocześnie medianą. Mamy

$$\sqrt{n}\big(\bar{X} - \mu\big) \to_{\mathrm{d}} \mathrm{N}\left(0, \sigma^2\right), \qquad \sqrt{n}\big(\hat{\mathrm{med}} - \mu\big) \to_{\mathrm{d}} \mathrm{N}\left(0, \frac{\pi\sigma^2}{2}\right),$$

zgodnie ze Stwierdzeniem 4.3.5, bo gęstość rozkładu $\mathrm{N}(\mu, \sigma^2)$ w punkcie μ jest równa $(2\pi)^{-1/2}\sigma^{-1}$. Średnia $\bar{X}$ jest ENW i jest estymatorem efektywnym, $\mathrm{as.ef}\bar{X} = 1$. Efektywność mediany jest mniejsza:

$$\mathrm{as.ef}(\hat{\mathrm{med}}) = \frac{2}{\pi} < 1.$$

(ii) *Rozkłady Laplace'a.* Rozważmy próbkę prostą z rozkładu $\mathrm{Lapl}(\mu, \lambda)$. Jest to rozkład symetryczny, o wartości oczekiwanej i medianie μ. Nietrudno przekonać się, że wariancja pojedynczej obserwacji jest równa $2/\lambda^2$. Mamy teraz

$$\sqrt{n}\big(\bar{X} - \mu\big) \to_{\mathrm{d}} \mathrm{N}\left(0, \frac{2}{\lambda^2}\right), \qquad \sqrt{n}\big(\hat{\mathrm{med}} - \mu\big) \to_{\mathrm{d}} \mathrm{N}\left(0, \frac{1}{\lambda^2}\right).$$

Wynika to, odpowiednio, z CTG i ze Stwierdzenia 4.3.5. Dla rozkładów Laplace'a $\hat{\text{med}} = \text{ENW}(\mu)$ i $\text{as.ef}(\hat{\text{med}}) = 1$ (Zadanie **??**). Średnia jest mniej efektywnym estymatorem μ niż mediana próbkowa:

$$\text{as.ef}(\bar{X}) = 1/2.$$

(iii) *Rozkłady Cauchy'ego.* Dla próbki z rozkładu $\text{Cauchy}(a, d)$, mediana z próbki jest zgodnym i asymptotycznie normalnym estymatorem „środka" rozkładu, czyli a. Stwierdzenie 4.3.5 daje

$$\sqrt{n}\big(\hat{\text{med}} - a\big) \to \text{N}\left(0, \frac{4\pi^2}{d^2}\right).$$

Zadanie **??** pokazuje, że średnia z próbki ma taki sam rozkład jak pojedyncza obserwacja: $\bar{X} \sim \text{Cauchy}(a, d)$. Zatem $\bar{X}$ nie jest zgodnym estymatorem a. Oczywiście, związane jest to z faktem, że rozkład Cauchy'ego nie ma wartości oczekiwanej ale ma medianę a.

4.4 Zadania

1. W modelu Hardy'ego-Weinberga, obliczyć asymptotyczną efektywność estymatorów z Przykładu 3.1.2.

2. Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu $\text{U}(0, \theta)$, z nieznanym parametrem $\theta > 0$, Przykład 3.2.8. Skonstruować estymator nieobciążony parametru θ postaci $\hat{\theta}_n = c_n \max(X_1, \ldots, X_n)$ z odpowiednio dobraną stałą c_n.

   (a) Obliczyć $\text{Var}_\theta \hat{\theta}_n$ i $\lim_{n\to\infty} n^2 \text{Var}_\theta \hat{\theta}_n$.

   (b) Obliczyć $\text{Var}_\theta \tilde{\theta}_n$ i $\lim_{n\to\infty} n \text{Var}_\theta \tilde{\theta}_n$ dla innego nieobciążonego estymatora, $\tilde{\theta}_n = 2\bar{X}_n$.

   (c) Podać rozkłady graniczne $n(\hat{\theta}_n - \theta)$ i $\sqrt{n}(\tilde{\theta}_n - \theta)$. Porównaj z Zadaniem 6 w Rozdziale 1.

Rozdział 5

Przedziały ufności

Pojęcie przedziału ufności precyzuje ideę estymacji z określoną dokładnością. Zamiast pojedynczego oszacowania nieznanego parametru, podajemy górną i dolną granicę oszacowania. Nie możemy co prawda zagwarantować, że parametr leży na pewno między tymi granicami. Możemy wymagać, by tak było z odpowiednio dużym prawdopodobieństwem.

5.0.1 DEFINICJA. *Niech $g(\theta)$ będzie funkcją nieznanego parametru. Rozważmy dwie statystyki $\underline{g} = \underline{g}(X)$ i $\overline{g} = \overline{g}(X)$. Mówimy, że $[\underline{g}, \overline{g}]$ jest* ***przedziałem ufności*** *dla $g(\theta)$ na poziomie ufności $1-\alpha$, jeśli*

$$\mathbb{P}_\theta \left(\underline{g}(X) \leqslant g(\theta) \leqslant \overline{g}(X) \right) \geqslant 1 - \alpha$$

dla każdego θ.

Typowo, α jest małą liczbą, na przykład $1-\alpha = 0.95$ lub $1-\alpha = 0.99$. Zauważmy, że warunek $\mathbb{P}_\theta(g(\theta) \in [\underline{g}, \overline{g}]) = 1-\alpha$ należy rozumieć tak: „losowy przedział $[\underline{g}, \overline{g}]$ pokrywa nieznaną liczbę $g(\theta)$ z dużym prawdopodobieństwem". Jeśli obliczone z próbki wartości statystyk są równe, powiedzmy, $\underline{g} = 5$ i $\overline{g} = 8$, to *nie ma sensu* sformułowanie "wielkość $g(\theta)$ należy do przedziału $[5, 8]$ z prawdopodobieństwem $1-\alpha$"! Doświadczenie losowe już zostało wykonane i zakończyło się albo „sukcesem" (to znaczy zaszło zdarzenie losowe $g(\theta) \in [\underline{g}, \overline{g}]$) lub „porażką" ($g(\theta) \notin [\underline{g}, \overline{g}]$). Osobliwość sytuacji polega na tym, że *nie wiemy*, która z tych ewentualności ma miejsce.

5.1 Przykłady

Rozpatrzymy typowe i ważne przykłady konstrukcji przedziałów ufności w modelu normalnym.

5.1.1 PRZYKŁAD (Przedział ufności dla średniej w modelu normalnym). Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu $\mathrm{N}(\mu, \sigma^2)$.

(i) *Znana wariancja.* Zakładamy, że znamy σ^2 i tylko μ jest nieznanym parametrem. Ponieważ

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1),$$

więc jeśli z jest kwantylem rzędu $1-\alpha/2$ standardowego rozkładu normalnego $\mathrm{N}(0,1)$, to mamy $\mathbb{P}_\mu(|\sqrt{n}(\bar{X}-\mu)/\sigma| \leqslant z) = \Phi(z) - \Phi(-z) = 1-\alpha$. Innymi słowy,

$$\mathbb{P}_\mu\left(\bar{X} - \frac{\sigma z}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + \frac{\sigma z}{\sqrt{n}}\right) = 1-\alpha,$$

Przedział ufności dla μ ma zatem postać $[\bar{X} - \sigma z/\sqrt{n}, \bar{X} + \sigma z/\sqrt{n}]$, gdzie $z = \mathrm{z}_{1-\alpha/2}$.

(ii) *Nieznana wariancja.* Założymy teraz, że wariancja σ^2 rozkładu normalnego jest nieznana. Interesuje nas przedział ufności dla μ. Niech $S^2 = \sum(X_i - \bar{X})^2/(n-1)$ będzie zwykle stosowanym nieobciążonym estymatorem wariancji i $S = \sqrt{S^2}$. Wiemy, że

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \mathrm{t}(n-1),$$

gdzie $\mathrm{t}(n-1)$ jest rozkładem t-Studenta z $n-1$ stopniami swobody (zobacz Wniosek 2.2.6 w Rozdziale 2). Jeśli więc t jest kwantylem rzędu $1-\alpha/2$ to $\mathbb{P}_{\mu,\sigma}(|\sqrt{n}(\bar{X}-\mu)/S| \leqslant t) = 1-\alpha$. Innymi słowy,

$$\mathbb{P}_{\mu,\sigma}\left(\bar{X} - \frac{St}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + \frac{St}{\sqrt{n}}\right) = 1-\alpha.$$

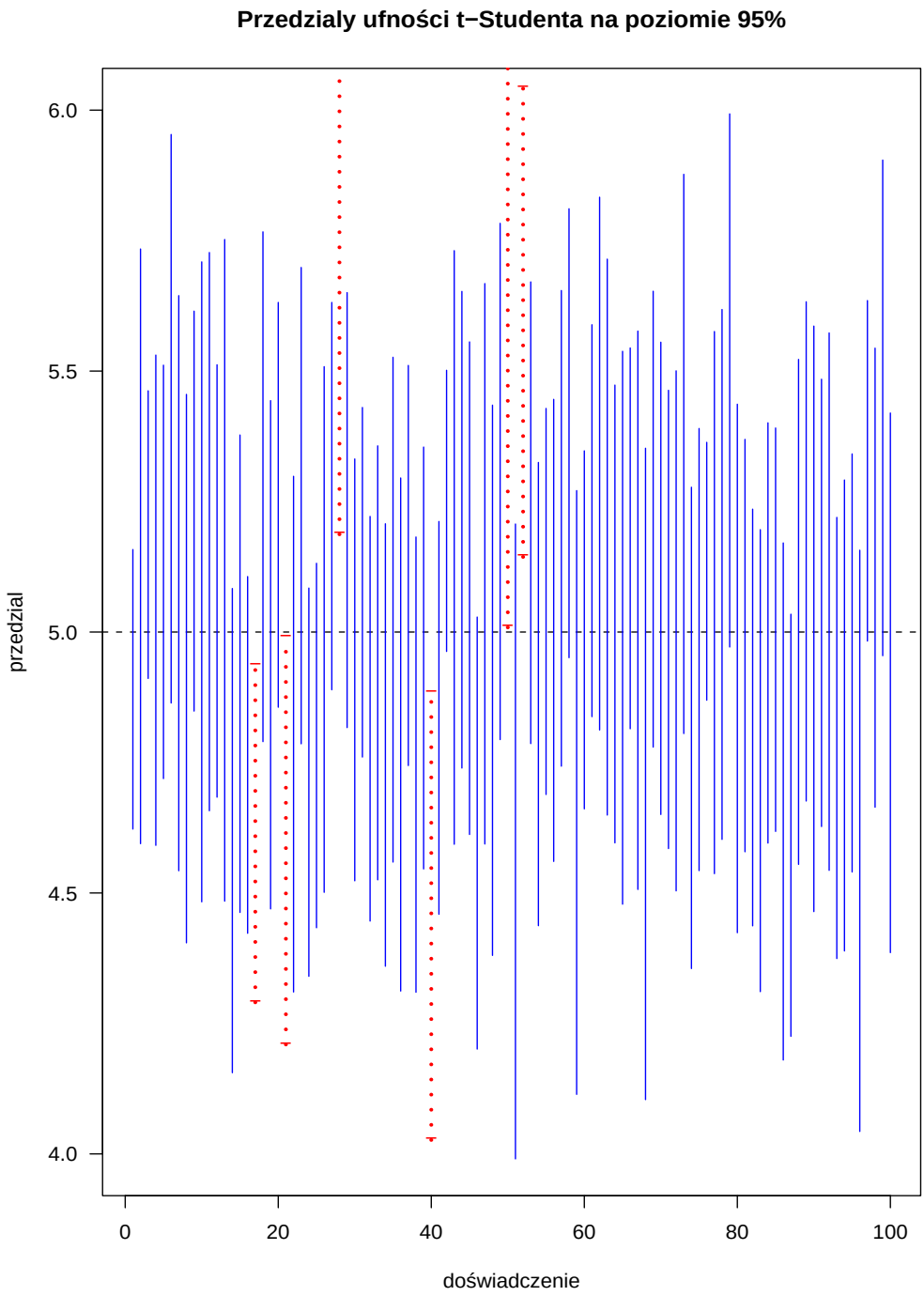
Przedział ufności ma postać $[\bar{X} - St/\sqrt{n}, \bar{X} + St/\sqrt{n}]$, gdzie $t = \mathrm{t}_{1-\alpha/2}(n-1)$.

Przedstawimy takie przedziały dla 20 symulowanych próbek.

nr. doświadczenia	przedział	$\ni \mu$?
1	[4.62, 5.16]	0
2	[4.59, 5.73]	0
3	[4.91, 5.46]	0
4	[4.59, 5.53]	0
5	[4.72, 5.51]	0
6	[4.86, 5.95]	0
7	[4.54, 5.64]	0
8	[4.40, 5.46]	0
9	[4.85, 5.61]	0
10	[4.48, 5.71]	0
11	[4.66, 5.73]	0
12	[4.68, 5.51]	0
13	[4.48, 5.75]	0
14	[4.16, 5.08]	0
15	[4.46, 5.38]	0
16	[4.42, 5.11]	0
17	[4.29, 4.94]	**1**
18	[4.79, 5.77]	0
19	[4.47, 5.44]	0
20	[4.86, 5.63]	0

Próbki rozmiaru $n = 20$ były generowane z rozkładu $\mathrm{N}(\mu, \sigma^2)$. Prawdziwe wartości parametrów użyte w symulacjach to $\mu = 5$ i $\sigma = 1$. Na 20 doświadczeń pokazanych w tabelce, raz zdarzyło się (dla 17-tej próbki), że przedział nie zawierał μ.

Rysunek przedstawia 100 doświadczeń, wśród których zdarzyło się sześć „porażek". Są one wyróżnione przerywaną linią i kolorem czerwonym.


Przedzialy ufności t−Studenta na poziomie 95%
przedział
6.0
5.5
5.0
4.5
4.0
0
20
40
60
80
100
doświadczenie

5.1.2 PRZYKŁAD (Przedział ufności dla wariancji). Rozważamy ten sam model, co poprzednio. Mamy próbkę $X_1, \ldots, X_n$ z rozkładu $\mathrm{N}(\mu, \sigma^2)$ z nieznanymi parametrami μ i σ^2. Zajmiemy się teraz konstrukcją przedziału ufności dla σ^2. Wiemy, że

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

(Twierdzenie Fishera, 2.2.4). Niech c_1 i c_2 będą kwantylami rzędu, odpowiednio, $\alpha/2$ i $1-\alpha/2$ rozkładu chi-kwadrat z $n-1$ stopniami swobody. Mamy $\mathbb{P}_{\mu,\sigma}(c_1 \leqslant (n-1)S^2/\sigma^2 \leqslant c_2) = 1-\alpha$. Inaczej,

$$\mathbb{P}_{\mu,\sigma}\left(\frac{(n-1)S^2}{c_2} \leqslant \sigma^2 \leqslant \frac{(n-1)S^2}{c_1}\right) = 1-\alpha.$$

Przedziałem ufności jest $[(n-1)S^2/c_2, (n-1)S^2/c_1]$, gdzie $c_1 = \chi^2_{\alpha/2}(n-1)$ i $c_2 = \chi^2_{1-\alpha/2}(n-1)$.

5.1.3 PRZYKŁAD (Przedział ufności dla ilorazu wariancji). Rozważmy dwie niezależne próbki $X_1, \ldots, X_k \sim \mathrm{N}(\mu_X, \sigma_X^2)$ i $Y_1, \ldots, Y_m \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$. Nieznanymi parametrami są μ_X, σ_X^2, μ_Y i σ_Y^2. Chcemy zbudować przedział ufności dla σ_Y^2/σ_X^2. Wyobraźmy sobie, że X_i i Y_j są wynikami pomiarów dokonanych przy użyciu dwóch przyrządów. Stosunek wariancji pozwala na porównanie dokładności obu przyrządów (przy tej interpretacji naturalne jest założenie, że $\mu_X = \mu_Y$; nie jest ono jednak potrzebne w dalszych rozważaniach). Wiemy, że

$$\frac{S_X^2\sigma_Y^2}{S_Y^2\sigma_X^2} \sim \mathrm{F}(k-1, m-1)$$

(zobacz Przykład 2.2.7). W powyższym wzorze, S_X^2 i S_Y^2 oznaczają standardowe nieobciążone estymatory wariancji obliczone, odpowiednio, dla próbki X-ów i Y-ów. Niech f_1 i f_2 będą kwantylami rzędu, odpowiednio, $\alpha/2$ i $1-\alpha/2$ rozkładu F-Snedecora z $k-1$ stopniami swobody licznika i $m-1$ stopniami swobody mianownika. Mamy $\mathbb{P}(f_1 \leqslant S_X^2\sigma_Y^2/(S_Y^2\sigma_X^2) \leqslant f_2) = 1-\alpha$. Inaczej,

$$\mathbb{P}\left(f_1\frac{S_Y^2}{S_X^2} \leqslant \frac{\sigma_Y^2}{\sigma_X^2} \leqslant f_2\frac{S_Y^2}{S_X^2}\right) = 1-\alpha.$$

Rzecz jasna, $\mathbb{P}$ oznacza tu rozkład prawdopodobieństwa zależny od wszystkich parametrów modelu, a więc $\mathbb{P}_{\mu_X,\mu_Y,\sigma_X,\sigma_Y}$. Przedziałem ufności jest $[f_1S_Y^2/S_X^2, f_2S_Y^2/S_X^2]$, gdzie $f_1 = \mathrm{F}_{\alpha/2}(k-1,m-1)$ i $f_2 = \mathrm{F}_{1-\alpha/2}(k-1,m-1)$.

5.2 Asymptotyczne przedziały ufności

Czasami wyznaczenie dokładnego przedziału ufności jest trudne i musimy się zadowolić rozwiązaniami przybliżonymi. Przybliżenia, o których będziemy mówić są uzasadnione w przypadku dużej liczności próbki.

5.2.1 DEFINICJA. *Rozważmy nieskończony ciąg obserwacji $X_1, X_2, \ldots$ i dwie statystyki: $\underline{g}_n = \underline{g}(X_1, \ldots, X_n)$ i $\overline{g}_n = \overline{g}(X_1, \ldots, X_n)$. Powiemy, że $[\underline{g}_n, \overline{g}_n]$ jest* ***asymptotycznym przedziałem ufności*** *dla $g(\theta)$ na poziomie ufności $1-\alpha$, jeśli*

$$\lim_{n\to\infty} \mathbb{P}_\theta \left(\underline{g}(X_1, \ldots, X_n) \leqslant g(\theta) \leqslant \overline{g}(X_1, \ldots, X_n) \right) \geqslant 1 - \alpha$$

dla każdego θ.

W praktyce, jeśli n jest „odpowiednio duże", oczekujemy, że warunek $\mathbb{P}_\theta(g(\theta) \in [\underline{g}_n, \overline{g}_n] \geqslant 1-\alpha$ jest *w przybliżeniu* spełniony.

5.2.2 PRZYKŁAD (Przedział ufności dla prawdopodobieństwa sukcesu w schemacie Bernoulliego). Niech $S_n \sim \mathrm{Bin}(n, \theta)$. Jeśli liczba prób n jest duża, S_n ma *w przybliżeniu* rozkład normalny $\mathrm{N}(n\theta, n\theta(1-\theta))$. Innymi słowy,

$$\mathbb{P}_\theta \left(\frac{\sqrt{n}\,|S_n/n - \theta|}{\sqrt{\theta(1-\theta)}} \leqslant z \right) \simeq \Phi(z) - \Phi(-z).$$

Weźmy za z kwantyl rzędu $1-\alpha/2$ standardowego rozkładu normalnego: $\Phi(z) - \Phi(-z) = 1-\alpha$. Możemy teraz ostatni wzór przepisać w postaci

$$\mathbb{P}_\theta \Big(z^2\theta(1-\theta) \geqslant n\,(S_n/n - \theta)^2 \Big) \simeq 1 - \alpha.$$

Nierówność w nawiasie jest kwadratowa względem θ i możemy ją rozwiązać w standardowy sposób. Wnioskujemy, że

$$\left[\frac{S_n + \dfrac{z^2}{n} - z\sqrt{\dfrac{S_n(n-S_n)}{n} + \dfrac{z^2}{4}}}{n+z^2}, \frac{S_n + \dfrac{z^2}{n} + z\sqrt{\dfrac{S_n(n-S_n)}{n} + \dfrac{z^2}{4}}}{n+z^2} \right]$$

jest przybliżonym (asymptotycznym) przedziałem ufności dla θ.

Nieparametryczne przedziały ufności dla kwantyli

Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu o dystrybuancie F. Interesować nas będzie estymacja kwantyla rzędu p tego rozkładu. Załóżmy, że dystrybuanta F jest ciągła i ściśle rosnąca w pewnym otoczeniu punktu ξ_p i $F(\xi_p) = p$. Niech $X_{1:n} \leqslant \cdots \leqslant X_{n:n}$ będą statystykami pozycyjnymi. Spróbujemy zbudować dla ξ_p przedział ufności postaci $[X_{\underline{k}:n}, X_{\overline{k}:n}]$, gdzie $0 \leqslant \underline{k} \leqslant \overline{k} \leqslant n$. Zauważmy, że

$$\mathbb{P}_F\big(X_{\underline{k}:n} \leqslant \xi_p \leqslant X_{\overline{k}:n}\big) = \sum_{i=\underline{k}}^{\overline{k}-1} \binom{n}{i} p^i (1-p)^{n-i}. \qquad (*)$$

Symbol $\mathbb{P}_F$ przypomina, że $X_1, \ldots, X_n$ jest próbką z rozkładu o dystrybuancie F. Aby uzasadnić równość $(*)$ rozważmy schemat Bernoulliego, w którym i-te doświadczenie kończy się „sukcesem", jeśli $X_i \leqslant \xi_p$ lub „porażką", jeśli $X_i > \xi_p$. Prawdopodobieństwo we wzorze $(*)$ jest równe

$$\mathbb{P}_F(\underline{k} \leqslant L_n < \overline{k}), \quad \text{gdzie } L_n = \sum_{i=1}^{n} \mathbb{I}(X_i \leqslant \xi_p).$$

Wzór $(*)$ można wykorzystać do zbudowania przedziału ufności na zadanym poziomie $1-\alpha$. Trzeba po prostu tak dobrać $\underline{k}$ i $\overline{k}$, żeby prawa strona była co najmniej równa $1-\alpha$. Zależność $\underline{k}$ i $\overline{k}$ od n jest, niestety, dość skomplikowana. Jeśli n jest duże, to można posłużyć się przybliżeniem normalnym. Wybierzmy dwa ciągi $\underline{k}_n$ i $\overline{k}_n$ w taki sposób, żeby

$$\lim_{n\to\infty} \frac{p - \underline{k}_n/n}{\sqrt{p(1-p)/n}} = z, \qquad \lim_{n\to\infty} \frac{\overline{k}_n/n - p}{\sqrt{p(1-p)/n}} = z, \qquad (**)$$

gdzie $z = \mathrm{z}_{1-\alpha/2}$. W praktyce można przyjąć na przykład

$$\underline{k} = \underline{k}_n = [np - z\sqrt{np(1-p)}];$$
$$\overline{k} = \overline{k}_n = [np + z\sqrt{np(1-p)}],$$

gdzie symbol „$[a]$" oznacza zaokrąglenie liczby a do najbliższej liczby całkowitej. Jeśli zachodzi warunek $(**)$, to

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}_F\big(X_{\underline{k}:n} \leqslant \xi_p \leqslant X_{\overline{k}:n}\big) \\
=&\lim_{n\to\infty} \mathbb{P}_F(\underline{k}_n \leqslant L_n \leqslant \overline{k}_n) \\
=&\lim_{n\to\infty} \mathbb{P}_F\left(\frac{\underline{k}_n/n - p}{\sqrt{p(1-p)/n}} \leqslant \frac{L_n - np}{\sqrt{np(1-p)}} \leqslant \frac{\overline{k}_n/n - p}{\sqrt{p(1-p)/n}}\right) \\
=&\lim_{n\to\infty} \mathbb{P}_F\left(-z \leqslant \frac{L_n - np}{\sqrt{np(1-p)}} \leqslant z\right) \\
=&\Phi(z) - \Phi(-z) = 1 - \alpha,
\end{aligned}$$

na mocy twierdzenia de Moivre'a-Laplace'a. Skonstruowaliśmy asymptotyczny przedział ufności $[X_{\underline{k}:n}, X_{\overline{k}:n}]$ dla kwantyla ξ_p. Zwróćmy uwagę, że zakładaliśmy bardzo mało o nieznanej dystrybuancie F. Nasze rozważania miały charakter "nieparametryczny", to znaczy nie wymagaliśmy, by dystrybuanta była dana wzorem o jakiejś szczególnej postaci.

5.2.3 PRZYKŁAD. Znajdziemy przedział ufności na poziomie 0.9 dla mediany. Z tablic rozkładu normalnego odczytujemy kwantyl $z = \mathrm{z}_{0.95} = 1.645$ i kładziemy

$$\underline{k} \simeq \frac{n}{2} - \frac{1.645}{2}\sqrt{n}, \qquad \overline{k} \simeq \frac{n}{2} + \frac{1.645}{2}\sqrt{n}.$$

Jeśli, powiedzmy, $n = 400$ to możemy przyjąć $\underline{k} = 183$ (zaokrąglenie liczby $200 - 16.45$) i $\overline{k} = 217$. Zatem,

$$\mathbb{P}(X_{183:400} \leqslant \mathrm{med}X \leqslant X_{217:400}) \simeq 0.9$$

dla próbki z rozkładu zmiennej losowej X.

Przedziały ufności i metoda największej wiarogodności

Jeśli $\hat{\theta} = \mathrm{ENW}(\theta)$ i spełnione są założenia Twierdzenia 4.3.1, to zgodnie z Wnioskiem 4.3.2,

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \to_{\mathrm{d}} \mathrm{N}\left(0, \frac{(g'(\theta))^2}{I_1(\theta)}\right),$$

dla dowolnej różniczkowalnej funkcji $g(\theta)$. Mamy więc

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(\frac{|g(\hat{\theta}_n) - g(\theta)|\sqrt{nI_1(\theta)}}{|g'(\theta)|} \leqslant z\right) = \Phi(z) - \Phi(-z).$$

Dla $z = \mathrm{z}_{1-\alpha/2}$, prawa strona staje się równa $1-\alpha$. Z Lematu Słuckiego wynika, że powyższy wzór pozostanie słuszny, gdy zastąpimy po lewej stronie $I_1(\theta)$ i $g'(\theta)$ przez $I_1(\hat{\theta}_n)$ i $g'(\hat{\theta}_n)$. Wynika to ze zgodności ENW: $\hat{\theta} \to_{\mathrm{P}} \theta$, przy założeniu ciągłości funkcji I_1 i g'. Podsumowując, stwierdzamy, że (przy pewnych założeniach)

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(g(\hat{\theta}_n) - \frac{z|g'(\hat{\theta}_n)|}{\sqrt{nI_1(\hat{\theta}_n)}} \leqslant g(\theta) \leqslant g(\hat{\theta}_n) + \frac{z|g'(\hat{\theta}_n)|}{\sqrt{nI_1(\hat{\theta}_n)}}\right) = 1-\alpha.$$

W ten sposób otrzymaliśmy bardzo ogólny sposób konstrukcji asymptotycznych przedziałów ufności. Niestety, w praktyce ten sposób nie zawsze jest zadowalający, bo wynikające z asymptotycznych rozważań przybliżenie może okazać się dostatecznie dobre dopiero dla bardzo dużych próbek.

Stabilizacja wariancji jest sposobem budowania asymptotycznych przedziałów ufności o lepszych własnościach (to jest dających zazwyczaj lepsze przybliżenia dla umiarkowanych liczności próbek). Metoda opisana poprzednio wymagała zastąpienia nieznanego parametru θ przez estymator $\hat{\theta}_n$ w wyrażeniu na asymptotyczną wariancję ENW. Dla *odpowiednio dobranej* funkcji $g(\theta)$ można tego uniknąć. Jeśli funkcja $g(\theta)$ spełnia równanie różniczkowe

$$g'(\theta)/\sqrt{I_1(\theta)} = c = \mathrm{const},$$

to asymptotyczna wariancja $(g'(\theta))^2/I_1(\theta) = c^2$ nie zależy od θ. Dla takiej funkcji mamy po prostu

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(g(\hat{\theta}) - \frac{zc}{\sqrt{n}} \leqslant g(\theta) \leqslant g(\hat{\theta}) - \frac{zc}{\sqrt{n}}\right) = 1-\alpha,$$

dla $z = \mathrm{z}_{1-\alpha/2}$. Jeśli $g(\theta)$ jest funkcją rosnącą, to przedział ufności dla $g(\theta)$ możemy bez trudu „przerobić" na przedział ufności dla samego parametru θ (i odwrotnie). Pokażemy to na przykładzie.

5.2.4 PRZYKŁAD (Przedziały ufności dla rozkładu Poissona). Zastosujmy metodę stabilizacji wariancji, biorąc za wyjściowy estymator $\mathrm{ENW}(\theta) = \bar{X}_n$, dla próbki z $\mathrm{Poiss}(\theta)$. Ponieważ $I_1(\theta) = 1/\theta$, więc funkcję $g(\theta)$ dobieramy tak, żeby $g'(\theta) = c/\sqrt{\theta}$. Stąd otrzymujemy $g(\theta) = \sqrt{\theta}$, przyjmując dla wygody $c = 1/2$. Mamy

$$\sqrt{n}\left(g(\hat{\theta}) - g(\theta)\right) \to_{\mathrm{d}} \mathrm{N}\left(0, 1/4\right).$$

Natychmiast wnioskujemy, że

$$\lim_{n\to\infty} \mathbb{P}_\theta \left(\sqrt{\bar{X}_n} - \frac{z}{2\sqrt{n}} \leqslant \sqrt{\theta} \leqslant \sqrt{\bar{X}_n} + \frac{z}{2\sqrt{n}} \right) = 1 - \alpha.$$

Stąd wynika, że asymptotyczny przedział ufności dla θ (a nie dla $\sqrt{\theta}$) można napisać w takiej postaci:

$$\left[\max\left\{ 0, \left(\sqrt{\bar{X}} - \frac{z}{2\sqrt{n}} \right)^2 \right\}, \left(\sqrt{\bar{X}} + \frac{z}{2\sqrt{n}} \right)^2 \right].$$

5.3 Zadania

1. Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu $\mathrm{U}(0, \theta)$, z nieznanym parametrem $\theta > 0$, Przykład 3.2.8. Skonstruować przedział ufności dla parametru θ postaci $[M_n, c_n M_n]$, gdzie $M_n = \max(X_1, \ldots, X_n)$. Dobrać stałą c_n tak, aby prawdopodobieństwo pokrycia było równe $1 - \alpha$.

2. Pokazać, że $\left[\hat{\theta}_n - z\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}/\sqrt{n}, \hat{\theta}_n + z\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}/\sqrt{n}\right]$, gdzie $\hat{\theta}_n = S_n/n$, $S_n \sim \mathrm{Bin}(n, \theta)$, jest asymptotycznym przedziałem ufności dla prawdopodobieństwa sukcesu θ w schemacie Bernoulliego.

3. Skonstruować asymptotyczny przedział ufności dla prawdopodobieństwa sukcesu θ w schemacie Bernoulliego metodą stabilizacji wariancji.

Część III

Testowanie hipotez statystycznych

Rozdział 6

Testy istotności

W tym rozdziale spróbujemy wyjaśnić, na czym polega zagadnienie testowania hipotez statystycznych. Pokażemy, jak konstruuje się tak zwane *testy istotności*. Skoncentrujemy się na kilku ważnych i typowych przykładach. Nasze rozważania będą miały charakter wstępny i heurystyczny. Bardziej systematyczną teorię przedstawimy w podrozdziale następnym.

6.1 Podstawy heurystyczne

Matematyczny model zjawiska losowego sprowadza się do określenia rozkładu prawdopodobieństwa $\mathbb{P}$ obserwacji X na przestrzeni $\mathcal{X}$. Przez *hipotezę statystyczną* rozumiemy, najogólniej mówiąc, pewną wypowiedź na temat rozkładu prawdopodobieństwa „rządzącego" zjawiskiem, które nas interesuje. Zajmiemy się najprostszą sytuacją. Rozważamy na razie jeden, ustalony rozkład $\mathbb{P}$. Pytamy, czy opisuje on poprawnie przebieg doświadczenia losowego. Stawiamy hipotezę: nasz model probabilistyczny jest zgodny z rzeczywistością. Ogólna zasada, wykraczająca daleko poza statystykę, jest taka: powinniśmy odrzucić hipotezę, jeśli rezultaty doświadczenia okażą się sprzeczne z przewidywaniami wynikającymi z modelu. Specyfika modeli probabilistycznych polega jednak na tym, że potrafimy przewidzieć tylko prawdopodobieństwo poszczególnych wyników. Zmodyfikujemy więc ogólną zasadę. Powinniśmy odrzucić hipotezę, *jeśli wynik doświadczenia jest bardzo mało*

prawdopodobny, to znaczy model przewiduje taki wynik z małym prawdopodobieństwem. Innymi słowy, wybieramy pewien zbiór $K \subset \mathcal{X}$ taki, że $\mathbb{P}(X \in K) \leqslant \alpha$, gdzie α jest odpowiednio „małą" liczbą, zwaną *poziomem istotności*. Jeśli w wyniku doświadczenia stwierdzimy, że zaszło zdarzenie „$X \in K$", to powinniśmy zwątpić w poprawność naszego modelu, czyli „odrzucić postawioną przez nas hipotezę statystyczną". Zbiór K nazywamy *obszarem krytycznym* testu, zaś cała procedura nosi nazwę testu *istotności*.

6.1.1 PRZYKŁAD (Czy prawdopodobieństwo „jedynki" jest równe 1/6?). Wykonujemy 300 rzutów kostką do gry. Jeśli kostka jest symetryczna i rzuty wykonujemy rzetelnie, to zmienna losowa $X = N_1$ równa liczbie wyrzuconych „jedynek" ma rozkład dwumianowy $\mathrm{Bin}(300, 1/6)$. Mamy jednak pewne wątpliwości, czy kasyno nie używa obciążonej kostki, w której jedynka ma prawdopodobieństwo *mniejsze* niż 1/6. Jeśli w wyniku doświadczenia ani ani razu nie otrzymamy jedynki, musimy podejrzewać, że coś jest nie w porządku. Może kostka ma w ogóle nie ma ściany oznaczonej jedną kropką? Co prawda, zdarzenie „$X = 0$" w modelu symetrycznej kostki jest *możliwe* ale skrajnie mało prawdopodobne, $\mathbb{P}(X = 0) = (5/6)^{300} \approx 1.76 * 10^{-24}$. Nie jesteśmy skłonni wierzyć w tak wyjątkowy traf. Odrzucamy, bez wielkiego wahania, hipotezę o zgodności modelu. Jak postąpić jeśli otrzymamy, powiedzmy, 33 jedynki, nie jest aż tak jasne. Czy to wynik zgodny z hipotezą o rzetelności kostki, czy też jest podstawą do odrzucenia tej hipotezy?

Rozpatrzymy sprawę bardziej systematycznie. Stawiamy hipotezę

$$\mathrm{H}_0 : X \sim \mathrm{Bin}(300, 1/6).$$

Zdecydujemy się na taki przepis postępowania:

$$\begin{aligned} &\text{jeśli } X < c \text{ to odrzucamy } \mathrm{H}_0; \\ &\text{jeśli } X \geqslant c \text{ to pozostajemy przy } \mathrm{H}_0, \end{aligned}$$

dla pewnej liczby „progowej" c. testu. Próg c ustalimy w następujący sposób. Wybierzmy *poziom istotności*, powiedzmy $\alpha = 0.01$. Uznamy, że zajście zdarzenia o prawdopodobieństwie mniejszym niż α jest dostatecznym powodem do odrzucenia H_0. Powinniśmy więc wybrać c tak, aby $\mathbb{P}(X < c) \leqslant 0.01$. Mamy przy tym na myśli prawdopodobieństwo obliczone przy założeniu, że H_0 jest prawdziwa. Łatwo przekonać się, że $\mathbb{P}(X < 36) = 0.00995 \leqslant 0.01$, a więc ostatecznie nasz test na poziomie istotności 0.01 jest następujący:

Odrzucamy H_0 jeśli $X < 36$.

Obszarem krytycznym naszego testu jest podzbiór $K = \{0, 1, \ldots, 35\}$ przestrzeni obserwacji $\mathcal{X} = \{0, 1, \ldots, 300\}$.

Jeżeli wynikiem doświadczenia są 33 jedynki, to opisana powyżej reguła decyzyjna odrzuca H_0. Inne spojrzenie na ten sam test jest następujące. Obliczamy prawdopodobieństwo otrzymania wyniku takiego jak 33 lub gorszego (za „gorszy" uznajemy zbyt małą liczbę jedynek), przy założeniu H_0. Mówimy, że *p-wartość* (lub *poziom krytyczny testu*) jest w naszym przypadku równa $p = \mathbb{P}(X \leqslant 33) = 0.00379$. Porównujemy p-wartość z założonym poziomem istotności.

Odrzucamy H_0 jeśli $p < 0.01$.

Obie reguły są równoważne, to znaczy prowadzą do takiej samej decyzji. Niemniej obliczanie p-wartości stwarza możliwość (lub pokusę) „mierzenia stopnia przekonania" o nieprawdziwości H_0.

Zwróćmy uwagę, że nasz test „reaguje na odstępstwa od prawdopodobieństwa 1/6 tylko w dół". Mówimy, że jest to test *hipotezy zerowej*

$$H_0 : \text{prawdopodobieństwo jedynki } = 1/6$$

przeciwko hipotezie alternatywnej

$$H_1 : \text{prawdopodobieństwo jedynki } < 1/6.$$

Jeśli rozważymy inną „jednostronną" alternatywę, H_1: prawdopodobieństwo jedynki $> 1/6$, to test powinien odrzucać H_0 jeśli $X > c$, dla odpowiednio dobranego c. Wybór $c = 65$ prowadzi do testu na poziomie istotności $\alpha = 0.01$, bo $\mathbb{P}(X > 65) = 0.00990$. Można również skonstruować test H_0 przeciwko „dwustronnej" alternatywie H_1: prawdopodobieństwo jedynki $\neq 1/6$. Do tej dyskusji jeszcze wrócimy. □

Podsumujmy rozważania z powyższego przykładu. Test hipotezy H_0 na poziomie α spełnia postulat

$$\mathbb{P}(\text{ odrzucamy } H_0) \leqslant \alpha,$$

gdzie prawdopodobieństwo jest obliczone przy założeniu prawdziwości H_0. Konstrukcja testu istotności dopuszcza dużą dowolność. Wybór obszaru krytycznego zależy od tego, „przed jaką ewentualnością chcemy się zabezpieczyć", czyli jaka jest hipoteza alternatywna H_1. Sposób obliczania p-wartości też zależy od wyboru H_1 (bo musimy sprecyzować, jakie wyniki uznajemy za „gorsze", bardziej świadczące przeciw H_0).

Jeśli hipoteza H_0 precyzuje dokładnie jeden rozkład prawdopodobieństwa $\mathbb{P}$, to mówimy, że jest to *hipoteza prosta*. Przypuśćmy, że pewien test H_0 przeciw H_1 używa statystyki $T = T(X)$, mianowicie odrzuca H_0 jeśli $T > c$. Zauważmy, że właściwie każdy test daje się w tej postaci zapisać. Niech $G(t) = \mathbb{P}(T(X) \leqslant t)$ będzie dystrybuantą statystyki T obliczoną przy założeniu, że H_0 jest prawdziwa, czyli $X \sim \mathbb{P}$. Wtedy p-wartość testu jest z definicji obliczana zgodnie ze wzorem

$$p = 1 - G(T(X)). \tag{6.1.2}$$

Używamy tu bardziej rozpowszechnionego terminu *p-wartość* (dosłowne tłumaczenie angielskiego „*p-value*") zamiast poprawnej ale wychodzącej z użycia nazwy *poziom krytyczny*. Następujący fakt wynika bezpośrednio z określenia (6.1.2).

6.1.3 Stwierdzenie. *Jeżeli* H_0 *jest prosta i rozkład statystyki testowej jest ciągły, to przy założeniu prawdziwości* H_0*, p-wartość testu jest zmienną losową o rozkładzie* $\mathrm{U}(0,1)$ *(jednostajnym na przedziale* $[0,1]$*).*

Dowód. Obliczmy dystrybuantę zmiennej losowej danej wzorem (6.1.2). Dla $0 < u < 1$ mamy

$$\begin{aligned}\mathbb{P}(1 - G(T) \leqslant u) &= \mathbb{P}(G(T) > 1 - u) = \mathbb{P}(T > G^{-1}(1-u)) \\ &= 1 - \mathbb{P}(T \leqslant G^{-1}(1-u)) = 1 - G(G^{-1}(1-u)) = u,\end{aligned}$$

przy dodatkowym założeniu, że istnieje funkcja odwrotna G^{-1}. W istocie wystarczy założyć, że G jest ciągła (można dość łatwo zmodyfikować powyższe rozumowanie, ale pomińmy szczegóły). □

6.2 Kilka typowych testów

Przegląd typowych testów istotności zacznę od testów, weryfikujących zgodność z zadanym rozkładem prawdopodobieństwa.

Test proporcji

W sytuacji takiej jak w Przykładzie 6.1.1, powszechnie używa się testów opartych na przybliżeniu rozkładu dwumianowego rozkładem normalnym. Zakładamy, że obserwujemy liczbę sukcesów w schemacie Bernoulli'ego i stawiamy hipotezę

$$\mathrm{H}_0 : \text{prawdopodobieństwo sukcesu } = \theta$$

Z Twierdzenia de Moivre'a - Laplace'a (CTG dla schematu Bernoulli'ego) wynika, że

$$\mathrm{Bin}(n, \theta) \simeq \mathrm{N}(n\theta, n\theta(1-\theta)),$$

przynajmniej dla „dostatecznie dużych n" (niektóre źródła zalecają stosowanie tego przybliżenia, jeśli $n\theta \geqslant 5$ i $n(1-\theta) \geqslant 5$).

Test H_0 przeciwko alternatywie

$$\mathrm{H}_1 : \text{prawdopodobieństwo sukcesu } < \theta$$

na poziomie istotności (w przybliżeniu) α ma postać

$$\text{odrzucamy } \mathrm{H}_0, \text{ jeśli } X - n\theta < -\mathrm{z}_{1-\alpha}\sqrt{n\theta(1-\theta)},$$

gdzie $\mathrm{z}_{1-\alpha}$ jest kwantylem rozkładu $\mathrm{N}(0,1)$, to znaczy $\Phi(\mathrm{z}_{1-\alpha}) = 1-\alpha$. Reguła obliczania p-wartości jest następująca:

$$p \simeq \Phi\left(\frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}\right).$$

Test H_0 przeciwko alternatywie

$$\mathrm{H}_1 : \text{prawdopodobieństwo sukcesu } > \theta$$

ma postać

$$\text{odrzucamy } \mathrm{H}_0, \text{ jeśli } X - n\theta > \mathrm{z}_{1-\alpha}\sqrt{n\theta(1-\theta)}.$$

Reguła obliczania p-wartości jest następująca:

$$p \simeq \Phi\left(-\frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}\right).$$

Test H_0 przeciwko alternatywie

$$\mathrm{H}_1 : \text{prawdopodobieństwo sukcesu } \neq \theta$$

ma postać

$$\text{odrzucamy } \mathrm{H}_0, \text{ jeśli } |X - np| > \mathrm{z}_{1-\alpha/2}\sqrt{np(1-p)} \qquad (*)$$

(zauważmy zamianę $\mathrm{z}_{1-\alpha}$ na $\mathrm{z}_{1-\alpha/2}$). Reguła obliczania p-wartości jest następująca:

$$p \simeq 2\Phi\left(-\frac{|X - n\theta|}{\sqrt{n\theta(1-\theta)}}\right).$$

Ze względu na późniejsze uogólnienia, napiszmy $(*)$ w innej, równoważnej formie:

$$\text{odrzucamy } \mathrm{H}_0, \text{ jeśli } \frac{(X-np)^2}{np(1-p)} > \chi^2_{1-\alpha}(1), \qquad (**)$$

gdzie $\chi^2_{1-\alpha}(1)$ jest kwantylem rozkładu chi-kwadrat z jednym stopniem swobody.

6.2.1 PRZYKŁAD. Wróćmy do Przykładu 6.2.1. Przypuśćmy, że w 300 rzutach kostką otrzymaliśmy 33 jedynek. Przeprowadzamy test H_0: $\theta = 1/6$ przeciw H_1: $\theta < 1/6$.

```
> n=300
> theta=1/6
> X=33
> prop.test(X,n,theta,alternative="less",correct=F)
```

Otrzymujemy rezulatat:

```
data:  X out of n, null probability theta
X-squared = 6.936, df = 1, p-value = 0.004224
alternative hypothesis:  true p is less than 0.1666667
```

Zauważmy, że podana powyżej p-wartość jest równa $\Phi((33-n\theta)/\sqrt{n\theta(1-\theta)})$:

```
> Z=(X-n*theta)/sqrt(n*theta*(1-theta))
> pnorm(Z) # = 0.004223891
```

Dokładna p-wartość (Przykład 6.1.1) jest równa $\sum_{k=0}^{33}\binom{300}{k}(1/6)^k(5/6)^{300-k}$. Można to obliczyć przy pomocy funkcji `pbinom()` (dystrybuanta rozkładu dwumianowego):

```
> pbinom(X,n,p) # = 0.00378605
```

Dla porównania, przeprowadźmy test H_0: $\theta = 1/6$ przeciw H_1: $\theta \neq 1/6$.

```
> n=300
> theta=1/6
> X=33
> prop.test(X,n,theta,alternative="less",correct=F)
```

Otrzymamy wynik:

```
data:  X out of n, null probability theta X-squared = 6.936, df =
1, p-value = 0.008448 alternative hypothesis:  true p is not equal
to 0.1666667
```

Zauważmy, że p-wartość jest teraz dwukrotnie większa.

6.2.2 Uwaga. Aby otrzymać dokładne p-wartości wystarczy zamiast funkcji `prop.test()` zastosować `binom.test()`. Proszę wypróbować. Przybliżony test oparty na CTG ma tę zaletę dydaktyczną, że w sposób naturalny uogólnia się na przypadek doświadczeń o $k > 2$ możliwych wynikach (zamiast dwóch: sukces/porażka). Tym uogólnieniem jest sławny test χ^2.

Test chi-kwadrat

Jest to jeden z klasycznych i najczęściej stosowanych testów statystycznych. Wyobraźmy sobie, że powtarzamy n-krotnie doświadczenie losowe, które ma k możliwych wyników. Wyniki kolejnych doświadczeń są zmiennymi losowymi $X_1, \ldots, X_n$ o wartościach w zbiorze $\mathcal{X} = \{w_1, \ldots, w_k\}$. Przypuśćmy, że te zmienne są niezależne i każda z nich ma ten sam rozkład prawdopodobieństwa dany tabelką

wynik	w_1	$\cdots$	w_i	$\cdots$	w_k
prawdopodobieństwo	p_1	$\cdots$	p_i	$\cdots$	p_k

gdzie $p_i = \mathbb{P}(X_j = w_i)$ (oczywiście, $p_1 + \cdots + p_k = 1$). Rezultaty doświadczeń podsumujmy w postaci „tabelki powtórzeń":

wynik	w_1	$\cdots$	w_i	$\cdots$	w_k
liczba doświadczeń	N_1	$\cdots$	N_i	$\cdots$	N_k

gdzie

$$N_i = \sum_{j=1}^{n} \mathbb{I}(X_j = w_i) = \text{ liczba doświadczeń, których wynikiem jest } w_i$$

(oczywiście, $N_1 + \cdots + N_k = n$). Jeśli nasze przypuszczenia są słuszne, to zmienne losowe $N_1, \ldots, N_k$ mają rozkład wielomianowy $\mathrm{Mult}(n, p_1, \ldots, p_k)$, to znaczy

$$\mathbb{P}(N_1 = n_1, \ldots, N_k = n_k) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}.$$

Najlepiej patrzeć na nasze doświadczenie jak na losowe wrzucanie n „kul" do k „komórek" („wielkość" i-tej komórki jest proporcjonalna do p_i). Test hipotezy

$$\mathrm{H}_0 : (N_1, \ldots, N_k) \sim \mathrm{Mult}(n, p_1, \ldots, p_k)$$

przeprowadzamy w następujący sposób. Obliczamy statystykę „chi-kwadrat":

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}. \tag{6.2.3}$$

Dla wyznaczenia wartości krytycznej testu lub p-wartości, korzystamy z następującego faktu, którego dowód pominiemy.

6.2.4 Stwierdzenie. *Jeżeli* H_0 *jest prawdziwa i* $n \to \infty$*, to rozkład statystyki* χ^2 *danej wzorem* (6.2.3) *zmierza do rozkładu* $\chi^2(k-1)$ *(chi-kwadrat z* $k-1$ *stopniami swobody).*

Test na poziomie istotności α (w przybliżeniu) otrzymamy gdy

$$\text{odrzucamy } \mathrm{H}_0 \text{ jeśli } \chi^2 > c,$$

gdzie $c = \chi^2_{1-\alpha}(k-1)$ jest kwantylem rzędu $1-\alpha$ rozkładu $\chi^2(k-1)$.

6.2.5 PRZYKŁAD (Czy kość jest rzetelna?). Przypuśćmy, że wykonaliśmy 150 rzutów kostką. Wyniki doświadczenia podsumujmy w następującej tabelce:

liczba oczek	1	2	3	4	5	6
liczba rzutów	15	27	36	17	26	29

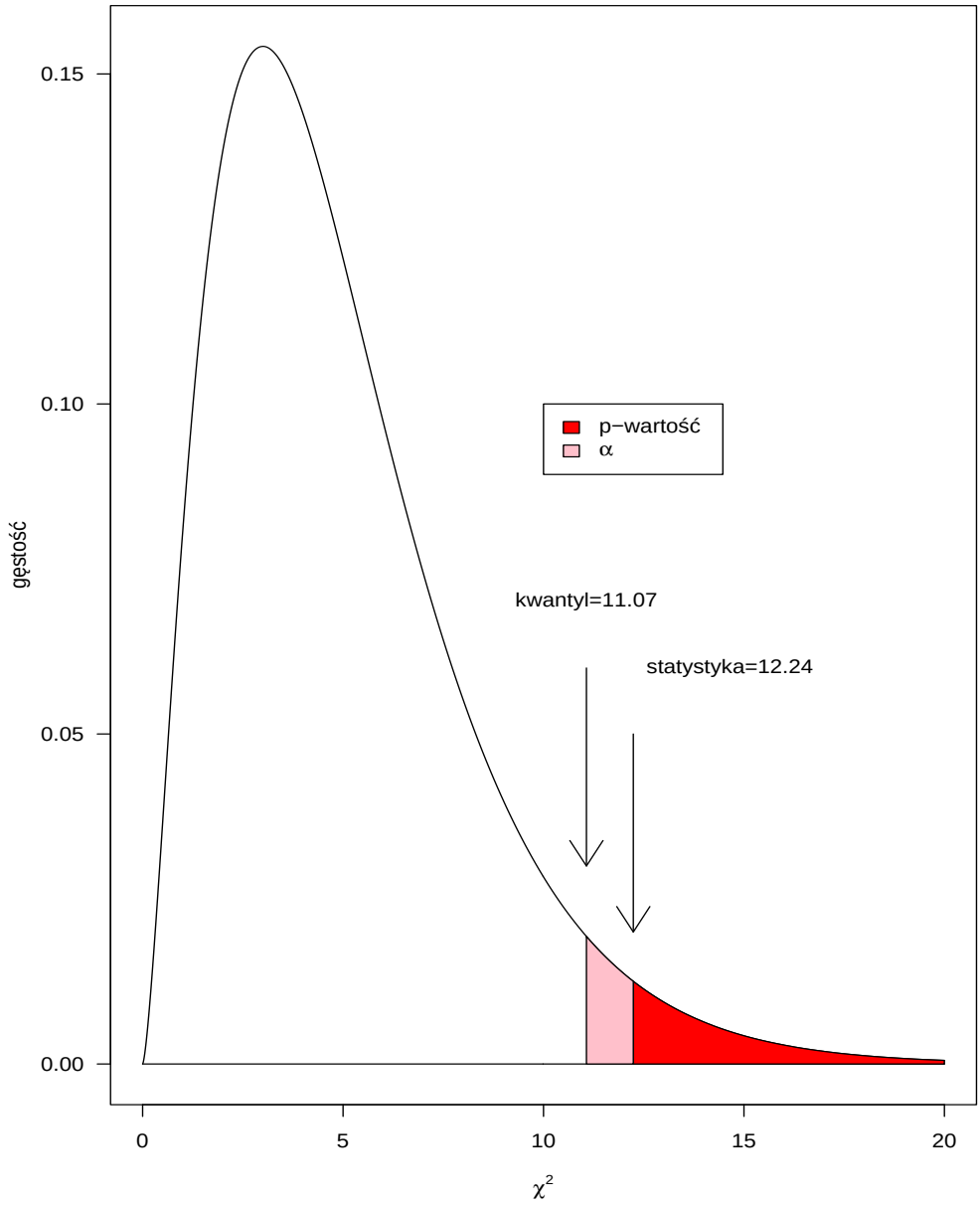
Stawiamy hipotezę, że kość jest rzetelna, czyli

$$\mathrm{H}_0 : (N_1, N_2, N_3, N_4, N_5, N_6) \sim \mathrm{Mult}(150, 1/6, 1/6, 1/6, 1/6, 1/6, 1/6).$$

Wartość statystyki testowej jest równa

$$\chi^2 = \frac{(15-25)^2}{25} + \frac{(27-25)^2}{25} + \frac{(36-25)^2}{25} + \frac{(17-25)^2}{25} + \frac{(26-25)^2}{25} + \frac{(29-25)^2}{25} = 12.24.$$

Przeprowadzimy test tej hipotezy na poziomie istotności 0.05. Najpierw zrobimy to tak, jak w czasach (nie tak dawnych) gdy komputery nie były powszechnie dostępne. Odczytujemy z tablic kwantyl rzędu 0.95 rozkładu chi-kwadrat z $6-1=5$ stopniami swobody: $\chi^2_{0.95}(5) = 11.07$. Ponieważ $12.24 > 11.7$, odrzucamy H_0. Należy przypuszczać, że kość nie jest symetryczna.



Rysunek pokazuje kwantyl $\chi^2_{0.95}(5) = 11.07$ i wartość statystyki testowej $\chi^2 = 12.24$ oraz gęstość rozkładu $\chi^2(5)$.

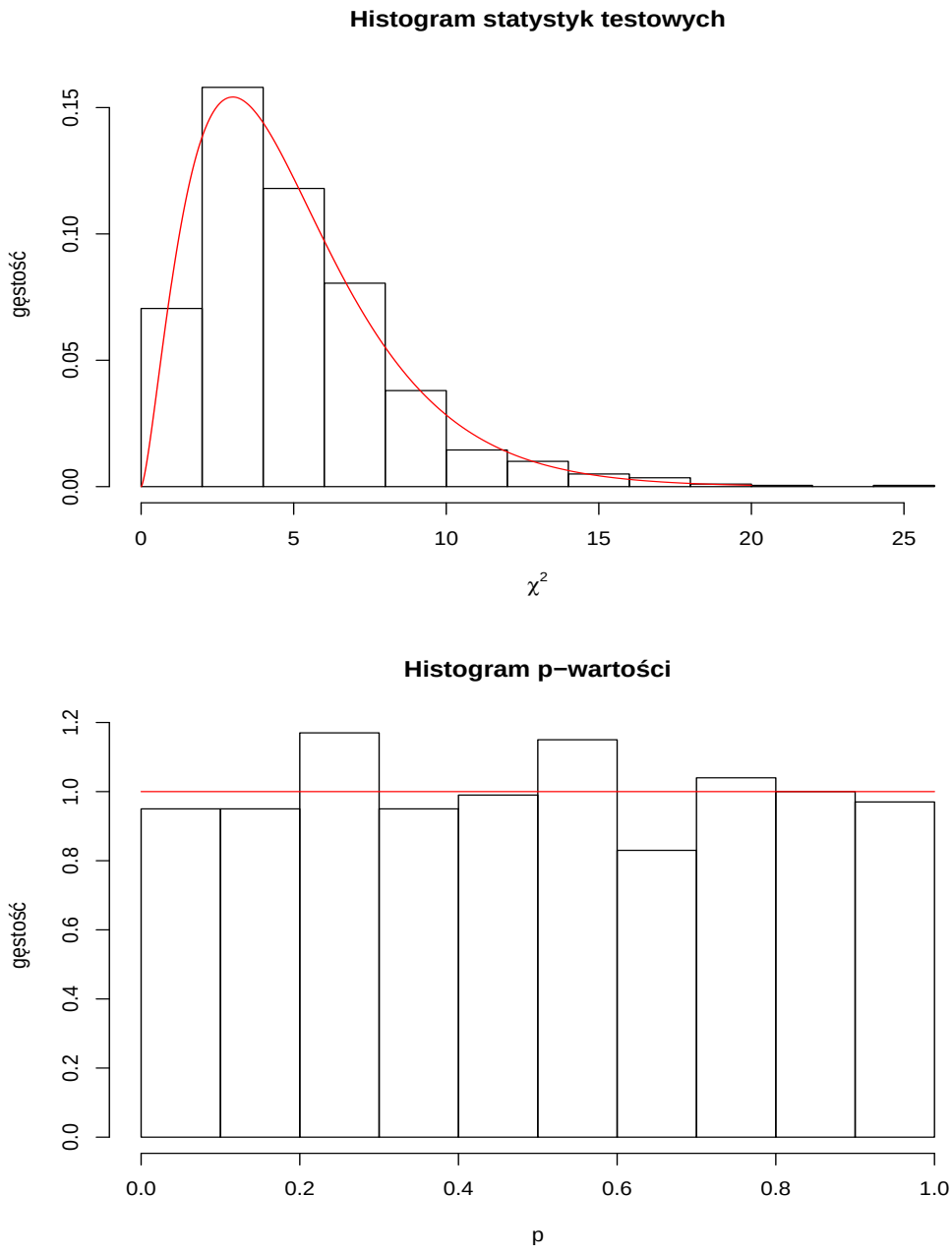
Bardziej współczesny sposób przeprowadzenia testu jest taki:

```
> Ni=c(15,27,36,17,26,29)
> chisq.test(Ni)

Chi-squared test for given probabilities

data:  Ni
X-squared = 12.24, df = 5, p-value = 0.03164
```

6.2.6 PRZYKŁAD (Rozkład statystyki testowej i p-wartości). Korzystając z ułatwienia, jakie daje komputer wyposażony w R, wykonajmy test opisany w poprzednim przykładzie wiele razy. Powtórzmy $m = 1000$ razy serię $n = 150$ rzutów rzetelną kostką (H_0 prawdziwa).

Rysunek pokazuje empiryczny rozkład statystyk testowych w $m = 1000$ powtórzeniach testu zgodności χ^2. Dane pochodziły z rozkładu spełniającego hipotezę zerową H_0. Dolny rysunek pokazuje empiryczny rozkład odpowiednich p-wartości.

Widzimy, że zgodnie ze Stwierdzeniem 6.2.4, wartości statystyki χ^2 mają rozkład zbliżony do $\chi^2(5)$ (gęstość tego rozkładu jest nałożona na histogram). Z kolei p-wartości mają rozkład zbliżony do rozkładu jednostajnego $U(0, 1)$. To jest oczywiście związane ze Stwierdzeniem 6.1.3. Zauważmy przy tym, że założenia Stwierdzenia 6.1.3 są spełnione tylko w przybliżeniu. Statystyka χ^2 jest zmienną dyskretną i tylko w granicy ma rozkład ciągły $\chi^2(5)$. Pomimo tego, zgodność p-wartości z rozkładem jednostajnym jest uderzająca. □



Rozkład wartości statystyki testowej χ^2 i p-wartości testu.

Test chi-kwadrat ma wiele odmian, stosowanych w różnych okolicznościach. Z pewnymi wersjami tego testu spotkamy się jeszcze w tych wykładach. Postać statystyki, którą tu rozpatrzyliśmy najłatwiej zrozumieć i zapamiętać w takiej postaci:

$$\chi^2 = \sum \frac{(\text{wielkość obserwowana} - \text{wielkość oczekiwana})^2}{\text{wielkość oczekiwana}}.$$

Sumowanie rozciągnięte jest na wszystkie „komórki", czyli wartości obserwacji. Oczywiście, w wersji tutaj omawianej „wielkość obserwowana" oznacza N_i, zaś „wielkość oczekiwana" (przy założeniu prawdziwości hipotezy) jest równa np_i.

Dyskretyzacja zmiennych ciągłych. Wspomnijmy jeszcze, że prosty chwyt pozwala stosować test zgodności χ^2 również do obserwacji „typu ciągłego". Rozpatrzmy zmienne losowe $X_1, \ldots, X_n$ i hipotezę

$$\mathrm{H}_0 : X_1, \ldots, X_n \text{ jest próbką z rozkładu o dystrybuancie } F,$$

gdzie F jest ciągła. Wystarczy rozbić zbiór wartości zmiennych losowych na rozłączne „komórki", na przykład wybrać punkty $-\infty = a_0 < a_< \cdots < a_k = \infty$ i zdefiniować

$$N_i = \sum_{j=1}^{n} \mathbb{I}(a_{i-1} < X_j \leqslant a_i), \qquad (i = 1, \ldots, k).$$

Dalej postępujemy tak jak poprzednio, przyjmując $p_i = F(a_i) - F(a_{i-1})$. Taki test dopuszcza dużo dowolności w wyborze „komórek".

Istnieją testy lepiej dostosowane do ciągłego typu danych. Należy do nich test Kołmogorowa-Smirnowa, który później omówimy.

Test chi-kwadrat hipotezy złożonej. Opiszemy modyfikację testu chi-kwadrat w sytuacji, gdy hipoteza zerowa precyzuje rozkład prawdopodobieństwa jedynie z dokładnością do nieznanego parametru. Podobnie jak dla „zwykłego" testu χ^2, rozważamy n niezależnych powtórzeń doświadczenia losowego o k możliwych wynikach. Hipoteza H_0 stwierdza, że rozkład prawdopodobieństwa pojedynczego doświadczenia jest dany tabelką:

wynik	w_1	$\cdots$	w_i	$\cdots$	w_k
prawdopodobieństwo	$p_1(\theta)$	$\cdots$	$p_i(\theta)$	$\cdots$	$p_k(\theta)$

gdzie $p_i(\theta)$ są, dla $i = 1, \cdots, k$, znanymi funkcjami nieznanego parametru θ. Ten parametr może być wektorem d-wymiarowym: $\theta = (\theta_1, \ldots, \theta_d)$. Statystykę χ^2 budujemy w dwóch etapach. Najpierw *estymujemy* θ metodą największej wiarogodności. Mamy

$$f_\theta(n_1, \ldots, n_k) = \mathbb{P}_\theta(N_1 = n_1, \ldots, N_k = n_k) = \text{const} \cdot p_1(\theta)^{n_1} \cdots p_k(\theta)^{n_k}$$

(tak jak w „zwykłym" teście χ^2, zmienna N_i oznacza liczbę powtórzeń i-tego wyniku). Logarytm wiarogodności ma więc postać

$$l(\theta) = \text{const} + \sum_{i=1}^{k} N_i \log p_i(\theta)$$

i estymator największej wiarogodności obliczamy rozwiązując układ równań

$$\frac{\partial l(\theta)}{\partial \theta_j} = \sum_{i=1}^{k} N_i \frac{1}{p_i(\theta)} \frac{\partial p_i(\theta)}{\partial \theta_j} = 0, \quad (j = 1, \ldots, d).$$

Statystykę χ^2 określamy tak jak w (6.2.3), wstawiając $\hat{\theta} = \text{ENW}(\theta)$ w miejsce nieznanego θ:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(N_i - np_i(\hat{\theta})\right)^2}{np_i(\hat{\theta})}.$$

Rozkład tej statystyki jest *inny*, niż statystyki obliczonej w przypadku ustalonych z góry prawdopodobieństw p_i. Intuicyjnie jest dość jasne, że statystyka z *estymowanymi* prawdopodobieństwami klatek jest średnio *mniejsza*, bo wartości oczekiwane są wyliczane na podstawie danych i mogą się „dopasować" do wartości obserwowanych. Okazuje się, że to „dopasowanie" zmniejsza *liczbę stopni swobody* granicznego rozkładu χ^2. Podamy ten fakt bez dowodu:

6.2.7 Stwierdzenie. *W opisanej wyżej sytuacji, rozkład statystyki χ^2 zmierza do rozkładu $\chi^2(k - d - 1)$, gdy $n \to \infty$.*

Przypomnijmy, że d jest liczbą współrzędnych wektora $(\theta_1, \ldots, \theta_d)$. Symbolicznie,

stopnie swobody = liczba klatek − *liczba estymowanych parametrów* − 1.

Oczywiście, wynika stąd następujący przepis na budowę testu na poziomie istotności α (w przybliżeniu, dla dużej liczby powtórzeń n): odrzucamy H_0 gdy

$$\chi^2 > c, \qquad c = \chi^2_{1-\alpha}(k - d - 1).$$

6.2.8 PRZYKŁAD (Zgodność danych z modelem Hardy'ego-Weinberga). Mówiliśmy już o tym, jak estymować nieznany parametr w tym modelu, patrz Przykłady 3.1.2 i 4.2.4. Teraz zajmiemy się sprawdzeniem hipotezy, że model poprawnie opisuje wyniki doświadczenia (genetycznego, w tym przypadku). Nasza hipoteza jest więc taka:

$$\mathrm{H}_0 : p_1 = \theta^2,\ \ p_2 = 2\theta(1-\theta),\ \ p_3 = (1-\theta)^2 \quad \text{dla pewnego } \theta \in (0,1).$$

Wykorzystamy estymator największej wiarogodności

$$\hat{\theta} = \mathrm{ENW}(\theta) = \frac{2N_1 + N_2}{2n}.$$

Statystyką testową jest

$$\chi^2 = \frac{\left(N_1 - n\hat{\theta}^2\right)^2}{n\hat{\theta}^2} + \frac{\left(N_2 - 2n\hat{\theta}(1-\hat{\theta})\right)^2}{2n\hat{\theta}(1-\hat{\theta})} + \frac{\left(N_3 - n(1-\hat{\theta})^2\right)^2}{n(1-\hat{\theta})^2}.$$

Proste, choć nieco żmudne przekształcenia pozwalają uprościć to wyrażenie:

$$\chi^2 = n\left(1 - \frac{N_2}{2n\hat{\theta}(1-\hat{\theta})}\right)^2.$$

Dla dużych n, ta statystyka ma w przybliżeniu rozkład $\chi^2(3-1-1) = \chi^2(1)$. □

Test niezależności chi-kwadrat. Przypuśćmy, że w pojedynczym doświadczeniu obserwujemy *parę* zmiennych losowych (X, Y) o wartościach w zbiorze $\{1,\dots,r\} \times \{1,\dots,s\}$ (te wartości należy traktować jako umowne „etykietki"). Rozkład prawdopodobieństwa jest opisany dwuwymiarową tabelką $(p_{ij}; i = 1,\dots,r; j = 1,\dots,s)$, gdzie $p_{ij} = \mathbb{P}(X = i, Y = j)$. Prawdopodobieństwa brzegowe $\mathbb{P}(X = i)$ i $\mathbb{P}(Y = j)$ zapiszemy w postaci

$$p_{i\bullet} = \sum_{j=1}^{s} p_{ij}, \qquad p_{\bullet j} = \sum_{i=1}^{r} p_{ij}.$$

Jeśli powtarzamy doświadczenie niezależnie n-krotnie, mamy niezależne wektory losowe $(X_1, Y_1), \dots, (X_n, Y_n)$, każdy o rozkładzie takim jak (X, Y). Możemy zbudować dwuwymiarową tabelkę (N_{ij}), gdzie

$$N_{ij} = \sum_{k=1}^{n} \mathbb{I}(X_k = i, Y_k = j),$$

$N_{ij} =$ liczba doświadczeń, które zakończyły się parą wyników (i, j).

Jest to tak zwana *tablica kontyngencji*. Wielkości „brzegowe" w tej tabelce oznaczymy

$$N_{i\bullet} = \sum_{j=1}^{s} N_{ij}, \qquad N_{\bullet j} = \sum_{i=1}^{r} N_{ij}.$$

Rozpatrzymy hipotezę, która stwierdza *niezależność* zmiennych X i Y:

$$\mathrm{H}_0 : p_{ij} = p_{i\bullet} p_{\bullet j}, \quad (i = 1, \ldots, r; j = 1, \ldots, s).$$

Może na pierwszy rzut oka tego nie widać, ale jest to specjalny przypadek schematu rozpatrzonego w poprzednim punkcie: prawdopodobieństwa „klatek" p_{ij} są znanymi funkcjami nieznanego parametru

$$\theta = (p_{1\bullet}, \ldots, p_{r-1\bullet}, p_{\bullet 1}, \ldots, p_{\bullet s-1}).$$

Zauważmy, że $p_{r\bullet} = 1 - p_{1\bullet} - \cdots - p_{r-1\bullet}$ i $p_{\bullet s} = 1 - p_{\bullet 1} - \cdots - p_{\bullet s-1}$, więc tych ostatnich prawdopodobieństw brzegowych nie włączyliśmy do wektora θ. W rezultacie θ ma wymiar $r + s - 2$. Zbudujemy statystykę chi-kwadrat, używając *estymowanych* prawdopodobieństw klatek

$$\hat{p}_{ij} = \frac{N_{i\bullet} N_{\bullet j}}{n^2}.$$

Przekonajmy się, że są to estymatory *największej wiarogodności* w modelu określonym hipotezą H_0. Wygodnie nam będzie potraktować wiarogodność jako funkcję *rozszerzonego* wektora parametrów $(p_{1\bullet}, \ldots, p_{r\bullet}, p_{\bullet 1}, \ldots, p_{\bullet s})$ i znaleźć maksimum *warunkowe*, przy ograniczeniach $\sum_{i=1}^{r} p_{i\bullet} = 1$ i $\sum_{j=1}^{s} p_{\bullet j} = 1$. Logarytm wiarogodności ma postać

$$l(p_{1\bullet}, \ldots, p_{r\bullet}, p_{\bullet 1}, \ldots, p_{\bullet s}) = \mathrm{const} + \sum_{i,j} N_{ij} [\log p_{i\bullet} + \log p_{\bullet j}].$$

Zgodnie z metodą Lagrange'a poszukujemy minimum funkcji $l - a - b$, gdzie

$$a = \lambda \left(\sum_i p_{i\bullet} - 1 \right), \quad b = \mu \left(\sum_j p_{\bullet j} - 1 \right),$$

λ i μ są „mnożnikami Lagrange'a". Ponieważ

$$\frac{\partial}{\partial p_{i\bullet}} (l - a - b) = \frac{N_{i\bullet}}{p_{i\bullet}} - \lambda, \qquad \frac{\partial}{\partial p_{\bullet j}} (l - a - b) = \frac{N_{\bullet j}}{p_{\bullet j}} - \mu,$$

więc przyrównując pochodne do zera dostajemy $p_{i\bullet} = N_{i\bullet}/\lambda$ i $p_{\bullet j} = N_{\bullet j}/\mu$. Wykorzystując postać ograniczeń, obliczamy $\lambda = \mu = n$ i ostatecznie otrzymujemy „oczywiste" estymatory $\hat{p}_{i\bullet} = \mathrm{ENW}(p_{i\bullet}) = N_{i\bullet}/n$ i $\hat{p}_{\bullet j} = \mathrm{ENW}(p_{\bullet j}) = N_{\bullet j}/n$. Wartość „oczekiwana" (przy założeniu H_0) w „klatce" (i, j) jest równa $n\hat{p}_{i\bullet}\hat{p}_{\bullet j} = N_{i\bullet}N_{\bullet j}/n$. Z tego co zostało powiedziane wynika, że statystyka chi-kwadrat przeznaczona do testowania hipotezy o niezależności jest następująca:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(N_{ij} - N_{i\bullet}N_{\bullet j}/n)^2}{N_{i\bullet}N_{\bullet j}/n}.$$

Dla $n \to \infty$, rozkład tej statystyki zmierza do rozkładu

$$\chi^2\left((r-1)(s-1)\right)$$

na mocy Stwierdzenia 6.2.7, bo $rs - (r-1) - (s-1) - 1 = (r-1)(s-1)$.

6.2.9 PRZYKŁAD (Zależność gustów muzycznych i poglądów politycznych). Rzecznik partii A twierdzi, że wśród zwolenników tej partii, miłośnicy muzyki disco-polo, rockowej i symfonicznej występują w tych samych proporcjach, co w całej populacji wyborców. Przeprowadzono sondaż. Wśród wylosowanych 100 wyborców (spośród osób o jednoznacznie sprecyzowanych preferencjach muzycznych), wyniki badania były następujące:

	Popieram A	Nie popieram A	Razem
Słucham Disco Polo	25	10	35
Słucham muzyki rockowej	20	20	40
Słucham muzyki symfonicznej	15	10	25
Razem	60	40	100

Sformułowana przez rzecznika hipoteza stwierdza niezależność dwóch „cech" wyborcy: stosunku do partii A i preferencji muzycznych. Przeprowadzimy test niezależności chi-kwadrat na poziomie istotności 0.05. Statystyka testowa ma wartość

$$\begin{aligned}\chi^2 &= \frac{(25 - 60*35/100)^2}{60*35/100} + \frac{(20 - 60*40/100)^2}{60*40/100} + \frac{(15 - 60*25/100)^2}{60*25/100} \\ &+ \frac{(10 - 40*35/100)^2}{40*35/100} + \frac{(20 - 40*40/100)^2}{40*40/100} + \frac{(10 - 40*25/100)^2}{40*25/100} \\ &= 2.53.\end{aligned}$$

Wartość krytyczną odczytujemy z tablicy rozkładu chi-kwadrat z $(2-1)(3-1) = 2$ stopniami swobody. Ponieważ $\chi^2_{0.95}(2) = 5.99 > 2.53$, więc zebrane dane nie dają podstaw, żeby zarzucić rzecznikowi kłamstwo. □

Porównanie kilku rozkładów wielomianowych. Test niezależności chi-kwadrat opisany w punkcie poprzednim stosuje się bez zmian do danych w postaci tablicy kontyngencji o ustalonych (nielosowych) sumach elementów w każdym wierszu. Tego rodzaju tablica odpowiada innemu sposobowi zbierania danych i innemu modelowi statystycznemu. Opiszemy rzecz dokładniej. Dla $i = 1, \ldots, r$, wykonujemy n_i doświadczeń zgodnie ze schematem wielomianowym $\mathrm{Mult}(n_i, p_{i1}, \ldots, p_{is})$. To znaczy, że każde z doświadczeń ma s możliwych wyników, przy tym j-ty wynik pojawia się z prawdopodobieństwem p_{ij}. Chcemy przeprowadzić test hipotezy stwierdzającej, że prawdopodobieństwa poszczególnych wyników są identyczne dla każdego z r rozkładów wielomianowych:

$$\mathrm{H}_0 : p_{1j} = \cdots = p_{rj} \qquad (j = 1, \ldots, s).$$

Dane są takiej samej postaci jak w punkcie poprzednim: jeśli $(N_{i1}, \ldots, N_{is}) \sim \mathrm{Mult}(n_i, p_{i1}, \ldots, p_{is})$ to mamy też dwuwymiarową tablicę (N_{ij}). Teraz jednak sumy $n_i = \sum_{j=1}^{s} N_{ij}$ są *nielosowe* (sumy $N_{\bullet j} = \sum_{i=1}^{r} N_{ij}$ są zmiennymi losowymi). Statystyka zbudowana tak samo jak w punkcie B:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - n_i N_{\bullet j}/n)^2}{n_i N_{\bullet j}/n},$$

gdzie $n = \sum_{i=1}^{r} n_i$. Okazuje się, że w nowej sytuacji rozkład asymptotyczny $(n_1 \to \infty, \ldots, n_s \to \infty)$ statystyki χ^2 jest też równy $\chi^2((r-1)(s-1))$, jeśli H_0 jest prawdziwa.

Zauważmy, że H_0 ma w rozpatrywanym tu schemacie ten sam intuicyjny sens, co hipoteza o niezależności zmiennych losowych (X, Y) rozpatrzona w poprzednim punkcie. Można powiedzieć, że mamy teraz dwie cechy (x, Y), ale pierwsza z nich jest „zwykłą", nielosową zmienną. Hipoteza zerowa mówi, że rozkład prawdopodobieństwa zmiennej losowej Y *nie zależy* od wartości x. W punkcie poprzednim hipoteza zerowa mówiła, że rozkład *warunkowy* zmiennej losowej Y nie zależy od wartości $X = x$. Oba modele różnią się sposobem zbierania danych. Albo losujemy n „osobników" z populacji zróżnicowanej pod względem cechy x, albo losujemy po n_i osobników spośród tych, dla których cecha x ma wartość i.

6.2.10 PRZYKŁAD (Porównanie kilku rozkładów dwumianowych). Dane dotyczą r klientów towarzystwa ubezpieczeniowego. Przypuśćmy, że i-ty klient ubezpieczał samochód przez n_i lat. W każdym roku mógł spowodować wypadek („sukces") lub nie spowodować („porażka"). Ignorujemy, dla uproszczenia, możliwość wielokrotnych wypadków w ciągu roku. Mamy do czynienia z r schematami Bernoulliego $\mathrm{Bin}(n_i, p_i)$, dla $i = 1, \ldots, r$. Istotne jest pytanie, czy prawdopodobieństwo „sukcesu" (wypadku) jest równe dla wszystkich klientów. Stawiamy hipotezę

$$\mathrm{H}_0 : p_1 = \cdots = p_r.$$

W języku aktuarialnym, hipoteza stwierdza, że rozpatrywana grupa polis jest jednorodna pod względem „poziomu ryzyka". Jest to, rzecz jasna, szczególny przypadek zagadnienia rozpatrywanego w poprzednio i zastosujemy test niezależności chi-kwadrat. Teraz statystyka testowa przybiera szczególnie prostą postać. Niech $K_i = N_{i1}$ będzie liczbą sukcesów w i-tym schemacie Bernoulliego. Wektor $(K_1, \ldots, K_r)$ opisuje całość obserwacji, bo $N_{i2} = n_i - K_i$. Proste przekształcenia prowadzą do następującego wzoru na statystykę testową:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{r} \left[\frac{(K_i - n_i\hat{p})^2}{n_i\hat{p}} + \frac{\left(n_i - K_i - n_i(1-\hat{p})\right)^2}{n_i\hat{p}} \right] \\ &= \sum_{i=1}^{r} \frac{(K_i - n_i\hat{p})^2}{n_i\hat{p}(1-\hat{p})},\end{aligned}$$

gdzie $\hat{p} = \sum_i K_i / \sum_i n_i$. Ta statystyka ma graniczny rozkład $\chi^2(r-1)$ przy $n_1 \to \infty, \ldots, n_r \to \infty$, jeśli H_0 jest prawdziwa.

Rozważmy takie (fikcyjne) dane, dotyczące przebiegu ubezpieczenia 10 klientów w ciągu 4 lat (gwiazdki oznaczają wypadki):

Klienci:	1	2	3	4	5	6	7	8	9	10
1 rok	*				*				*	*
2 rok			*						*	*
3 rok										
4 rok	*						*			*
liczba wypadków	2	0	1	0	1	0	1	0	2	3

Mamy tu $r = 10$, $n_1 = \cdots = n_{10} = 4$, $\hat{p} = 10/40 = 0.25$ i $n_i\hat{p} = 1$. Dla wygody zapisaliśmy tabelkę w postaci „odwróconej", zamieniając wiersze na kolumny. Przeprowadzimy test hipotezy $\mathrm{H}_0 : p_1 = \cdots = p_{10}$. Przyjmijmy poziom istotności 0.05. Wartość statystyki testowej

$$\chi^2 = \frac{1}{0.75}\Big[(2-1)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (2-1)^2 + (3-1)^2\Big] = 13.33$$

porównujemy z wartością krytyczną 16.9 (odczytaną z tablic $\chi^2(9)$). Test nie odrzuca H_0. Różnice w ilości wypadków dla poszczególnych klientów mieszczą się w granicach losowych fluktuacji, zdarzających się rozsądnie często w sytuacji, gdy klienci są „jednakowi". „Rozsądnie często" znaczy dla nas: z prawdopodobieństwem przynajmniej 0.05. Tyle może powiedzieć statystyk. Na szczęście, to nie statystyk decyduje o tym, czy obciążyć wszystkich 10 klientów *jednakową składką*! □

Test Kołmogorowa-Smirnowa

Niech F będzie ustaloną, ciągłą dystrybuantą. Rozpatrzmy hipotezę stwierdzającą, że obserwacje $X_1, \ldots, X_n$ są próbką z rozkładu o dystrybuancie F. Symbolicznie,

$$\mathrm{H}_0 : X_1, \ldots, X_n \sim F.$$

Niech $\hat{F}_n$ będzie dystrybuantą empiryczną. Statystyka testowa jest równa

$$D_n = \sup_{-\infty<x<\infty} |\hat{F}_n(x) - F(x)|.$$

Łatwo zauważyć, że

$$D_n = \max(D_n^+, D_n^-),$$

gdzie

$$D_n^+ = \max_{i=1,\ldots,n}\left(\frac{i}{n} - F(X_{i:n})\right), \qquad D_n^- = \max_{i=1,\ldots,n}\left(F(X_{i:n}) - \frac{i-1}{n}\right).$$

Idea testu jest prosta. Odrzucamy H_0, jeśli odchylenia dystrybuanty empirycznej od dystrybuanty teoretycznej (sprecyzowanej w hipotezie) są duże, czyli gdy $D_n > c$. Sposób doboru poziomu krytycznego c opiera się na dwóch faktach. Pierwszy z nich jest prosty.

6.2.11 Stwierdzenie. *Dla dowolnego rozkładu ciągłego F, rozkład statystyki D_n jest taki sam.*

Naturalnie, mamy tu na myśli rozkład statystyki D_n przy prawdziwej hipotezie H_0.

Dowód. Dla uproszczenia przeprowadzimy dowód przy założeniu, że istnieje dobrze zdefiniowana funkcja odwotna F^{-1} do dystrybuanty. Rozpatrzmy próbkę $U_1, \ldots, U_n$ z rozkładu *jednostajnego* $\mathrm{U}(0,1)$. Zauważmy, że

$$\sup_{0<u<1} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(U_i \leqslant u) - u \right| = \sup_{-\infty<x<\infty} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(U_i \leqslant F(x)) - F(x) \right|$$
$$= \sup_{-\infty<x<\infty} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(F^{-1}(U_i) \leqslant x) - F(x) \right|.$$

Wiemy, że zmienne losowe $X_i = F^{-1}(U_i)$ mają rozkład o dystrybuancie $F(x)$. Ze wzoru napisanego powyżej wynika, że rozkład statystyki D_n jest taki sam dla rozkładu $\mathrm{U}(0,1)$ i dla rozkładu o dystrybuancie F. □

Stwierdzenie 6.2.11 pozwala budować tablice rozkładu statystyki D_n i obliczać p-wartości. Potrzebna jest, w zasadzie, oddzielna tablica dla każdego n. Dla dużych rozmiarów próbki, sytuacja jeszcze bardziej się upraszcza. Kołmogorow udowodnił następujące twierdzenie.

6.2.12 TWIERDZENIE. *Jeśli $n \to \infty$, to*

$$\mathbb{P}\left(\sqrt{n} D_n \leqslant d\right) \to K(d) = \sum_{k=-\infty}^{\infty} (-1)^k \mathrm{e}^{-2k^2 d^2}, \quad (d > 0). \qquad \square$$

Dla praktyka znaczenie ma fakt, że $K(d)$ jest znaną dystrybuantą. Napisaliśmy rozwinięcie tej dystrybuanty w postaci szeregu tylko po to, żeby Czytelnik mógł docenić elegancję wyniku Kołmogorowa.

Wróćmy do naszego testu zgodności. Załóżmy, że n jest „dostatecznie duże". W klasycznej, przedkomputerowej postaci test przeprowadzamy następująco.

Ustalamy poziom istotności α i odczytujemy z tablic odpowiedni kwantyl dystrybuanty K, to znaczy taką liczbę $\mathrm{d}_{1-\alpha}$, że $K(\mathrm{d}_{1-\alpha}) = 1-\alpha$.

$$\text{odrzucamy } \mathrm{H}_0, \text{ jeśli } \sqrt{n}D_n > \mathrm{d}_{1-\alpha}.$$

6.2.13 PRZYKŁAD (Testowanie generatora liczb losowych). Chcemy sprawdzić, czy ciąg produkowany przez R-owską funkcję `runif()` zachowuje się jak ciąg niezależnych zmiennych losowych o rozkładzie jednostajnym $\mathrm{U}(0,1)$. Test Kołmogorowa jest jednym z podstawowych sprawdzianów. Dla ilustracji rozważmy 10-cio elementową „próbkę" $(*)$:

$$0.4085, 0.5267, 0.3751, 0.8329, 0.0846, 0.8306, 0.6264, 0.3086, 0.3662, 0.7952$$

(w rzeczywistości bada się znacznie dłuższe ciągi). Przeprowadzimy test hipotezy

$$\mathrm{H}_0 : (*) \text{ jest próbką losową z } \mathrm{U}(0,1)$$

na poziomie istotności 0.1. Dystrybuanta rozkładu jednostajnego jest równa $F(x) = x$ dla $0 < x < 1$. Sposób obliczania statystyki D_{10} pokazuje następująca tabelka:

$X_{i:10}$	$(i-1)/10$	$i/10$	$i/10 - X_{i:10}$	$X_{i:10} - (i-1)/10$
0.0846	0.0	0.1	0.0154	0.0846
0.3086	0.1	0.2	−0.1086	0.2086
0.3662	0.2	0.3	−0.0662	0.1662
0.3751	0.3	0.4	0.0249	0.0751
0.4085	0.4	0.5	0.0915	0.0085
0.5267	0.5	0.6	0.0733	0.0267
0.6264	0.6	0.7	0.0736	0.0264
0.7952	0.7	0.8	0.0048	0.0952
0.8306	0.8	0.9	0.0694	0.0306
0.8329	0.9	1.0	0.1671	−0.0671

Widać, że $D_{10}^{+} = 0.1671$ i $D_{10}^{-} = 0.2086$ (największe liczby w dwóch ostatnich kolumnach). Stąd $D_{10} = 0.2086$. Wartość krytyczna statystyki D_{10}, odczytana z tablic, jest równa $\mathrm{d}_{0.9} = 0.369$. Test Kołmogorowa nie odrzuca hipotezy, że nasz ciąg jest próbką z rozkładu $\mathrm{U}(0,1)$. Doświadczenie nie dało powodów, by kwestionować generator.

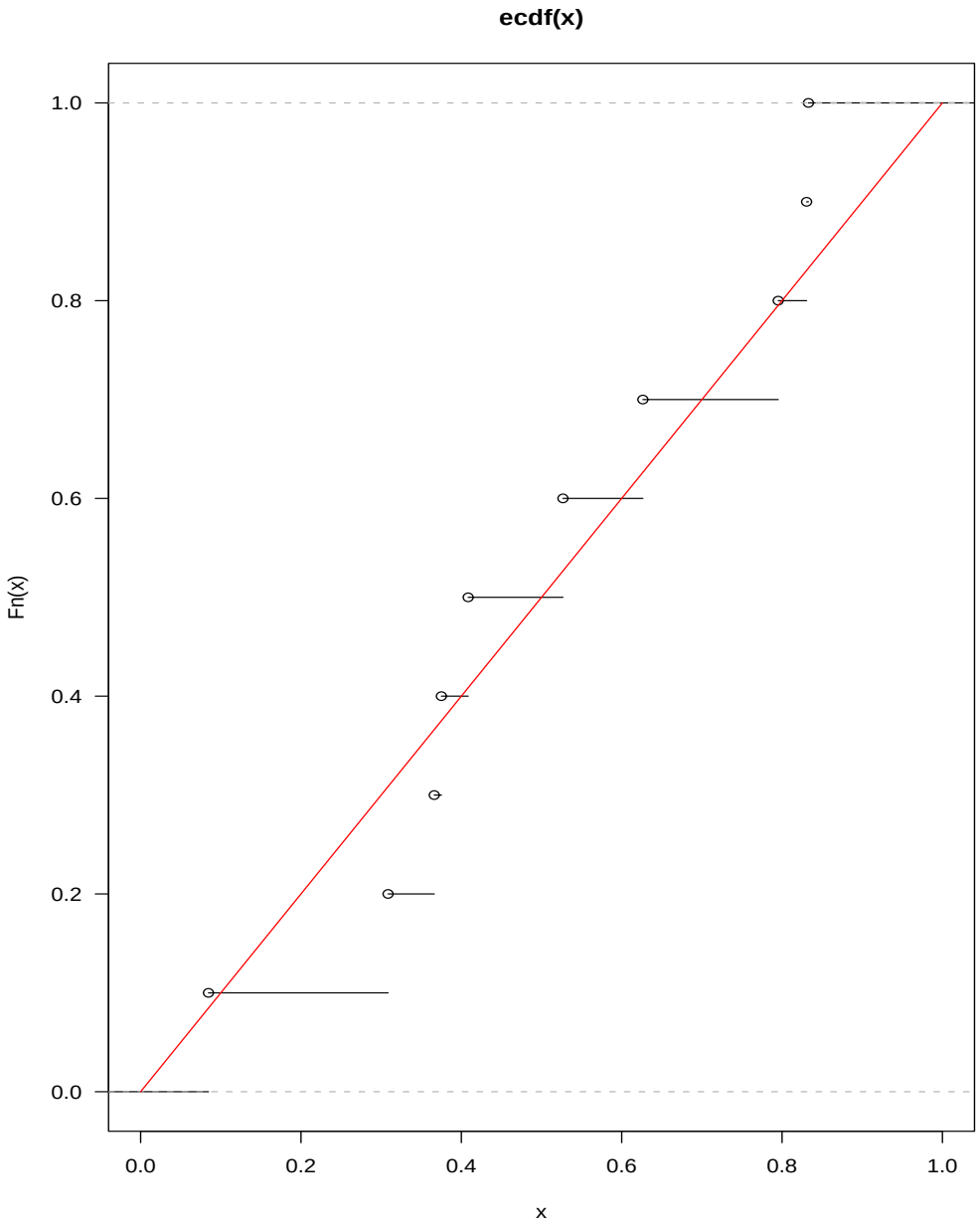
Następny przykład ilustruje użycie R-owskiej funkcji `ks.test()`

```
> X=c(0.4085, 0.5267, 0.3751, 0.8329, 0.0846, 0.8306, 0.6264,
0.3086, 0.3662, 0.7952)
> ks.test(X,punif)


One-sample Kolmogorov-Smirnov test


data:  X
D = 0.2086, p-value = 0.7037
alternative hypothesis:  two.sided
```

Drugim parametrem funkcji `ks.test()` jest nazwa funkcji obliczającej dystrybuantę rozkładu sprecyzowanego w H_0, w naszym przypadku `punif()`. Implementacja testu Kołmogorowa-Smirnowa w R automatycznie oblicza dokładną p-wartość, jeśli $n \leqslant 100$, zaś dla $n > 100$ używa asymptotycznej aproksymacji.

Warto jeszcze obejrzeć dystrybuantę empiryczną na tle „prawdziwej" dystrybuantu $U(0,1)$.

Komputer pozwala powtarzać całą zabawę niemal bez ograniczeń. Opisane wyżej „doświadczenie losowe" (10-krotne wywołanie funkcji `random`) wykonaliśmy 10000 razy. Za każdym razem obliczyliśmy wartość statystyki D_{10}. W 1012 przypadkach wartość statystyki przekroczyła 0.369. Test Kołmogorowa „odrzucił hipotezę H_0" 1012 razy. Ponieważ poziom istotności testu jest równy $0.1 = 1000/10000$, zgodność z przewidywaniami teorii jest uderzająca. To oczywiście świadczy o tym, że wyprodukowany przez generator ciąg 100000 liczb bardzo dobrze naśladuje ciąg niezależnych zmiennych losowych o rozkładzie jednostajnym. Gdyby test „odrzucał H_0" zbyt rzadko, wtedy mielibyśmy powód do niepokoju.



Dystrybuabta empiryczna i „hipotetyczna” dla 10-cio elementowej próbki.

□

Testy dla dwóch próbek

Rozpatrzmy dwa ciągi obserwacji: $X_1, \ldots, X_n$ i $Y_1, \ldots, Y_m$. Zakładamy, że są to próbki z rozkładów o dystrybuantach, odpowiednio, F i G:

$$X_1, \ldots, X_n \sim F, \qquad Y_1, \ldots, Y_m \sim G.$$

Stawiamy hipotezę, że obie próbki pochodzą z *tego samego* rozkładu prawdopodobieństwa:

$$\mathrm{H}_0 : F = G.$$

Sens tej hipotezy jest taki, że dwa doświadczenia losowe nie różnią się w istotny sposób. Najczęściej oba doświadczenia przeprowadzamy w nieco odmiennych warunkach. Chcemy skontrolować przypuszczenie, że ta odmienność nie wpływa na przebieg zjawiska.

6.2.14 PRZYKŁAD. Zbadano iloraz inteligencji n uczniów szkoły „X" i m uczniów szkoły „Y". Przykładowe dane, dla $n = 7$ i $m = 6$, wyglądają następująco:

X	110	112	115	98	130	123	141
Y	88	135	140	95	125	138	

Stawiamy hipotezę, że poziom inteligencji uczniów obu szkół jest taki sam. Matematycznie, hipoteza sprowadza się do przypuszczenia, że X-y i Y-i są próbkami losowymi *z tego samego* rozkładu prawdopodobieństwa. □

Przedstawimy kilka testów hipotezy $\mathrm{H}_0 : F = G$ w problemie dwóch próbek. Sposób stosowania każdego z tych testów zilustrujemy na Przykładzie 6.2.14. Przyznajmy od razu, że przedstawimy testy, przeznaczone w zasadzie do porównywania rozkładów *ciągłych*. Iloraz inteligencji jest zmienną *dyskretną*, bo przyjmuje tylko wartości całkowite. Liczba tych wartości jest jednak na tyle duża, że rozkład ciągły możemy uznać za rozsądne przybliżenie.

Dwupróbkowy test Kołmogorowa-Smirnowa. Idea tego testu jest podobna do testu jednopróbkowego, omówionego poprzednio. Porównujemy teraz *dwie dystrybuanty empiryczne* – tę dla X-ów i tę dla Y-ów:

$$D_{n,m} = \sup_{-\infty<z<\infty} |\hat{F}_n(z) - \hat{G}_m(z)|,$$

gdzie

$$\hat{F}_n(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leqslant z), \qquad \hat{G}_m(z) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}(Y_i \leqslant z).$$

Jeśli hipoteza H_0 jest prawdziwa, to rozkład statystyki $D_{n,m}$ nie zależy od F pod warunkiem, że dystrybuanta F jest ciągła. Zachodzi przy tym zbieżność

$$\mathbb{P}\left(\sqrt{\frac{nm}{n+m}} D_{n,m} \leqslant d\right) \to K(d), \qquad (n, m \to \infty).$$

Dystrybuanta K jest tu ta sama, co w Twierdzeniu 6.2.12. Sposób postępowania jest podobny, jak dla testu jednopróbkowego. Odrzucamy H_0, jeśli $D_{n,m} > c$. W praktyce korzystamy z gotowej funkcji, która oblicza $D_{n,m}$ i (dokładną lub przybliżoną) p-wartość.

6.2.15 PRZYKŁAD. Prześledźmy zastosowanie testu Kołmogorowa-Smirnowa na danych przedstawionych w Przykładzie 6.2.14. Sposób obliczania statystyki $D_{n,m}$ pokazuje następująca tabelka:

z		$\hat{F}_n(z)$	$\hat{G}_m(z)$	$\lvert\hat{F}_n(z) - \hat{G}_m(z)\rvert$
88	Y	0/7	1/6	7/42
95	Y	0/7	2/6	14/42
98	X	1/7	2/6	8/42
110	X	2/7	2/6	2/42
112	X	3/7	2/6	4/42
115	X	4/7	2/6	10/42
123	X	5/7	2/6	16/42
125	Y	5/7	3/6	9/42
130	X	6/7	3/6	15/42
135	Y	6/7	4/6	8/42
138	Y	6/7	5/6	1/42
140	Y	6/7	6/6	6/42
141	X	7/7	6/6	0/42

Widać, że $D_{7,6} = 16/42$. Test na poziomie istotności 0.05 odrzuca H_0 jeśli statystyka przekracza 30/42 (to można odczytać z tablic). Konkluzja jest taka, że nie ma podstaw do odrzucenia hipotezy zerowej. Dane nie są sprzeczne z założeniem, że poziom inteligencji uczniów w obu szkołach ma ten sam rozkład prawdopodobieństwa. □

Test serii. Porządkujemy połączoną próbkę $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ w kolejności rosnącej. W tym uporządkowanym ciągu liczymy serie. *Serią* nazywamy ciąg sąsiadujących ze sobą elementów z tej samej próbki, to znaczy X-ów lub Y-ów. Niech zmienna losowa K będzie liczbą serii. Jeśli H_0 jest prawdziwa, czyli $F = G$, to uporządkowanie połączonej próbki jest równoważne ustawieniu n klocków z literą X i m klocków z literą Y w losowej kolejności. Jeśli liczba serii K jest *zbyt mała*, znaczy to, że X-y i Y-i są „zbyt słabo przemieszane". W takiej sytuacji jesteśmy skłonni odrzucić H_0. Policzenie $\mathbb{P}(K = k)$ w „klockowym" schemacie jest elementarnym (co nie znaczy, że łatwym!) zadaniem kombinatorycznym. Dostępne są tablice, podające rozkład liczby serii dla danych n i m. Pozwalają one odpowiednio dobrać obszar krytyczny testu serii.

Wróćmy do danych z Przykładu 6.2.14. Test serii na poziomie istotności 0.05 prowadzi do odrzucenia $\mathrm{H}_0 : F = G$ jeśli $K < 5$ (dla $n = 7$ i $m = 6$). Uporządkowanie połączonej próbki obrazuje tabelka:

88	95	98	110	112	115	123	125	130	135	138	140	141
Y	Y	X	X	X	X	X	Y	X	Y	Y	Y	X

Mamy $K = 6$ serii. Nie odrzucamy H_0 na poziomie istotności 0.05. Test serii prowadzi do takiej samej konkluzji, co test Kołmogorowa-Smirnowa. □

Test Wilcoxona-Manna-Whitneya. Jest to kolejny test „kombinatoryczny" hipotezy $\mathrm{H}_0 : F = G$. Podobnie jak poprzednio, statystyka testowa zależy od kolejności X-ów i Y-ów w połączonej, uporządkowanej próbce. Niech W oznacza *sumę rang* X-ów. Rangą nazywamy numer elementu w uporządkowanej próbce. Niech U oznacza *sumę inwersji* X-ów. Inwersją X-a nazywamy liczbę Y-ów poprzedzających go w uporządkowanej próbce. Dla danych z Przykładu A, rangi poszczególnych elementów są napisane w dolnym wierszu tabelki:

88	95	98	110	112	115	123	125	130	135	138	140	141
Y	Y	X	X	X	X	X	Y	X	Y	Y	Y	X
1	2	3	4	5	6	7	8	9	10	11	12	13

Zatem $W = 3{+}4{+}5{+}6{+}7{+}9{+}13 = 47$. Z kolei $U = 2{+}2{+}2{+}2{+}2{+}3{+}6 = 19$.

Oto kilka spostrzeżeń na temat statystyk W i U. Po pierwsze, obie te statystyki są równoważne, bo jest między nimi prosta zależność:

$$W = \frac{n(n+1)}{2} + U.$$

Istotnie, jeśli oznaczymy przez R_i oraz I_i odpowiednio, rangę i inwersję i-tego co do wielkości X-a to łatwo się przekonać, że $W = \sum_{i=1}^{n} R_i = \sum_{i=1}^{n}(i + I_i) = n(n+1)/2 + U$. Każdy test zależący od statystyki W można wyrazić przy pomocy U i odwrotnie. Jest tylko sprawą gustu, którą z nich wolimy się posługiwać. Napiszmy teraz $W = W_X$ i $U = U_X$. Niech W_Y i U_Y oznaczają sumę rang i sumę inwersji Y-ów (określone tak samo, tylko obie próbki zamieniają się rolami). Mamy

$$U_X + U_Y = mn,$$

bo $W_X + W_Y = \sum_{i=1}^{n+m} i = (n+m)(n+m+1)/2 = U_X + n(n+1)/2 + U_Y + m(m+1)/2$. Wreszcie zauważmy, że rozkład zmiennej losowej U jest symetryczny w tym sensie, że

$$\mathbb{P}(U = u) = \mathbb{P}(U = nm - u).$$

Żeby to uzasadnić, wystarczy rozważyć uporządkowanie połączonej próbki w odwrotnej, malejącej kolejności.

Rzecz jasna, odrzucamy $\mathrm{H}_0 : F = G$ wtedy, gdy U_X lub U_Y jest zbyt duże. Ze względu na wspomnianą własność symetrii, znajdujemy takie u, że

$$\mathbb{P}(U < u) = \mathbb{P}(U > mn - u) \leqslant \alpha/2,$$

przy prawdziwości H_0. Odrzucamy H_0 jeśli $U < u$ lub $U > mn - u$. W naszym przykładzie, dla poziomu istotności $\alpha = 0.05$, dla $n = 7$ i $m = 6$, wartość krytyczna u jest równa $u = 7$. Nie odrzucamy H_0.

W praktyce, tak jak w naszym przykładzie, test Wilcoxona bywa stosowany do próbek z rozkładów dyskretnych (choć teoretycznie ten test wymaga założenia o ciągłości rozkładów). W takiej sytuacji mogą pojawić się tak zwane „węzły", czyli równe obserwacje pochodzące z różnych próbek. Wyobraźmy sobie, że w naszym Przykładzie A, w próbce X-ów pojawiła się dodatkowa, ósma obserwacja, równa 140. Ponieważ w próbce Y-ów jest też obserwacja równa 140, nie wiadomo, której z nich przypisać rangę 12, a której 13. Stosuje

się wyjście kompromisowe: każdej z dwóch obserwacji przypisuje się „rangę" 12.5. W rezultacie, suma rang X-ów w połączonej, 14-elementowej próbce przyjmuje teraz wartość $W = 3+4+5+6+7+9+12.5+14 = 60.5$. Suma inwersji X-ów przyjmuje warość $U = 2+2+2+2+2+3+5.5+6 = 24.5$. Czytelnik zechce to sprawdzić. Dodajmy, że taki sposób postępowania ma tylko heurystyczne uzasadnienie, ale jest stosowany w praktyce.

Na zakończenie wspomnijmy, że dla dużych n i m, rozkład statystyki U jest w przybliżeniu normalny,

$$\mathrm{N}\left(\frac{nm}{2}, \frac{nm(n+m+1)}{12}\right).$$

Przybliżony test na poziomie istotności α jest więc taki: odrzucamy H_0 jeśli

$$\left|U - \frac{nm}{2}\right| > \mathrm{z}_{1-\alpha/2}\sqrt{\frac{nm(n+m+1)}{12}},$$

gdzie $\mathrm{z}_{1-\alpha/2}$ jest kwantylem rozkładu $\mathrm{N}(0,1)$. □

Zwróćmy uwagę, że odeszliśmy już w ostatnich przykładach od prostego schematu przedstawionego we wstępie do tego rozdziału. W problemie dwóch próbek, hipoteza zerowa nie precyzuje dokładnie rozkładu prawdopodobieństwa (modelu probabilistycznego). Mamy do czynienia z całą *rodziną* rozkładów prawdopodobieństwa (modelem statystycznym), w którym $F = G$ ale wspólna dystrybuanta może być dowolna. Mimo to, potrafiliśmy skonstruować statystyki testowe, których rozkład prawdopodobieństwa jest taki sam dla wszystkich rozkładów prawdopodobieństwa spełniających hipotezę zerową. To jest typowa sytuacja.

Testowanie zgodności z typem rozkładu

Obserwujemy próbkę $X_1, \ldots, X_n$ z pewnego nieznanego rozkładu o dystrybuancie F. Przypuszczamy, że jest to rozkład normalny. Stawiamy hipotezę

$$\mathrm{H}_0 : X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2) \quad \text{dla pewnych } \mu, \sigma.$$

Inaczej,

$$\mathrm{H}_0 : F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \quad \text{dla pewnych } \mu, \sigma.$$

Poznane przez nas testy zgodności wymagają pewnych modyfikacji, bo rozpatrywana teraz hipoteza zerowa jest złożona, to znaczy nie precyzuje jednego rozkładu prawdopodobieństwa, tylko całą rodzinę rozkładów.

Test chi-kwadrat. Najpierw dane ciągłe musimy zdyskretyzować, w sposób już wcześniej wyjaśniony. Wybieramy punkty $-\infty = a_0 < a_1 < \cdots < a_k = \infty$ i dzielimy prostą na rozłączne przedziały. Oznaczamy przez N_i liczbę obserwacji wpadających do i-tego przedziału. Wartości obserwowane porównujemy z wartościami oczekiwanymi $n\hat{p}_i$, gdzie

$$\hat{p}_i = \Phi\left(\frac{a_i - \hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{a_{i-1} - \hat{\mu}}{\hat{\sigma}}\right).$$

W zasadzie, należy za estymatory μ i σ przyjąć ENW obliczone dla *zdyskretyzowanych* zmiennych. Zgodnie z tym co powiedzieliśmy, te ENW są rozwiązaniem układu równań

$$\sum_i N_i \frac{1}{p_i}\frac{\partial p_i}{\partial \mu} = 0, \quad \sum_i N_i \frac{1}{p_i}\frac{\partial p_i}{\partial \sigma} = 0.$$

Statystyka χ^2 ma asymptotyczny rozkład $\chi^2(k-3)$ (bo estymujemy 2 parametry). W praktyce, rozwiązywanie napisanych wyżej równań jest zbyt kłopotliwe i przyjmuje się, że wstawienie „zwyczajnych" estymatorów $\hat{\mu} = \bar{X}$ i $\hat{\sigma} = S$ we wzory na p_i niewiele zmienia rozkład statystyki χ^2.

Dokładnie w ten sam sposób można testować inne hipotezy „zgodności z typem rozkładu", czyli

$$\mathrm{H}_0 : F = F_\theta \quad \text{dla pewnego } \theta,$$

gdzie $\{F_\theta\}$ jest daną parametryczną rodziną dystrybuant (rozkładów wykładniczych, gamma itp.). □

Test Kołmogorowa-Lilieforsa. Hipotezę o *normalności* rozkładu można testować przy pomocy zmodyfikowanej statystyki Kołmogorowa

$$D'_n = \sup_{-\infty<x<\infty} \left|\hat{F}_n(x) - \Phi\left(\frac{x - \bar{X}}{S}\right)\right|.$$

W tej formie, statystyka związana jest z nazwiskiem Lilieforsa. Wartości krytyczne są stablicowane. Jak łatwo zgadnąć, są mniejsze od wartości krytycznych „zwykłego" testu Kołmogorowa-Smirnowa (tego dla prostej hipotezy zerowej). Sytuacja jest faktycznie znacznie gorsza, niż dla testu χ^2:

rozkład statystyki Kołmogorowa-Lilieforsa zależy w istotny sposób od faktu, że obserwacje mają rozkład *normalny*. Porównaj, dla kontrastu, Stwierdzenie 6.2.7. Jeśli chcielibyśmy w podobny sposób testować (powiedzmy) hipotezę, że rozkład jest wykładniczy, to łatwo napisać odpowiednio zmodyfikowaną statystykę Kołmogorowa-Smirnowa – ale trudno ustalić wartości krytyczne. □

Rozdział 7

Teoria testowania hipotez

Przeanalizujemy zagadnienie testowania hipotez statystycznych w sposób bardziej formalny, niż w poprzednim rozdziale. Punktem wyjścia jest model statystyczny, a więc przestrzeń $\Omega = \mathcal{X}$ wyposażona w rodzinę $\{\mathbb{P}_\theta, \theta \in \Theta\}$ rozkładów prawdopodobieństwa.

7.1 Definicje

Nasze rozważania będą miały dość ogólny i abstrakcyjny charakter. Hipotezy statystyczne, czyli wypowiedzi na temat rozkładu prawdopodobieństwa, utożsamiamy z *podzbiorami przestrzeni parametrów* Θ (tak, jak zdarzenia losowe utożsamiamy z podzbiorami przestrzeni próbkowej Ω). Mówiąc o zagadnieniu testowania, zawsze będziemy rozważali *hipotezę zerową*

$$\mathrm{H}_0 : \theta \in \Theta_0$$

i *hipotezę alternatywną*

$$\mathrm{H}_1 : \theta \in \Theta_1.$$

Zakładamy, że Θ_0 i Θ_1 są podzbiorami przestrzeni Θ takimi, że $\Theta_0 \cap \Theta_1 = \emptyset$. Znaczy to, że H_0 i H_1 wzajemnie się wykluczają. Obie „konkurencyjne" hipotezy traktujemy nierównoprawnie. Zasadniczo, interpretacja jest taka: H_0 jest założeniem obowiązującym do czasu, gdy pojawią się dane doświadczalne sprzeczne (lub raczej „bardzo trudne do pogodzenia") z tą hipotezą. Z kolei, H_1 jest „ewentualnością, z którą powinniśmy się liczyć", jeśli przyjdzie nam zrezygnować z hipotezy H_0. *Testem* hipotezy H_0 przeciw alternatywie H_1 nazywamy statystykę

$$\delta : \mathcal{X} \to \{0, 1\}.$$

Wartość „1" interpretujemy jako decyzję o odrzuceniu H_0, zaś „0" oznacza, że nie odrzucamy H_0. Zbiór $K = \{x \in \mathcal{X} : \delta(x) = 1\}$ nazywamy *obszarem krytycznym* testu. Przeważnie test ma postać $\delta(X) = \mathbb{I}(T(X) > c)$, gdzie $T(X)$ jest pewną „wygodną" statystyką, zwaną *statystyką testową*, zaś c jest liczbą zwaną *wartością krytycznym*. Tak więc,

$$\begin{aligned} &\text{jeśli } T(X) > c \text{ czyli } \delta(X) = 1 \text{ czyli } X \in K, \text{ to odrzucamy } H_0; \\ &\text{jeśli } T(X) \leqslant c \text{ czyli } \delta(X) = 0 \text{ czyli } X \notin K, \text{ to nie odrzucamy } H_0. \end{aligned}$$

Oczywiście, jest tylko sprawą wygody, czy określamy funkcję δ, T i c, czy zbiór K: są to trzy równoważne sposoby opisu testu. Istotne jest to, że test jest *regułą podejmowania decyzji* w zależności od wyniku obserwacji. Ponieważ obserwacje są losowe, musimy czasem popełniać błędy. Skutki działania testu przedstawimy tak:

stan rzeczy \ decyzja	$\delta = 0$	$\delta = 1$
H_0 prawdziwa	O.K.	błąd I rodzaju
H_1 prawdziwa	błąd II rodzaju	O.K.

Będziemy również mówić, że decyzja „0" oznacza akceptację H_0 a decyzja „1" – akceptację H_1. Niektórzy statystycy starannie unikają takich sformułowań. To jednak kwestia interpretacji decyzji. Statystyka nie zajmuje się skutkami błędnych decyzji, tylko bada *częstość* podejmowania takich decyzji. Niech

$$1 - \beta(\theta) = \mathbb{P}_\theta\left(\delta(X) = 1\right)$$

oznacza prawdopodobieństwo odrzucenia H_0. Nazwiemy $1 - \beta(\theta)$ *funkcją mocy* testu δ. Interpretacja tej funkcji jest odmienna dla $\theta \in \Theta_0$ i $\theta \in \Theta_1$:

Jeśli $\theta \in \Theta_0$ to
$1 - \beta(\theta) = \mathbb{P}_\theta\left(\delta(X) = 1\right)$ jest prawdopodobieństwem błędu I rodzaju.

Jeśli $\theta \in \Theta_1$ to
$\beta(\theta) = \mathbb{P}_\theta\left(\delta(X) = 0\right)$ jest prawdopodobieństwem błędu II rodzaju.

Za klasyczną teorią testowania stoi ważna idea metodologiczna. Jest to zasada konserwatyzmu: nie należy rezygnować z ustalonej teorii (hipotezy zerowej), jeśli nie ma po temu koniecznych lub przynajmniej bardzo wyraźnych powodów. Wobec tego staramy się przede wszystkim *kontrolować prawdopodobieństwo błędu I rodzaju*. Interesować nas będą testy, dla których prawdopodobieństwo błędu I rodzaju nie przekracza zadanej z góry, małej liczby. Spośród takich i tylko takich testów postaramy się wybrać te, które mają możliwie małe prawdopodobieństwo błędu II rodzaju.

Jest jeszcze inny powód kontroli błędu I rodzaju. Często w zastosowaniach za H_0 przyjmuje się hipotezę, której błędne odrzucenie ma poważniejsze skutki niż błędne jej przyjęcie. O przykłady takich sytuacji praktycznych nie jest trudno. Wyobraźmy sobie, że zadaniem statystyka jest porównanie skuteczności dwóch leków: A i B. Przypuśćmy, że lek A jest od dawna stosowany, jego działanie i skutki uboczne są dobrze znane. Lek B jest nowy i jego stosowanie wiąże się z pewnym ryzykiem. W takiej sytuacji statystyk powinien za hipotezę zerową przyjąć H_0: „lek A jest nie mniej skuteczny, niż B". W istocie, błąd I rodzaju polegający na pochopnym odrzuceniu H_0 może narazić pacjentów na niebezpieczeństwo – tego staramy się przede wszystkim uniknąć. Błąd II rodzaju może tylko trochę opóźnić postęp w metodach leczenia i jest mniej groźny w skutkach.

7.1.1 DEFINICJA. *Mówimy, że δ jest testem* ***na poziomie istotności*** *α, jeśli*

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta\left(\delta(X) = 1\right) \leqslant \alpha.$$

Za poziom istotności α przyjmuje się zwykle „okrągłą, małą" liczbę, na przykład $\alpha = 0.01$ lub $\alpha = 0.05$. Zgodnie z tym co powiedzieliśmy wcześniej, wybór poziomu istotności odzwierciedla nasz „stopień konserwatyzmu": im bardziej przywiązani jesteśmy do wyjściowej hipotezy H_0, tym mniejsze wybieramy α.

Załóżmy, że test ma postać $\delta(X) = \mathbb{I}(T(X) > c)$. **Poziom krytyczny** testu lub inaczej p-**wartość** jest to n*ajmniejszy poziom istotności przy którym odrzucamy* H_0. Formalnie, jeśli zaobserwujemy $x \in \mathcal{X}$, to

$$p = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(X) > T(x)).$$

7.1.2 DEFINICJA. *Test δ^* jest* ***jednostajnie najmocniejszy*** *na poziomie istotności α, jeśli*

(i) δ^* jest testem na poziomie istotności α;

(ii) dla każdego testu δ na poziomie istotności α, mamy

$$\mathbb{P}_\theta\left(\delta^*(X) = 1\right) \geqslant \mathbb{P}_\theta\left(\delta(X) = 1\right)$$

dla każdego $\theta \in \Theta_1$.

W skrócie, δ^* jest TJNM (hipotezy $\mathrm{H}_0 : \theta \in \Theta_0$ przeciw alternatywie $\mathrm{H}_1 : \theta \in \Theta_1$ na poziomie istotności α).

7.2 Lemat Neymana-Pearsona

Niech θ_0 i θ_1 będą ustalonymi punktami przestrzeni parametrów Θ. Rozważymy dwie *hipotezy proste*:

$$\mathrm{H}_0 : \theta = \theta_0, \qquad \mathrm{H}_1 : \theta = \theta_1.$$

Znaczy to, że $\Theta_0 = \{\theta_0\}$ i $\Theta_1 = \{\theta_1\}$ są zbiorami jednopunktowymi. Niech $\mathbb{P}_0$ i $\mathbb{P}_1$ oznaczają rozkłady prawdopodobieństwa na przestrzeni próbkowej,

odpowiadające wartościom parametru θ_0 i θ_1. Załóżmy, że te rozkłady mają gęstości f_0 i f_1. Niech X oznacza wektor obserwacji. Innymi słowy, obie hipotezy możemy sformułować tak:

$$\mathrm{H}_0 : X \sim f_0, \qquad \mathrm{H}_1 : X \sim f_1.$$

7.2.1 TWIERDZENIE (Lemat Neymana - Pearsona). *Niech*

$$K^* = \left\{ x \in \mathcal{X} : \frac{f_1(x)}{f_0(x)} > c \right\}.$$

Załóżmy, że $\mathbb{P}_0(K^*) = \alpha$ *i* $\mathbb{P}_1(K^*) = \beta$. *Jeśli* $K \subset \mathcal{X}$ *jest takim zbiorem, że* $\mathbb{P}_0(K) \leqslant \alpha$, *to* $\mathbb{P}_1(K) \leqslant \beta$.

Dowód. Ponieważ $\int_{K^*} f_0 \geqslant \int_K f_0$, więc $\int_{K^*\setminus K} f_0 \geqslant \int_{K\setminus K^*} f_0$. Pomnóżmy tę ostatnią nierówność obustronnie przez c i skorzystajmy z faktu, że $f_1 > cf_0$ na zbiorze $K^*\setminus K$ i $cf_0 \geqslant f_1$ na $K\setminus K^*$. Otrzymujemy $\int_{K^*\setminus K} f_1 \geqslant \int_{K\setminus K^*} f_1$ i stąd

$$\int_{K^*} f_1 \geqslant \int_K f_1.$$

Dodajmy, że nierówność jest *ostra*, jeśli tylko $\int_{K^*\setminus K} f_0 > 0$ i $c > 0$. □

Popatrzmy na zbiory K^* i K jak na obszary krytyczne testów δ^* i δ. Z oczywistych względów, δ^* nazywamy *testem ilorazu wiarogodności*. Lemat Neymana - Pearsona stwierdza, że test ilorazu wiarogodności jest *najmocniejszy* na poziomie istotności α (przymiotnik „jednostajnie" możemy opuścić, bo hipoteza alternatywna jest prosta).

7.2.2 Uwaga. Przypuśćmy, że f_1 i f_2 oznaczają gęstości prawdopodobieństwa w „zwykłym" sensie, to znaczy obserwacja X jest zmienną losową „typu ciągłego". Jeśli α jest zadaną z góry liczbą ($0 < \alpha < 1$), to *zazwyczaj* można dobrać poziom krytyczny c testu ilorazu wiarogodności tak, żeby $\mathbb{P}_0(f_1(X)/f_0(X) > c) = \alpha$. Nie będziemy tego stwierdzenia ani uściślać ani dowodzić, ale zobaczymy wkrótce na kilku przykładach jak to się robi.

Z drugiej strony, „gęstości" w Twierdzeniu 7.2.1 mogą oznaczać dyskretne funkcje prawdopodobieństwa: $f_i(x) = \mathbb{P}_i(X = x)$. Dowód jest taki sam, tylko całki trzeba zamienić na sumy. Lemat Neymana - Pearsona pozostaje prawdziwy dla obserwacji typu dyskretnego, ale przy jego stosowaniu pojawiają się drobne trudności. Dla danego α, może nie istnieć c takie, że $\mathbb{P}_0(f_1(X)/f_0(X) > c) = \alpha$. Wtedy najmocniejszy test na poziomie istotności α może nie być testem ilorazu wiarogodności (Zadanie **??**). Pewnym wyjściem jest użycie tak zwanych *testów zrandomizowanych*, w których uzależnia się decyzję od przypadku. W naszych rozważaniach, w szczególności w Definicji 7.1.2, ograniczyliśmy się dla uproszczenia do testów niezrandomizowanych. Dokładniejsza dyskusja wymagałaby więc bardziej ogólnej definicji testu. Nie będziemy się tym zajmować. □

Wspomnijmy, że obszar krytyczny testu ilorazu wiarogodności można (i zazwyczaj wygodnie jest) napisać w równoważnej, „zlogarytmowanej" postaci $K^* = \{x : \log f_1(x) - \log f_0(x) > \log c = \text{const}\}$. Wprowadźmy umowę, że „const" jest „ogólnym" oznaczeniem poziomu krytycznego i może w trakcie rachunków reprezentować *różne* liczby.

7.2.3 PRZYKŁAD (Model normalny, test „jednostronny"). Załóżmy, że mamy próbkę $X_1, \ldots, X_n$ z rozkładu $\mathrm{N}(\mu, \sigma^2)$, gdzie σ^2 jest znane. Rozważmy najpierw dwie hipotezy *proste*: $\mathrm{H}_0 : \mu = \mu_0$ przeciw $\mathrm{H}_1 : \mu = \mu_1$, gdzie $\mu_0 < \mu_1$. Zbudujemy test ilorazu wiarogodności. Niech

$$f_h(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu_h)^2\right],$$

dla $h = 0, 1$. Mamy więc

$$\begin{aligned} \log f_1(X_1, \ldots, X_n) &- \log f_0(X_1, \ldots, X_n) \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_1)^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_0)^2 \\ &= \frac{1}{2\sigma^2}\left[2\sum_{i=1}^{n} X_i(\mu_1 - \mu_0) + n(\mu_0^2 - \mu_1^2)\right]. \end{aligned}$$

Test ilorazu wiarogodności odrzuca H_0 na rzecz H_1, jeśli powyższe wyrażenie przekracza pewną stałą. Ale tak się dzieje wtedy i tylko wtedy, gdy $\bar{X} >$

c dla pewnej (innej) stałej. Dopiero teraz dokończymy naszą konstrukcję, stosownie wybierając c. Niech $c = z\sigma/\sqrt{n}$, gdzie z jest kwantylem rzędu $1-\alpha$ standardowego rozkładu normalnego, czyli $\Phi(z) = 1-\alpha$:

$$\text{odrzucamy } \mathrm{H}_0, \text{ jeśli } \bar{X} > \mu_0 + \frac{z\sigma}{\sqrt{n}}.$$

Niech $1-\beta(\mu)$ będzie funkcją mocy tego testu. Nasz wybór stałej c zapewnia, że $1-\beta(\mu_0) = \alpha$. Z lematu N-P wnioskujemy, że jest to *najmocniejszy* test $\mathrm{H}_0 : \mu = \mu_0$ przeciw $\mathrm{H}_1 : \mu = \mu_1$ na poziomie istotności α. Mówiliśmy na razie o dwóch hipotezach prostych, ale ten sam test jest również TJNM $\mathrm{H}_0 : \mu \leqslant \mu_0$ przeciw $\mathrm{H}_1 : \mu > \mu_0$, na poziomie istotności α. Żeby to uzasadnić, posłużymy się bardzo prostą argumentacją. Jeśli test ma poziom istotności α, to w szczególności $1-\beta(\mu_0) \leqslant \alpha$. Dla *dowolnego* $\mu_1 > \mu_0$, spośród testów spełniających warunek $1-\beta(\mu_0) \leqslant \alpha$, nasz test ma największą moc w punkcie μ_1. Ale to znaczy, że jest on *jednostajnie* najmocniejszym testem przeciw alternatywie $\mu > \mu_0$.

Na koniec przyjrzyjmy się funkcji mocy naszego TJNM:

$$1-\beta(\mu) = 1 - \Phi\left(z - \sqrt{n}\frac{\mu-\mu_0}{\sigma}\right).$$

□

Rodziny z monotonicznym ilorazem wiarogodności. W Przykładzie 7.2.3, konstrukcja TJNM opierała się na spostrzeżeniu, że najmocniejszy test test hipotezy prostej μ_0 przeciw μ_1 na poziomie istotności α jest *taki sam* dla wszystkich alternatyw $\mu_1 > \mu_0$. To rozumowanie można uogólnić. Jeśli mamy rodzinę gęstości $\{f_\theta\}$ taką, że dla wszystkich $\theta_0 < \theta_1$ iloraz $f_{\theta_1}(x)/f_{\theta_0}(x)$ jest rosnącą funkcją pewnej statystyki $T(x)$ to mówimy o rodzinie z *monotonicznym ilorazem wiarogodności*. W zagadnieniu „jednostronnym" $\mathrm{H}_0 : \theta \leqslant \theta_0$ przeciw $\mathrm{H}_1 : \theta > \theta_0$, TJNM na poziomie istotności α ma postać $T(X) > c$, gdzie c jest stałą, dobraną tak żeby $\mathbb{P}_{\theta_0}(T(X) > c) = \alpha$.

Testy nieobciążone. Dla rodziny z monotonicznym ilorazem wiarogodności, w zagadnieniu „dwustronnym" $\mathrm{H}_0 : \theta = \theta_0$ przeciw $\mathrm{H}_1 : \theta \neq \theta_0$, TJNM po prostu *nie istnieje*. Rzeczywiście, test na poziomie istotności α *musi* być postaci $T(X) > c_1$, żeby być najmocniejszy w punkcie $\theta_1 > \theta_0$ i *musi* być postaci $T(X) < c_2$, żeby być najmocniejszy w punkcie $\theta_2 < \theta_0$. Powstaje podobny kłopot jak w teorii estymacji. Mamy do dyspozycji różne testy, z których żaden nie jest (uniwersalnie) najlepszy. Pewnym wyjściem jest ograniczenie rozważań do tak zwanych testów *nieobciążonych*.

7.2.4 DEFINICJA. *Rozważmy zagadnienie testowania* $\mathrm{H}_0 : \theta \in \Theta_0$ *przeciw* $\mathrm{H}_1 : \theta \in \Theta_1$. *Test* δ *nazywamy* ***nieobciążonym****, jeśli dla dowolnych* $\theta_0 \in \Theta_0$ *i* $\theta_1 \in \Theta_1$ *mamy*

$$\mathbb{P}_{\theta_1}(\delta(X) = 1) \geqslant \mathbb{P}_{\theta_0}(\delta(X) = 1).$$

Innymi słowy, dla testu nieobciążonego moc nigdzie nie spada poniżej poziomu istotności. Jest to bardzo naturalne wymaganie.

7.2.5 PRZYKŁAD (Model normalny, test „dwustronny"). W modelu z Przykładu 7.2.3, rozważmy zagadnienie testowania

$$\mathrm{H}_0 : \mu = \mu_0 \text{ przeciw } \mathrm{H}_1 : \mu \neq \mu_0.$$

TJNM, jak wiemy, *nie istnieje*. Rozważmy test

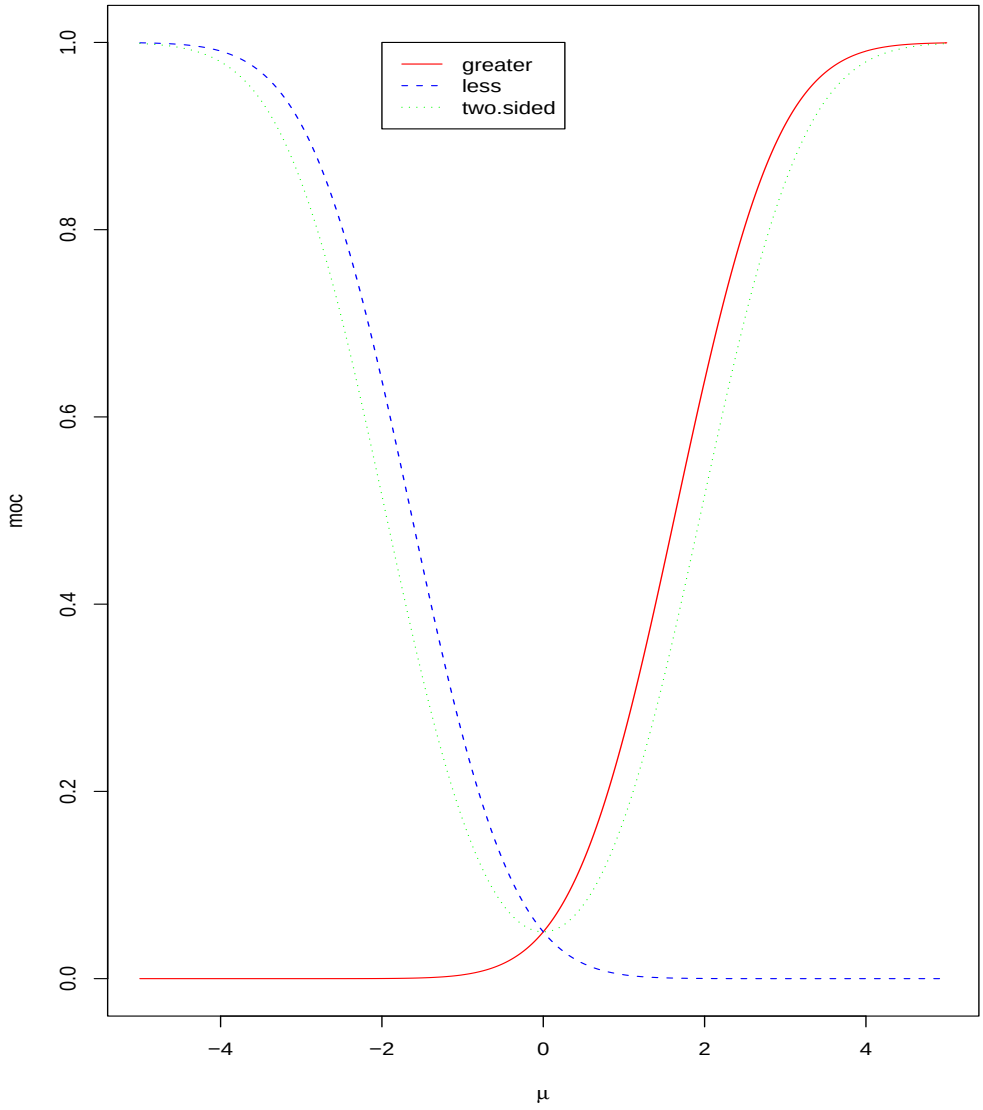
$$\text{odrzucamy } \mathrm{H}_0, \text{ jeśli } |\bar{X} - \mu_0| > \frac{z\sigma}{\sqrt{n}},$$

gdzie z jest kwantylem rzędu $1 - \alpha/2$, czyli $\Phi(z) = 1 - \alpha/2$. Funkcja mocy rozważanego teraz testu jest następująca:

$$1 - \beta(\mu) = 1 - \Phi\left(z - \sqrt{n}\frac{\mu - \mu_0}{\sigma}\right) + \Phi\left(-z - \sqrt{n}\frac{\mu - \mu_0}{\sigma}\right).$$

Zauważmy, że $1 - \beta(\mu_0) = \alpha$ i $1 - \beta(\mu) > \alpha$ dla $\mu \neq \mu_0$. Zatem nasz test jest testem *nieobciążonym* $\mathrm{H}_0 : \mu \leqslant \mu_0$ przeciw $\mathrm{H}_1 : \mu > \mu_0$ na poziomie istotności α. W istocie, można pokazać więcej: jest to test najmocniejszy *spośród testów nieobciążonych* na ustalonym poziomie istotności. □

Funkcje mocy dwóch testów jednostronnych $H_0 : \mu = 0$ (przeciw $H_1 : \mu > 0$ i przeciw $H_1 : \mu < 0$, Przykład 7.2.3) oraz testu dwustronnego (przeciw $H_1 : \mu \neq 0$, Przykład 7.2.5) są pokazane na rysunku. Przyjęty poziom istotności jest równy $\alpha = 0.05$.



7.3 Parametryczne testy istotności

Rozpatrzmy **model normalny**. Zakładamy, że $X_1, \ldots, X_n$ jest próbką z rozkładu $\mathrm{N}(\mu, \sigma^2)$ z nieznanymi paramerami μ i σ. Przedstawimy kilka typowych testów w tym modelu. Będziemy rozpatrywali *złożone* hipotezy zerowe i przedstawimy testy, które nie są jednostajnie najmocniejsze, bo TJNM po prostu nie istnieją. Estymatory $\bar{X}$ i S^2 oznaczają to, co zwykle.

Test Studenta

Test $\mathrm{H}_0 : \mu = \mu_0$ przeciwko $\mathrm{H}_1 : \mu > \mu_0$, gdzie μ_0 jest ustaloną liczbą. Na poziomie istotności α, odrzucamy H_0, gdy

$$\sqrt{n}\frac{\bar{X} - \mu_0}{S} > t, \qquad t = \mathrm{t}_{1-\alpha}(n-1).$$

Test $\mathrm{H}_0 : \mu = \mu_0$ przeciwko $\mathrm{H}_1 : \mu \neq \mu_0$, gdzie μ_0 jest ustaloną liczbą. Na poziomie istotności α, odrzucamy H_0, gdy

$$\sqrt{n}\frac{|\bar{X} - \mu_0|}{S} > t, \qquad t = \mathrm{t}_{1-\alpha/2}(n-1).$$

Zajmiemy się teraz problemem porównania populacji, z których pochodzą dwie (niezależne) próbki. Niech $X_1, \ldots, X_n \sim \mathrm{N}(\mu_X, \sigma_X^2)$ i $Y_1, \ldots, Y_m \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$. Znaczenie symboli $\bar{X}$, $\bar{Y}$, S_X^2 i S_Y^2 jest oczywiste (i takie samo, jak w Przykładzie 2.2.7).

Dwupróbkowy test Studenta.

Test $\mathrm{H}_0 : \mu_X = \mu_Y$ przeciwko $\mathrm{H}_1 : \mu_X > \mu_Y$. *Zakładamy*, że $\sigma_X^2 = \sigma_Y^2$. Testujemy więc hipotezę o równości wartości oczekiwanych, nie kwestionując założenia o równości wariancji. Odrzucamy H_0, gdy

$$\frac{\bar{X} - \bar{Y}}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}}\sqrt{\frac{nm}{n+m}(n+m-2)} > t, \qquad t = \mathrm{t}_{1-\alpha}(n+m-2).$$

Pozostawiamy Czytelnikowi oczywistą modyfikację procedury testowania, gdy alternatywa jest dwustronna: $\mathrm{H}_0 : \mu_X = \mu_Y$ przeciwko $\mathrm{H}_1 : \mu_X \neq \mu_Y$.

Testy dotyczące wariancji

Test $\mathrm{H}_0 : \sigma = \sigma_0$ przeciwko $\mathrm{H}_1 : \sigma > \sigma_0$, gdzie σ_0 jest ustaloną liczbą. Odrzucamy H_0, gdy

$$\frac{n-1}{\sigma_0^2} S^2 > c, \qquad c = \chi^2_{1-\alpha}(n-1).$$

Test $\mathrm{H}_0 : \sigma = \sigma_0$ przeciwko $\mathrm{H}_1 : \sigma \neq \sigma_0$, gdzie σ_0 jest ustaloną liczbą. Odrzucamy H_0, gdy

$$\frac{n-1}{\sigma_0^2} S^2 > c_2 \text{ lub } \frac{n-1}{\sigma_0^2} S^2 < c_1, \qquad c_1 = \chi^2_{\alpha/2}(n-1), c_2 = \chi^2_{1-\alpha/2}(n-1).$$

Zajmiemy się teraz problemem porównania populacji, z których pochodzą dwie (niezależne) próbki. Niech $X_1, \ldots, X_n \sim \mathrm{N}(\mu_X, \sigma_X^2)$ i $Y_1, \ldots, Y_m \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$. Znaczenie symboli $\bar{X}$, $\bar{Y}$, S_X^2 i S_Y^2 jest oczywiste (i takie samo, jak w Przykładzie 2.2.7).

Test $\mathrm{H}_0 : \mu_X = \mu_Y$ przeciwko $\mathrm{H}_1 : \mu_X > \mu_Y$. *Zakładamy*, że $\sigma_X^2 = \sigma_Y^2$. Testujemy więc hipotezę o równości wartości oczekiwanych, nie kwestionując założenia o równości wariancji. Odrzucamy H_0, gdy

$$\frac{\bar{X} - \bar{Y}}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}} \sqrt{\frac{nm}{n+m}(n+m-2)} > t, \qquad t = \mathrm{t}_{1-\alpha}(n+m-2).$$

Pozostawiamy Czytelnikowi oczywistą modyfikację procedury testowania, gdy alternatywa jest dwustronna: $\mathrm{H}_0 : \mu_X = \mu_Y$ przeciwko $\mathrm{H}_1 : \mu_X \neq \mu_Y$. □

7.4 Test ilorazu wiarogodności

Rozważmy dwa modele statystyczne (na tej samej przestrzeni obserwacji). W modelu pierwszym mamy do czynienia z rodziną rozkładów prawdopodobieństwa o gęstościach $f_0(\theta_0, x)$, gdzie $\theta_0 \in \Theta_0$, zaś w modelu drugim mamy gęstości $f_1(\theta_1, x)$, gdzie $\theta_1 \in \Theta_1$. Chcemy zdecydować, który z dwóch konkurujących modeli wybrać do opisu doświadczenia losowego, w którym pojawiła się obserwacja X. Zgodnie z klasycznym schematem testowania

hipotez, oba modele nie będą traktowane równoprawnie: pierwszy z nich odgrywa rolę „hipotezy zerowej", z której nie mamy ochoty rezygnować bez wyraźnej konieczności. Rozważamy zatem zagadnienie testowania

$$\mathrm{H}_0 : X \sim f_0(\theta_0, \cdot) \textit{ dla pewnego } \theta_0 \in \Theta_0$$

przeciw

$$\mathrm{H}_1 : X \sim f_1(\theta_1, \cdot) \textit{ dla pewnego } \theta_1 \in \Theta_1.$$

W istocie, jesteśmy w podobnej sytuacji jak w Lemacie Neymana - Pearsona z tą różnicą, że dwa konkurencyjne modele są *statystyczne*, a nie *probabilistyczne*. Modele zawierają nieznane parametry (θ_0 i θ_1). Rzecz jasna, to bardzo komplikuje sprawę i nie możemy bezpośrednio odwołać się do Lematu N-P. Spróbujemy naśladować ideę testu ilorazu wiarogodności, z koniecznymi modyfikacjami. Za statystykę testową przyjmiemy

$$\lambda = \frac{\sup_{\theta_1 \in \Theta_1} f_1(\theta_1, X)}{\sup_{\theta_0 \in \Theta_0} f_0(\theta_0, X)}.$$

Innymi słowy,

$$\lambda = \frac{f_1(\hat{\theta}_1, X)}{f_0(\hat{\theta}_0, X)},$$

gdzie $\hat{\theta}_0 = \hat{\theta}_0(X)$ i $\hat{\theta}_1 = \hat{\theta}_1(X)$ są *estymatorami największej wiarogodności* odpowiednio w pierwszym i drugim modelu. Odrzucamy H_0, jeśli $\lambda > c$, gdzie c jest pewną stałą. Konstrukcję tego testu można nieformalnie uzasadnić tak: porównujemy „największą szansę otrzymania obserwacji X, gdy prawdziwa jest hipoteza H_1" i „największą szansę, gdy prawdziwa jest hipoteza H_0". Dostatecznie duża wartość ilorazu tych największych szans sugeruje odrzucenie H_0 na rzecz H_1. Nasze rozważania są tak ogólne i abstrakcyjne, że niewiele można powiedzieć o własnościach testu. Nie możemy podać ogólnego sposobu wyznaczania stałej c tak, żeby test był na założonym poziomie istotności. Czasami to może być bardzo trudne, chociaż w niektórych konkretnych sytuacjach daje się zrobić.

7.4.1 PRZYKŁAD (Rozkład wykładniczy, czy rozkład Gamma?). Chcemy rozstrzygnąć, czy próbka $X_1, \ldots, X_n$ pochodzi z rozkładu wykładniczego $\mathrm{Ex}(\theta) = \mathrm{Gamma}(1, \theta)$, czy raczej z rozkładu $\mathrm{Gamma}(2, \theta)$. W obu modelach, θ odgrywa rolę nieznanego „parametru skali". Jeśli $f(x)$ jest gęstością (pojedynczej) obserwacji, to nasze zagadnienie możemy zapisać tak: testujemy hipotezę zerową

$$\mathrm{H}_0 : f(x) = f_0(\theta, x) = \theta \mathrm{e}^{-\theta x} \text{ dla pewnego } \theta > 0,$$

przeciw alternatywie

$$\mathrm{H}_1 : f(x) = f_1(\theta, x) = \theta^2 x \mathrm{e}^{-\theta x} \text{ dla pewnego } \theta > 0.$$

Dla H_0, czyli w modelu wykładniczym, $\mathrm{ENW}(\theta) = \hat{\theta}_0 = 1/\bar{X}$. Dla H_1, w alternatywnym modelu Gamma, $\mathrm{ENW}(\theta) = \hat{\theta}_1 = 2/\bar{X}$. Zatem

$$\begin{aligned} \lambda &= \frac{\prod_i f_1(\hat{\theta}_1, X_i)}{\prod_i f_0(\hat{\theta}_0, X_i)} = \frac{\prod_i (\hat{\theta}_1^2 X_i \exp\left[-\hat{\theta}_1 X_i\right]}{\prod_i (\hat{\theta}_0 \exp\left[-\hat{\theta}_0 X_i\right]} \\ &= \mathrm{const} \prod_i \frac{X_i}{\sum_k X_k}. \end{aligned}$$

Oznaczmy $S_n = \sum_i X_i$. Test ma postać:

wybieramy model $\mathrm{Gamma}(2, \theta)$, jeśli $\prod_i (X_i/S_n) > c$.

Jeśli ustalimy poziom istotności α, to można dobrać stałą c tak, żeby w modelu wykładniczym równość $\mathbb{P}_\theta \left(\prod_i (X_i/S_n) > c\right) = \alpha$ zachodziła *dla każdego* θ. Statystyka testowa ma *dla dowolnego* θ taki sam rozkład prawdopodobieństwa jak dla $\theta = 1$. Jest tak dlatego, że θ jest „parametrem skali". Znaczy to tyle: jeśli $X_i \sim \mathrm{Ex}(\lambda)$ to $Y_i = \lambda X_i \sim \mathrm{Ex}(1)$. Ale $\prod (X_i/\sum X_k) = \prod (Y_i/\sum Y_k)$. Statystyka testowa, policzona dla X-ów (czyli dla próbki z rozkładu $\mathrm{Ex}(\lambda)$) ma tę samą wartość, co dla Y-ów (dla próbki z rozkładu $\mathrm{Ex}(1)$). Z tego co powiedzieliśmy, nie wynika, że wyznaczenie stałej c w zależności od α jest numerycznie łatwe. Jest jednak w zasadzie proste. Wystarczy znaleźć rozkład statystyki testowej przy założeniu, że $X_i \sim \mathrm{Ex}(1)$ i za c wziąć kwantyl rzędu $1 - \alpha$ tego rozkładu. □

Modele „zagnieżdżone". Rozważmy model statystyczny opisany przez rodzinę gęstości prawdopodobieństwa $\{f_\theta : \theta \in \Theta\}$ i „zanurzony" w nim mniejszy model $\{f_\theta : \theta \in \Theta_0\}$, gdzie $\Theta_0 \subset \Theta$. Interesuje nas pytanie: czy możemy założyć, że mniejszy model Θ_0 opisuje poprawnie doświadczenie losowe, czy potrzebny jest większy model Θ? Formalnie, zagadnienie polega na zbudowaniu testu hipotezy $\mathrm{H}_0 : \theta \in \Theta_0$ przeciw alternatywie $\mathrm{H}_1 : \theta \in \Theta \backslash \Theta_0 = \Theta_1$. Możemy wykorzystać test ilorazu wiarogodności opisany w poprzednim punkcie. Skupimy się na najbardziej typowej sytuacji, kiedy mniejszy model jest określony przez *nałożenie dodatkowych ograniczeń w postaci układu równań.* Załóżmy, że $\Theta \subset \mathbb{R}^d$, dana jest funkcja $h : \mathbb{R}^d \to \mathbb{R}^p$ i $\Theta_0 = \{\theta \in \Theta : h(\theta) = 0\}$. Zauważmy, że równanie $h(\theta) = 0$ jest faktycznie układem p równań z d „niewiadomymi": współrzędnymi wektora $\theta = (\theta_1, \ldots, \theta_d)$. Jest raczej jasne, że w „typowej" sytuacji Θ_0 jest podzbiorem „$d-p$-wymiarowym" zbioru „d-wymiarowego" Θ. W konsekwencji, zazwyczaj mamy $\sup_{\theta\in\Theta_1} f(\theta, X) = \sup_{\theta\in\Theta} f(\theta, X)$. Możemy statystykę ilorazu wiarogodności napisać w nieco prostszej postaci

$$\lambda = \frac{\sup_{\theta\in\Theta} f(\theta, X)}{\sup_{\theta\in\Theta_0} f(\theta, X)}.$$

W liczniku mamy maksimum *bezwarunkowe*, po całej przestrzeni parametrów Θ. To maksimum jest osiągane w punkcie $\hat{\theta}$, czyli w $\mathrm{ENW}(\theta)$ obliczonym dla większego modelu. W mianowniku jest maksimum *warunkowe.* Punkt, w którym to maksimum jest osiągane oznaczamy przez $\dot{\theta}$ i nazywamy *estymatorem największej wiarogodności z ograniczeniami.* Oczywiście $\dot{\theta}$ jest $\mathrm{ENW}(\theta)$ w mniejszym modelu i mamy $h(\dot{\theta}) = 0$. Reasumując, w zagadnieniu

$$\mathrm{H}_0 : h(\theta) = 0 \text{ przeciw } \mathrm{H}_1 : h(\theta) \neq 0.$$

statystyka ilorazu wiarogodności przybiera postać

$$\lambda = \frac{f(\hat{\theta}, X)}{f(\dot{\theta}, X)},$$

gdzie

$$f(\hat{\theta}) = \sup_{\theta} f(\theta, X),$$
$$f(\dot{\theta}) = \sup_{\theta: h(\theta)=0} f(\theta, X), \qquad (h(\dot{\theta}) = 0).$$

7.4.2 PRZYKŁAD (Zagadnienie dwóch próbek dla rozkładu wykładniczego). Mamy dwie niezależne próbki: $X_1, \ldots, X_n \sim \mathrm{Ex}(\theta_X)$ i $Y_1, \ldots, Y_n \sim \mathrm{Ex}(\theta_Y)$. Zbudujemy test ilorazu wiarogodności $\mathrm{H}_0 : \theta_X = \theta_Y$ przeciw $\mathrm{H}_1 : \theta_X \neq \theta_Y$. Parametrem jest para $\theta = (\theta_X, \theta_Y)$ Przestrzeń parametrów jest „ćwiartką" płaszczyzny: $\Theta =]0, \infty[^2$. To jest „duży model". Hipoteza zerowa wyznacza „mały model": zbiór $\Theta_0 = \{(\theta_X, \theta_Y) : \theta_X = \theta_Y\}$. Wiarogodność jest określona wzorem

$$\begin{aligned} f_{\theta_X,\theta_Y}(x_1, \ldots, x_n, y_1, \ldots, y_n) &= \prod \theta_X \mathrm{e}^{-\theta_X x_i} \theta_Y \mathrm{e}^{-\theta_Y y_i} \\ &= (\theta_X \theta_Y)^n \exp\left[-\theta_X \sum x_i - \theta_Y \sum y_i\right]. \end{aligned}$$

ENW bez ograniczeń i ENW z ograniczeniami są takie:

$$\hat{\theta} = \left(\frac{1}{\bar{X}}, \frac{1}{\bar{Y}}\right), \qquad \dot{\theta} = \left(\frac{2}{\bar{X}+\bar{Y}}, \frac{2}{\bar{X}+\bar{Y}}\right).$$

To po prostu dlatego, że w „dużym" modelu mamy dwie niezwiązane ze sobą próbki, a w „małym" modelu – jedną, połączoną próbkę z tego samego rozkładu. W obu przypadkach skorzystaliśmy z prostych wzorów na ENW dla modelu wykładniczego. Wprowadźmy oznaczenie $\bar{Z} = (\bar{X}+\bar{Y})/2$. Statystyka ilorazu wiarogodności jest równa

$$\begin{aligned} \lambda &= \frac{(\bar{X}\bar{Y})^{-n} \exp\left[-\bar{X}^{-1} \sum X_i - \bar{Y}^{-1} \sum Y_i\right]}{(\bar{Z})^{-2n} \exp\left[-\bar{Z}^{-1} \sum X_i - \bar{Z}^{-1} \sum Y_i\right]} \\ &= \frac{\bar{Z}^{2n}}{\bar{X}^n \bar{Y}^n} = \left(\frac{\bar{Z}}{\bar{X}} \frac{\bar{Z}}{\bar{Y}}\right)^n. \end{aligned}$$

Test jest zatem takiej postaci:

$$\text{odrzucamy } \mathrm{H}_0, \text{ jeśli } 4R(1-R) < c,$$

gdzie $R = \bar{X}/(\bar{X}+\bar{Y})$, zaś c jest pewną stałą. Łatwe jest dobranie wartości krytycznej tak, żeby otrzymać test na poziomie istotności α. Można wykorzystać fakt, że przy prawdziwości hipotezy zerowej mamy $R \sim \mathrm{Beta}(n,n)$ i kwantyle rozkładu Beta są łatwo dostępne.

Rozkład asymptotyczny

Tak jak w punkcie poprzednim, rozpatrujemy *dwa modele zagnieżdżone* i zagadnienie testowania

$$\mathrm{H}_0 : h(\theta) = 0 \text{ przeciw } \mathrm{H}_1 : h(\theta) \neq 0.$$

Zakładamy, że funkcja $h : \mathbb{R}^d \to \mathbb{R}^p$ jest „dostatecznie porządna", $\Theta \subset \mathbb{R}^d$ jest „zbiorem d-wymiarowym" i $\Theta_0 = \{\theta \in \Theta : h(\theta) = 0\}$ jest „zbiorem o wymiarze $d-p$":

$$\dim\Theta = d, \qquad \dim\Theta_0 = d - p < d.$$

Nie będziemy się starali uściślić tych sformułowań. W większości konkretnych przypadków i tak wiadomo, o co chodzi. Przy pewnych dodatkowych założeniach, prawdziwy jest następujący fakt.

7.4.3 TWIERDZENIE. *Jeśli* H_0 *jest prawdziwa, to przy* $n \to \infty$ *rozkład statystyki* $2\log\lambda$ *zmierza do rozkładu* $\chi^2(p)$ *(chi-kwadrat z* p *stopniami swobody).*

Mnemotechniczna regułka jest taka:

liczba stopni swobody = liczba ograniczeń = $\dim\Theta - \dim\Theta_0$.

Nie podamy dowodu w pełnej ogólności, ale naszkicujemy rozumowanie dla pewnego przypadku szczególnego. Rozważmy mianowicie model *probabilistyczny* zanurzony w większym modelu *statystycznym*. Testowanie $\mathrm{H}_0 : \theta = \theta_0$ przeciw $\mathrm{H}_1 : \theta \neq \theta_0$ mieści się w rozpatrywanym schemacie modeli zagnieżdżonych, z *jednopunktowym* zbiorem $\Theta_0 = \{\theta_0\}$. Możemy uważać, że θ_0 jest rozwiązaniem układu równań $h(\theta) = 0$, gdzie $h : \mathbb{R}^d \to \mathbb{R}^d$ (d równań dla d parametrów). Oczywiście, „estymator z ograniczeniami" jest równy $\dot{\theta} = \theta_0$ (gdy „mały" model zawiera tylko jeden punkt, to nie ma co estymować). Faktycznie testy „dwustronne" w modelu normalnym, o których mówiliśmy w poprzednim rozdziale można otrzymać w ten właśnie sposób. Zamiast wracać do tych testów, rozpatrzymy nowy, prosty przykład.

7.4.4 PRZYKŁAD. Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu wykładniczego $\mathrm{Ex}(\theta)$ z nieznanym parametrem θ. Chcemy przeprowadzić test hipotezy

$\mathrm{H}_0 : \theta = 1$ przeciw $\mathrm{H}_1 : \theta \neq 1$.

ENW bez ograniczeń jest, oczywiście, równy $\hat{\theta} = 1/\bar{X}$. Zatem

$$\begin{aligned}\lambda &= \frac{\prod f_{\hat{\theta}}(X_i)}{\prod f_1(X_i)} = \frac{\bar{X}^{-n}\exp\left[-\bar{X}^{-1}\sum X_i\right]}{\exp\left[-\sum X_i\right]} \\ &= \bar{X}^{-n}\exp\left[-n(\bar{X}-1)\right].\end{aligned}$$

$$\begin{aligned}\lambda &= \frac{\prod f_{\hat{\theta}}(X_i)}{\prod f_1(X_i)} = \frac{\bar{X}^{-n}\exp\left[-\bar{X}^{-1}\sum X_i\right]}{\exp\left[-\sum X_i\right]} \\ &= \bar{X}^{-n}\exp\left[-n(\bar{X}-1)\right].\end{aligned}$$

Do wyznaczenia wartości krytycznej przybliżonego testu na poziomie istotności α można skorzystać z granicznego rozkładu statystyki $2\log\lambda$. Wiemy z Twierdzenia 7.4.3, że jest to rozkład $\chi^2(1)$. Można zresztą ten fakt udowodnić niezależnie (Zadanie 6). □

Szkic dowodu Twierdzenia 7.4.3. Ograniczymy się do przypadku $d = 1$ i $p = 1$. Przyjmijmy, że spełnione są założenia Twierdzenia 4.3.1 o asymptotycznej normalności $\hat{\theta} = \mathrm{ENW}(\theta)$.

Podobnie jak w dowodzie Twierdzenia 4.3.1, napiszemy $l'(\theta) = \sum_1^n (\log f)'(\theta, X_i)$ i wykorzystamy rozwinięcie Taylora, tym razem dla funkcji $l(\theta)$:

$$\log \lambda = l(\hat{\theta}) - l(\theta_0) \simeq l'(\theta_0)(\hat{\theta} - \theta_0) + \frac{1}{2} l''(\theta_0)(\hat{\theta} - \theta_0)^2.$$

Powołamy się teraz na przybliżoną równość $\hat{\theta} - \theta_0 \simeq l'(\theta_0)/l''(\theta_0)$. Otrzymujemy

$$\log \lambda \simeq -\frac{1}{2} \frac{(l'(\theta_0))^2}{l''(\theta_0)} = \frac{1}{2} \frac{(\sum (\log f)'(\theta_0, X_i)/\sqrt{n})^2}{\sum (\log f)''(\theta_0, X_i)/n}. \qquad (*)$$

Zarówno $l'(\theta_0)$, jak i $l''(\theta_0)$ są sumami niezależnych zmiennych losowych o jednakowym rozkładzie. Zastosujmy CTG do licznika i PWL do mianownika we wzorze $(*)$: mamy $l'(\theta_0)/\sqrt{n} \to_{\mathrm{d}} \mathrm{N}(0, I_1(\theta_0))$ i $l''(\theta_0)/n \to_{\mathrm{P}} -I_1(\theta_0)$, tak jak w dowodzie Twierdzenia 4.3.1. W rezultacie $2 \log \lambda \to_{\mathrm{d}} \chi^2(1)$, co należało wykazać. □

7.5 Zgodność testów

Zgodność jest podstawowym pojęciem, dotyczącym asymptotycznych własności testów. Rozważmy zagadnienie testowania w najogólniejszej postaci:

$$\mathrm{H}_0 : \theta \in \Theta_0 \text{ przeciw } \mathrm{H}_1 : \theta \in \Theta_1.$$

Niech $X_1, \ldots, X_n, \ldots$ będzie nieskończonym ciągiem obserwacji o rozkładzie $\mathbb{P}_\theta$ i $\delta_n = \delta(X_1, \ldots, X_n)$ będzie ciągiem testów takich, że δ_n zależy od początkowych n obserwacji. Myślimy o $\delta(X_1, \ldots, X_n)$ jak o *jednym* teście, stosowanym do próbek o różnej liczności. Pod tym względem sytuacja jest podobna jak w Rozdziale 3, gdzie interesowaliśmy się granicznymi własnościami estymatorów.

7.5.1 DEFINICJA. *Mówimy, że test δ_n jest* ***zgodny*** *na poziomie istotności α, jeśli*

(i) $\lim_{n\to\infty}\mathbb{P}_\theta(\delta_n = 1) \leqslant \alpha$ *dla każdego* $\theta \in \Theta_0$*;*

(ii) $\lim_{n\to\infty}\mathbb{P}_\theta(\delta_n = 1) = 1$ *dla każdego* $\theta \in \Theta_1$.

Innymi słowy, test jest zgodny, jeśli przy zwiększającej się próbce utrzymuje się ustalony poziom istotności, a moc dąży do jedności. Chodzi o to, że test „w końcu z całą pewnością" odrzuci hipotezę zerową, jeśli jest fałszywa. Sens określenia zgodności dla testów i dla estymatorów jest podobny. Zgodny estymator „w końcu z całą pewnością" identyfikuje nieznany parametr. Zgodność jest raczej słabym, minimalnym wymaganiem (przynajmniej w sytuacji, gdy obserwacje $X_1, \ldots, X_n, \ldots$ są niezależnymi zmiennymi losowymi o jednakowym rozkładzie). Zadanie 1 dotyczy zgodności testów rozważanych w Przykładach 7.2.3 i 7.2.5. Oto inny, nieco mniej oczywisty przykład.

7.5.2 PRZYKŁAD (Zgodność testu Kołmogorowa). Zakładamy, że ciąg obserwacji $X_1, \ldots, X_n, \ldots$ jest próbką z rozkładu o nieznanej dystrybuancie F. Niech F_0 będzie ustaloną dystrybuantą. Rozważamy zagadnienie testowania

$$\mathrm{H}_0 : F = F_0 \text{ przeciw } \mathrm{H}_1 : F \neq F_0.$$

Napiszemy statystykę Kołmogorowa w postaci $D_n = D_n(F_0; X_1, \ldots, X_n) = \sup_x |\sum \mathbb{I}(X_i \leqslant x)/n - F_0(x)|$. Test na poziomie istotności α odrzuca H_0 jeśli $\sqrt{n}D_n > c$ (powiedzmy, że wartość krytyczną ustalamy przybliżoną metodą opartą na Twierdzeniu 6.2.12). Zgodność testu Kołmogorowa znaczy tyle, że dla $F \neq F_0$,

$$\mathbb{P}_F\left(\sqrt{n}D_n(F_0; X_1, \ldots, X_n)\right) \to 1, \qquad (n \to \infty).$$

Symbol $\mathbb{P}_F$ znaczy, że $X_1, \ldots, X_n \sim F$. Zauważmy, że Definicja 7.5.1 pracuje tu w niezwykle abstrakcyjnej sytuacji: za „parametr" rozkładu prawdopodobieństwa uważamy tu jego *dystrybuantę*. Zgodność testu Kołmogorowa wynika z twierdzenia Gliwienki-Cantelliego (Twierdzenie 1.1.7). □

7.6 Zadania

1. Obliczyć granicę $1-\beta(\mu)$ przy $n \to \infty$ dla testów rozważanych w Przykładach 7.2.3 i 7.2.5. Wskazać związek orzymanego wyniku z pojęciem zgodności testu.

2. Uzasadnić dokładnie fakt wspomniany w Przykładzie 7.5.2: zgodność testu Kołmogorowa wynika z twierdzenia Gliwienki-Cantelliego (Twierdzenie 1.1.7).

3. Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu $\mathrm{N}(\mu, \sigma^2)$ ze *znaną wartością oczekiwaną* μ. Pokazać, że TJNM $\mathrm{H}_0 : \sigma \leqslant \sigma_0$ przeciw $\mathrm{H}_1 : \sigma > \sigma_0$ na poziomie istotności α ma postać $\sum (X_i - \mu)^2 > c$. Jak wybrać poziom krytyczny c?

4. Wyprowadzić test ilorazu wiarogodności dla dwóch hipotez prostych: $\mathrm{H}_0 : f(x) = \mathrm{e}^{-x}$ przeciw $\mathrm{H}_1 : f(x) = x\mathrm{e}^{-x}$. Zauważmy, że tutaj staramy się rozpoznać, czy próbka pochodzi z rozkładu $\mathrm{Ex}(1) = \mathrm{Gamma}(1,1)$, czy z rozkładu $\mathrm{Gamma}(2,1)$. Porównać z Przykładem 7.4.1.

5. Wykonać symulacje komputerowe, które pozwolą wyznaczyć wartość krytyczną c dla testu w Przykładzie 7.4.1. Przyjąć $\alpha = 0.1$ i $n = 10$. Wygenerować dużą próbkę z rozkładu statystyki testowej i przyjąć za c odpowiedni kwantyl empiryczny.

6. Wyprowadzić bez odwoływania się do Twierdzenia 7.4.3 asymptotyczny rozkład statystyki testowej w Przykładzie 7.4.4 przy prawdziwej hipotezie H_0. Innymi słowy pokazać, że jeśli $X_1, \ldots, X_n \sim \mathrm{Ex}(1)$, to

$$2n \left[(\bar{X} - 1) - \log \bar{X} \right] \to_{\mathrm{d}} \chi^2(1), \qquad (n \to \infty).$$

   Wskazówka: Z CTG wynika, że zmienna losowa $\bar{X} - 1$ ma w przybliżeniu taki rozkład prawdopodobieństwa jak $Z/\sqrt{n}$, gdzie $Z \sim \mathrm{N}(0,1)$. Użyć rozwinięcia Taylora, odrzucając wyrazy zmierzające do zera.

Część IV

Regresja

Rozdział 8

Regresja liniowa

8.1 Wstęp

Modele regresji zajmują szczególne miejsce w statystyce. Mają niebywałą ilość różnorodnych zastosowań. Używa się ich powszechnie w chemii, biologii, ekonomii, doświadczalnictwie rolniczym i właściwie w każdej z nauk empirycznych. Z konieczności ograniczymy się do paru najprostszych modeli i nasza dyskusja będzie bardzo pobieżna. Regresja opisuje, mówiąc najogólniej, statystyczną zależność tak zwanej zmiennej „objaśnianej" od zmiennych „objaśniających".

Przypuśćmy, że interesuje nas związek pomiędzy dwiema zmiennymi, które oznaczymy przez x i Y. Mierzymy lub obserwujemy wielokrotnie wartości tych zmiennych. Dane mają postać par (x_i, Y_i) i możemy je zapisać w takiej tabelce:

przypadki \ zmienne	niezależna (objaśniająca) x	zależna (objaśniana) Y
1	x_1	Y_1
2	x_2	Y_2
$\vdots$	$\vdots$	$\vdots$
n	x_n	Y_n

Na przykład, możemy badać zależność pomiędzy parami zmiennych (x, Y) takiego typu:

x		Y
wielkość produkcji	–	zużycie energii
wiek dziecka	–	wzrost
stężenie katalizatora	–	wydajność procesu
dawka nawozu	–	plony
...	–	...

„Przypadki" odpowiadają pomiarom lub obserwacjom zmiennej Y dla różnych wartości zmiennej x. Poszczególne pomiary mogą dotyczyć różnych obiektów lub tego samego, ewoluującego procesu.

Przypuszczamy, że zmienna Y jest „w zasadzie" funkcją x, ale „zaburzoną losowymi błędami". Nasz model zależności będzie taki:

$$Y = \phi(x) + \varepsilon,$$

gdzie ε jest *błędem losowym*. Funkcję $y = \phi(x)$ nazywamy *funkcją regresji*. Dla poszczególnych „przypadków", czyli uzyskanych doświadczalnie punktów (x_i, Y_i) mamy

$$Y_i = \phi(x_i) + \varepsilon_i, \qquad (i = 1, \ldots, n).$$

Punkty doświadczalne (x_i, Y_i) nie leżą dokładnie na krzywej regresji, ale znajdują się „w pobliżu" wykresu funkcji $y = \phi(x)$. Zakładamy, że wielkości x_i są znane i nielosowe. Odpowiada to sytuacji, gdy zmienna x jest „pod kontrolą eksperymentatora" i jest mierzona bezbłędnie. Wartości zmiennej Y są losowymi obserwacjami (ze względu na wpływ losowego składnika ε). Funkcja regresji ϕ jest nieznana i będziemy ją estymować na podstawie danych.

Wspomnimy później o nieco innnym modelu regresji, w którym zmienna objaśniająca też jest losowa. Na razie jednak pozostaniemy przy napisanych wyżej założeniach. *Nota bene*, oznaczenie zmiennej niezależnej *małą* literą x, a zmiennej zależnej *dużą* literą Y ma nam stale przypominać, gdzie „tkwi losowość". Czytelnik powinien wiedzieć, że w literaturze ta konwencja nie jest powszechnie przyjęta.

Metoda najmniejszych kwadratów

Sprecyzowanie modelu regresji wymaga przyjęcia konkretnych założeń o funkcji ϕ oraz o błędach losowych ε_i. Założymy, że funkcja regresji ma znaną postać, natomiast zależy od nieznanego parametru β. Napiszemy zatem $\phi(x) = \phi(\beta, x)$. Zwrócmy uwagę, że wartość β dla poszczególnych „przypadków" $i = 1, \ldots, n$ jest taka sama (zależność opisuje jedna funkcja, tylko błędy losowe są różne). W ten sposób powstają *parametryczne* modele regresji[1].

Przyjmiemy klasyczne założenie, że błędy są niezależne i mają jednakowy rozkład normalny. Podsumujmy i uzupełnijmy opis modelu:

$$Y_i = \phi(\beta, x_i) + \varepsilon_i, \qquad (i = 1, \ldots, n) \tag{8.1.1}$$

gdzie

i - numer „przypadku",

x_i - wartość zmiennej „objaśniającej" (znana i nielosowa),

ε_i - błąd losowy (nieobserwowana zmienna losowa),

Y_i - obserwowana zmienna losowa „objaśniana",

β - nieznany parametr (nielosowy).

8.1.2 Założenie. *Spełniona jest zależność* (8.1.1). *Błędy* $\varepsilon_1, \ldots, \varepsilon_n$ *są niezależnymi zmiennymi losowymi o jednakowym rozkładzie normalnym* $\mathrm{N}(0, \sigma^2)$.

Schemat opisany powyżej można łatwo uogólnić uwzględniając wpływ *wielu* zmiennych objaśniających na zmienną objaśnianą. Na przykład, wydajność procesu chemicznego może zależeć od stężenia katalizatora i od ciśnienia. Na wysokość plonów może mieć wpływ intensywność nawożenia, poziom opadów i jeszcze inne czynniki (zmienne). Nie musimy zakładać, że x_i są skalarami; mogą to być wektory. Również parametr β może być wektorem. Pozostaniemy natomiast przy założeniu, że wartości zmiennej objaśnianej Y_i są skalarne.

[1] Istnieją też nieparametryczne estymatory regresji, o których przelotnie wspomnimy.

Łączna gęstość prawdopodobieństwa obserwacji $Y_1, \ldots, Y_n$ jest następująca:

$$f_{\beta,\sigma}(y_1, \ldots, y_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \phi(\beta, x_i))^2\right].$$

W ten sposób określona jest rodzina rozkładów prawdopodobieństwa na przestrzeni próbkowej $\Omega = \mathbb{R}^n$; przestrzenią parametrów jest $\Theta = \mathbb{R}^p \times]0, \infty[$, gdzie p jest wymiarem parametru β. Ten opis modelu mieści się w ogólnym schemacie wprowadzonym w Rozdziale 2.

Ze wzoru na postać gęstości natychmiast wynika prosty wniosek.

8.1.3 Stwierdzenie. *Jeśli spełnione jest Założenie 8.1.2, to estymator największej wiarogodności parametru β jest rozwiązaniem zadania minimalizacji*

$$\mathrm{SSE}(\beta) := \sum_{i=1}^{n}(Y_i - \phi(\beta, x_i))^2 \to \min_{\beta}.$$

Skrót „SSE" pochodzi od angielskiego zwrotu oznaczającego *sumę kwadratów błędów*, „Sum of Squares of Errors". Będziemy nazywać $\mathrm{SSE} = \min_\beta \mathrm{SSE}(\beta)$ *resztową sumą kwadratów*. Estymator wprowadzony w Stwierdzeniu 8.1.3 nazywamy *estymatorem najmniejszych kwadratów* i w skrócie napiszemy $\hat{\beta} = \mathrm{ENK}(\beta)$. Niezależnie od Założenia 8.1.2, ENK ma bardzo przekonującą interpretację. Dopasowujemy krzywą do punktów doświadczalnych w ten sposób, żeby suma kwadratów „odchyłek" punktów od krzywej była minimalna. Przy tym „odchyłki" mierzymy *wzdłuż osi Y*. Metoda najmniejszych kwadratów sprowadza się do metody największej wiarogodności przy założeniu o normalnym rozkładzie błędów, ale ma samodzielny sens i może być stosowana bez tego założenia.

8.2 Model liniowy

Ograniczymy się do najprostszej, liniowej postaci funkcji regresji. Mimo, że założenie o liniowości wydaje się bardzo ograniczające, rozmaitość i zakres zastowań modeli liniowych są zaskakująco duże.

Prosta regresja liniowa

Rozpatrzymy na początek model z jedną (skalarną) zmienną objaśniającą. Model liniowy z wyrazem wolnym ma postać

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad (i = 1, \ldots, n).$$

Wykresem funkcji regresji jest linia prosta $y = \beta_0 + \beta_1 x$. Wzory przybierają prostą i przejrzystą formę. Estymatory najmniejszych kwadratów parametrów β_0 i β_1 są następujące:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}, \tag{8.2.1}$$

gdzie

$$\bar{x} = \frac{1}{n}\sum x_i, \qquad \bar{Y} = \frac{1}{n}\sum Y_i.$$

Istotnie,

$$\mathrm{SSE}(\beta) = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2.$$

Rozwiązujemy układ równań:

$$\frac{1}{2}\frac{\partial \mathrm{SSE}(\beta)}{\partial \beta_0} = \sum_{i=1}^{n}(\beta_0 + \beta_1 x_i - Y_i) = 0,$$
$$\frac{1}{2}\frac{\partial \mathrm{SSE}(\beta)}{\partial \beta_1} = \sum_{i=1}^{n}(\beta_0 + \beta_1 x_i - Y_i)x_i = 0.$$

Rachunki są elementarne i łatwe (Zadanie 1).

Niech $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, gdzie $\hat{\beta}_0$ i $\hat{\beta}_1$ są ENK danymi wzorem (8.2.1). Punkty $(x_i, \hat{Y}_i)$ leżą na *dopasowanej* (wyestymowanej) prostej regresji. Mówimy, że $\hat{Y}_i$ są *przewidywanymi* wartościami zmiennej objaśnianej. Różnice $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ pomiędzy wartościami obserwowanymi i przewidywanymi nazywamy *resztami* albo *residuami*.

8.2.2 PRZYKŁAD (Ilość produktu i stężenie katalizatora). Badamy zależność ilości produktu w pewnej reakcji chemicznej (zmienna Y) od stężenia katalizatora (zmienna x). Przeprowadzono doświadczenie 15 razy, wybierając różne stężenia katalizatora i otrzymano takie wyniki:

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_i	3	5	7	8	10	11	12	12	13	15	15	16	18	19	20
Y_i	14	40	40	30	50	34	40	70	52	50	70	64	88	72	90

Zakładamy, że ilość produktu zależy w sposób liniowy od stężenia katalizatora (w interesującym nas zakresie wartości obu zmiennych). Odchylenia od dokładnej zależności liniowej traktujemy jako „błędy losowe". Mówiąc porządniej, decydujemy się na opis zależności Y od x przy pomocy modelu prostej regresji liniowej.

Estymowane wartości współczynników są, dla naszych danych, równe $\hat{\beta}_0 = 7.61$ i $\hat{\beta}_1 = 3.75$. Przyjmujemy więc, że funkcja

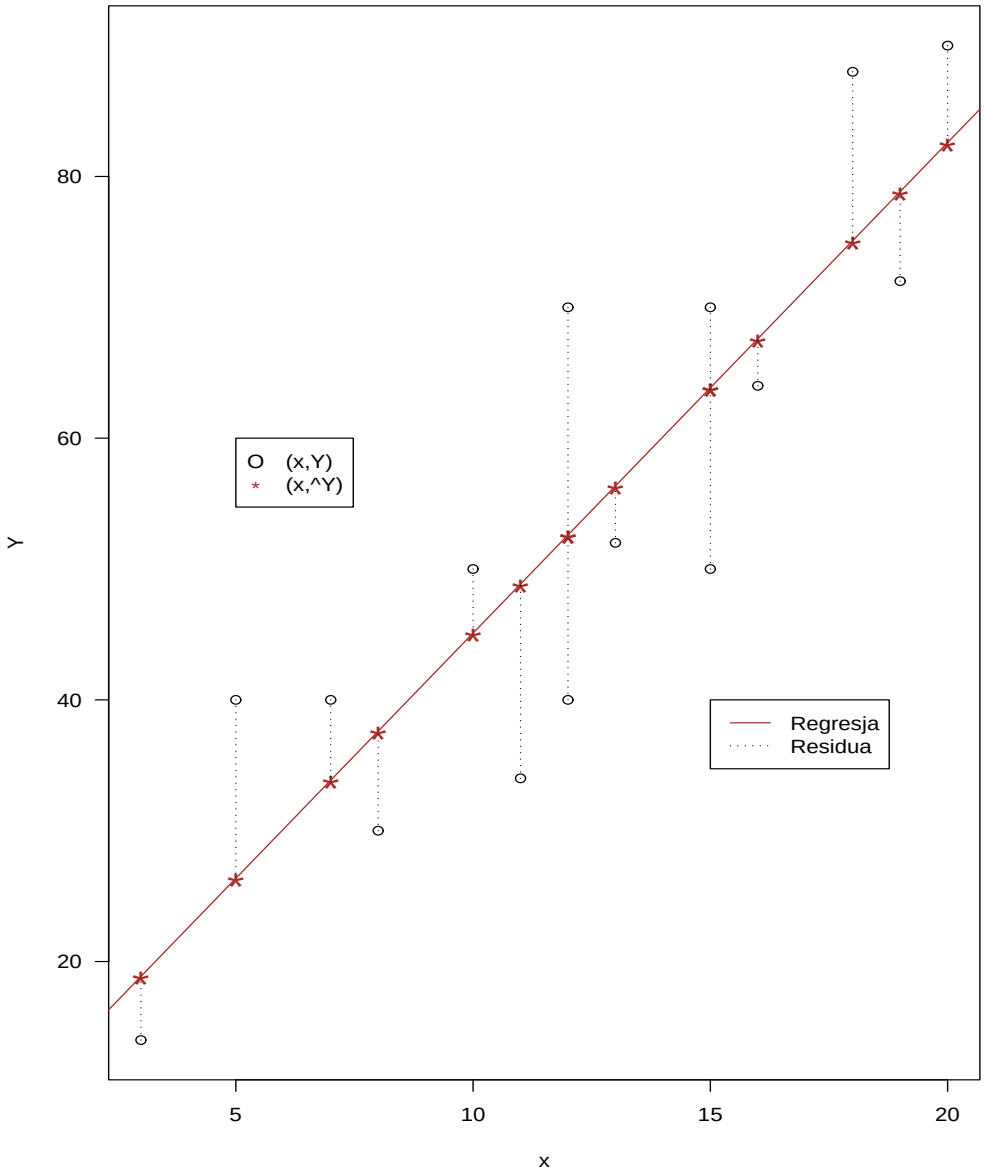
$$\hat{Y} = 7.61 + 3.75x$$

opisuje w przybliżeniu interesującą nas zależność. Obliczyliśmy to przy pomocy programiku napisanego w języku R, który wygląda tak:

```
> x=c(3, 5, 7, 8, 10, 11, 12, 12, 13, 15, 15, 16, 18, 19, 20)
> Y=c(14, 40, 40, 30, 50, 34, 40, 70, 52, 50, 70, 64, 88, 72, 90)
> fit=lm(Y ~ x)
> beta=fit$coefficients
> beta
Intercept       X
7.613670 3.748886
```

Funkcja `lm()` (*linear model*) estymuje parametry regresji liniowej metodą najmniejszych kwadratów. Wyniki działania tej funkcji, zapamiętane w postaci listy `fit`, zawierają oprócz estymatorów (`fit$coefficients`) inne ciekawe wielkości, na przykład residua (`fit$residuals`). Zwróćmy uwagę, że R (domyślnie) włącza wyraz wolny do modelu regresji.

Punkty doświadczalne wraz z dopasowaną prostą regresji pokazuje następujący rysunek.



Regresja liniowa wieloraka

Rozpatrzymy teraz model z *wieloma* zmiennymi objaśniającymi. Ich liczbę oznaczmy przez r. Zmienna objasniana jest jedna, skalarna, tak jak poprzednio. Wskaźnik $i = 1, \ldots, n$ będzie, tak jak dotąd, numerował kolejne „przypadki" lub „obiekty". Zmienne opisujące i-ty obiekt oznaczymy przez $x_{i1}, \ldots, x_{ir}$ i Y_i. Model regresji liniowej z wyrazem wolnym przybiera postać

$$Y_i = \beta_0 + \sum_{j=1}^{r} \beta_j x_{ij} + \varepsilon_i, \qquad (i = 1, \ldots, n).$$

Prosty chwyt pozwala włączyć wyraz wolny do funkcji liniowej. Przyjmijmy umownie, że $x_{i0} = 1$. Zmienne objaśniające dla i-tego obiektu ustawimy w wektor wierszowy, dołączając jedynkę: $x_i^\top = (1, x_{i1}, \ldots, x_{ip})$. Można teraz zapisać bardziej zwięźle model w postaci wektorowej:

$$Y_i = \sum_{j=0}^{r} \beta_j x_{ij} + \varepsilon_i = x_i^\top \beta, \qquad (i = 1, \ldots, n),$$

gdzie $\beta = (\beta_0, \beta_1, \ldots, \beta_r)^\top$. W postaci macierzowej to można przepisać tak:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1r} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nr} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Będziemy konsekwentnie stosowali notację wektorowo-macierzową. Wektory i macierze w powyższym wzorze oznaczymy pojedynczymi literami Y, X, β i ε. Przyjmijmy, dla jednolitości oznaczeń, że symbol p oznaczać będzie wymiar wektora β. Dla regresji liniowej z r zmiennymi objaśniającymi i wyrazem wolnym mamy zatem

$$p = r + 1.$$

Model liniowy przybiera zwięzłą postać:

$$\underset{n \times 1}{Y} \quad = \quad \underset{n \times p}{\mathsf{X}} \quad \cdot \quad \underset{p \times 1}{\beta} \quad + \quad \underset{n \times 1}{\varepsilon}.$$

Pod spodem napisaliśmy *wymiary* poszczególnych obiektów. Znana i nielosowa macierz X jest zwana *macierzą planu*, β jest wektorem nieznanych parametrów, Y jest wektorem obserwacji, ε jest losowym wektorem „błędów".

8.2.3 Uwaga. Zauważmy, że do macierzy X dołączyliśmy „zerową" kolumnę złożoną z samych jedynek. W większości zastosowań jest to naturalne (ta operacja jest wykonywana w R „domyślnie"). Czasami trzeba rozważyć model regresji bez wyrazu wolnego. Należy wtedy pamiętać, że $p = r$, a nie $p = r + 1$. Przyjmijmy umowę, że liczba kolumn macierzy X i wymiar wektora β będą zawsze równe p. W ogólnych, teoretycznych rozważaniach, będziemy pisać $\beta = (\beta_1, \ldots, \beta_p)^\top$, bo wygodniej numerować współrzędne wektora od 1, nie od 0. Wzory dla regresji z wyrazem wolnym wymagają oczywistej modyfikacji.

W dalszym ciągu ograniczymy sie do rozważania następującej sytuacji.

8.2.4 Założenie. *Mamy $p < n$ i macierz X jest pełnego rzędu, to znaczy* $\mathrm{rz}(\mathsf{X}) = p$.

Sens powyższego założenia jest jasny. Wydaje się, że do wyestymowania p nieznanych parametrów, potrzeba więcej niż p obserwacji [2].

Ważna część teorii wymaga wprowadzonego w Założeniu 8.1.2 warunku: $\varepsilon_1, \ldots, \varepsilon_n$ są niezależnymi zmiennymi losowymi o jednakowym rozkładzie $\mathrm{N}(0, \sigma^2)$. Zreasumujmy nasze rozważania w następującej postaci.

8.2.5 Założenie. *Model jest opisany równaniem $Y = \mathsf{X}\beta + \varepsilon$, gdzie $\varepsilon \sim \mathrm{N}(0, \sigma^2\mathsf{I})$.*

Część teorii nie wymaga założenia o normalności. Wystarczy, że zmienne losowe $\varepsilon_1, \ldots, \varepsilon_n$ spełniają warunki $\mathbb{E}\varepsilon_i = 0$ i $\mathrm{Var}\varepsilon_i = \sigma^2$ dla $i = 1, \ldots, n$ oraz $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ dla $i \neq j$. Sformułujmy to w postaci następującego, słabszego założenia.

8.2.6 Założenie. *Model jest opisany równaniem $Y = \mathsf{X}\beta + \varepsilon$, gdzie $\mathbb{E}\varepsilon = 0$ i $\mathrm{VAR}\varepsilon \sim \sigma^2\mathsf{I}$.*

[2] W ostatnich latach coraz więcej uwagi poświęca się w statystyce modelom, w których $p > n$. Ale to już inna historia, wykraczająca poza zakres naszych rozważań.

Poniższy przykład pokazuje, że założenie o liniowości funkcji regresji jest mniej ograniczające, niż się wydaje.

8.2.7 PRZYKŁAD (Regresja wielomianowa). Rozpatrzmy model z pojedynczą zmienną objaśniającą, w którym funkcja regresji jest wielomianem r-tego stopnia:

$$Y_i = \beta_0 + \sum_{j=1}^{r} \beta_j x_i^j + \varepsilon_i, \qquad (i = 1, \ldots, n).$$

To jest *model liniowy*, w którym i-ty wiersz macierzy planu jest równy

$$x_i^\top = (1, x_i, \ldots, x_i^j, \ldots, x_i^r) \qquad (i = 1, \ldots, n).$$

Estymacja w modelu liniowym

Pracujemy w ogólnym modelu liniowym $Y = \mathsf{X}\beta + \varepsilon$. Przy Założeniach 8.2.6 i 8.2.4 można napisać jawne, macierzowe wzory na estymator najmniejszych kwadratów, ENK(β). Rozwiązujemy zadanie minimalizacji

$$\mathrm{SSE}(\beta) = \sum_{i=1}^{n} (Y_i - x_i^\top \beta)^2 = (\mathsf{X}\beta - Y)^\top (\mathsf{X}\beta - Y) = \min_{\beta}.$$

Obliczając gradient lewej strony względem β dostajemy $\mathsf{X}^\top(\mathsf{X}\beta - Y) = 0$, czyli

$$\mathsf{X}^\top \mathsf{X}\beta = \mathsf{X}^\top Y.$$

Jest to tak zwany układ równań *normalnych* w postaci macierzowej. Założenie 8.2.4 gwarantuje, że macierz $\mathsf{X}^\top\mathsf{X}$ jest odwracalna i mamy prosty wzór:

$$\mathrm{ENK}(\beta) = \hat{\beta} = (\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top Y.$$

Ponieważ $\mathbb{E}Y = \mathsf{X}\beta$, więc $\mathbb{E}\hat{\beta} = (\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top\mathbb{E}Y = \beta$. ENK($\beta$) jest estymatorem *nieobciążonym*. Policzymy macierz kowariancji ENK. Mamy

$$\mathrm{VAR}(\hat{\beta}) = (\mathrm{Cov}(\beta_j, \beta_k); j, k = 1, \ldots, p) = \sigma^2(\mathsf{X}^\top\mathsf{X})^{-1}.$$

Istotnie,

$$\begin{aligned}
\mathrm{VAR}(\hat{\beta}) &= \mathbb{E}(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top \\
&= (\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top\mathbb{E}(Y - \mathsf{X}\beta)(Y - \mathsf{X}\beta)^\top\mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1} \\
&= (\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top(\sigma^2\mathsf{I})\mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1} = \sigma^2(\mathsf{X}^\top\mathsf{X})^{-1}.
\end{aligned}$$

W Rozdziale 3 wprowadziliśmy pojęcie estymatora nieobciążonego o minimalnej wariancji. Obecnie mamy do czynienia z *wektorowym* parametrem $\beta = (\beta_1, \ldots, \beta_p)$. Zajmiemy się estymacją *liniowej funkcji* tego parametru, to znaczy wyrażenia postaci

$$g^\top \beta = \sum_{j=1}^{p} g_j \beta_j,$$

sposób przechodzimy do *jednowymiarowego* zagadnienia estymacji i możemy odwołać się do znajomych pojęć. Z definicji,

$$\mathrm{ENK}(g^\top \beta) = \hat{g} = g^\top \hat{\beta}$$

jest *estymatorem najmniejszych kwadratów* funkcji $g^\top \beta$ (po prostu, wstawiamy $\mathrm{ENK}(\beta)$ w miejsce nieznanego β). Okazuje się, że ENK mają *najmniejszą wariancję pośród estymatorów liniowych i nieobciążonych.* Mówi się, że $\hat{\beta}$ jest najlepszym liniowym nieobciążonym estymatorem β, w skrócie BLUE (Best Linear Unbiased Estimator). Taka jest treść klasycznego twierdzenia, które teraz sformułujemy dokładniej.

8.2.8 TWIERDZENIE (Gaussa – Markowa). *Przyjmijmy Założenie 8.2.6. Rozważmy dowolny nieobciążony i liniowy estymator funkcji $g^\top \beta$, to znaczy estymator postaci $\tilde{g} = c^\top Y$ taki, że $\mathbb{E}\tilde{g} = g^\top \beta$. Jeżeli $\hat{g} = ENK(g^\top \beta)$, to*

$$\mathrm{Var}\hat{g} \leq \mathrm{Var}\tilde{g}.$$

Dowód. Ponieważ $\mathbb{E}\tilde{g} = c^\top \mathsf{X}\beta = g^\top \beta$ dla każdego β, więc $c^\top \mathsf{X} = g^\top$. Oczywiście, $\mathrm{Var}\tilde{g} = \sigma^2 c^\top c$. Dowód będzie zakończony gdy pokażemy, że

$$0 \leq c^\top c - g^\top (\mathsf{X}^\top \mathsf{X})^{-1} g.$$

Możemy tę nierówność przepisać w postaci

$$0 \leq c^\top c - c^\top \mathsf{X}(\mathsf{X}^\top \mathsf{X})^{-1} \mathsf{X}^\top c = c^\top (\mathsf{I} - \mathsf{H}) c,$$

gdzie $\mathsf{H} = \mathsf{X}(\mathsf{X}^\top \mathsf{X})^{-1} \mathsf{X}^\top$. Wystarczy teraz zauważyć, że macierz $\mathsf{I} - \mathsf{H}$ jest *nieujemnie określona.* Jest tak, bo jest ona *symetryczna i idempotentna.* □

Jeśli przyjmiemy silniejsze Założenie 8.2.5 zamiast 8.2.6 to można pokazać, że ENK jest nie tylko BLUE (najlepszy wśród liniowych estymatorów nieobciążonych) ale także ENMW (najlepszy wśród *wszystkich* estymatorów nieobciążonych). Przyjmiemy ten fakt bez dowodu.

Geometria ENK

W dalszym ciągu rozważamy ogólny model $Y = \mathsf{X}\beta + \varepsilon$. Będziemy w istotny sposób korzystać z Założeń 8.2.5 i 8.2.4. Współrzędne wektorów p-wymiarowych numerujemy od 1 do p. Zauważmy, że $\hat{Y}_i = x_i^\top \hat{\beta}$ jest współrzędną Y-ową punktu odpowiadającego wektorowi x_i i leżącemu na wykresie *dopasowanej* (estymowanej metodą NK) funkcji regresji. Odpowiednią resztą jest $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$. Wektorowo napiszemy $\hat{Y} = \mathsf{X}\hat{\beta} = (\hat{Y}_1, \ldots, \hat{Y}_n)^\top$ i $\hat{\varepsilon} = Y - \hat{Y}$. Mamy

$$\hat{Y} = \mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top Y = \mathsf{H}Y,$$

gdzie H jest macierzą rzutu ortogonalnego (w przestrzeni $\mathbb{R}^n$) na p-wymiarową podprzestrzeń liniową $\mathcal{R}(\mathsf{X})$ generowaną przez kolumny macierzy X (czyli obraz przekształcenia liniowego o macierzy X). Wystarczy sprawdzić, że H jest macierzą *symetryczną* ($\mathsf{H}^\top = \mathsf{H}$) i *idempotentną* ($\mathsf{H}^2 = \mathsf{HH} = \mathsf{H}$). Rzut na dopełnienie ortogonalne $\mathcal{R}(\mathsf{X})^\perp$ ma macierz $\mathsf{I} - \mathsf{H}$.

Geometryczna interpretacja metody najmniejszych kwadratów staje się przejrzysta, jeśli przejdziemy do takiego ortogonalnego układu współrzędnych, którego pierwsze p wersorów jest bazą podprzestrzeni $\mathcal{R}(\mathsf{X})$ a następne $n-p$ wersorów jest bazą $\mathcal{R}(\mathsf{X})^\perp$. Taki układ można napisać w jawnej postaci stosując procedurę ortogonalizacji Hilberta-Schmidta do bazy (nieortogonalnej) przestrzeni $\mathbb{R}^n$, złożonej z p kolumn macierzy X oraz $n-p$ innych wektorów. Pamiętajmy, że macierz X o wymiarach $n \times p$ jest pełnego rzędu p.

Potrzebny fakt sformułujemy w następującej postaci [3].

[3] Dla naszych celów istotne będą tylko współrzędne kolumn X w nowej bazie.

8.2.9 Stwierdzenie. *Istnieje macierz ortogonalna* Q *o wymiarze* $n \times n$ *oraz macierz górna trójkątna* R *o wymiarze* $p \times p$ *takie, że*

$$\mathsf{X} = \mathsf{Q} \begin{pmatrix} \mathsf{R} \\ \cdots \\ \mathsf{O} \end{pmatrix}.$$

W tym wzorze O *jest zerową macierzą o wymiarze* $(n-p) \times p$.

Kolumny macierzy Q tworzą ortonormalną bazę $\mathcal{R}(\mathsf{X})$. Element r_{jk} macierzy R jest iloczynem skalarnym k-tej kolumny X i j-tej kolumny Q. Tak więc R zawiera współrzędne kolumn X w nowej bazie. Fakt, że macierz R jest trójkątna, $r_{jk} = 0$ dla $k < j$ oznacza, że początkowe k kolumn Q jest bazą w przestrzeni rozpiętej przez k pierwszych kolumn X.

Współrzędne wektora Y w nowej bazie oznaczymy $\underline{Y} = \mathsf{Q}^\top Y$. Rzut na przestrzeń $\mathcal{R}(\mathsf{X})$ ma w nowej bazie macierz współrzędnych $\underline{\mathsf{H}} = \mathsf{Q}^\top \mathsf{H} \mathsf{Q}$:

$$\underline{Y} \mapsto Y = \mathsf{Q}\underline{Y} \mapsto \hat{Y} = \mathsf{H}Y = \mathsf{H}\mathsf{Q}\underline{Y} \mapsto \underline{\hat{Y}} = \mathsf{Q}^\top \mathsf{H} \mathsf{Q} \underline{Y}.$$

Zauważmy, że

$$\begin{aligned} \underline{\mathsf{H}} = \mathsf{Q}^\top \mathsf{X} (\mathsf{X}^\top \mathsf{X})^{-1} \mathsf{X}^\top \mathsf{Q} &= \begin{pmatrix} \mathsf{R} \\ \cdots \\ \mathsf{O} \end{pmatrix} \left(\mathsf{R}^\top \mathsf{R}\right)^{-1} \left(\mathsf{R}^\top \vdots \mathsf{O}\right) \\ &= \begin{pmatrix} \mathsf{I} & \vdots & \mathsf{O} \\ \cdots & & \cdots \\ \mathsf{O} & \vdots & \mathsf{O} \end{pmatrix}, \end{aligned}$$

gdzie I jest macierzą jednostkową wymiaru $p \times p$. To znaczy, że w nowym układzie współrzędnych, rzutowanie wektora Y na podprzestrzeń $\mathcal{R}(\mathsf{X})$ polega na zastąpieniu $n-p$ ostatnich współrzędnych zerami:

$$\text{jeśli } \underline{Y} = (\underline{Y}_1, \ldots, \underline{Y}_p, \ldots, \underline{Y}_n)^\top \text{ to } \underline{\hat{Y}} = (\underline{Y}_1, \ldots, \underline{Y}_p, 0, \ldots, 0)^\top.$$

Pamiętajmy przy tym, że $Y = \mathsf{Q}\underline{Y}$ i $\hat{Y} = \mathsf{Q}\underline{\hat{Y}}$. Wzór $Y = \mathsf{X}\beta = \varepsilon$ w nowym układzie współrzędnych przybiera postać

$$\underline{Y} = \begin{pmatrix} \mathsf{R} \\ \cdots \\ \mathsf{O} \end{pmatrix} \beta + \underline{\varepsilon},$$

gdzie $\underline{\varepsilon} = \mathsf{Q}\varepsilon$. Następujące proste spostrzeżenie odgrywa w dalszych rozważaniach zasadniczą rolę. *Przy Założeniu 8.2.5, wektor losowy $\underline{\varepsilon}$ ma łączny rozkład normalny* $\mathrm{N}(0, \sigma^2\mathsf{I})$.

Geometryczne rozważania prowadzą do bardzo prostego dowodu następującej ogólnej wersji Twierdzenia Fishera.

8.2.10 TWIERDZENIE (Fishera). *Jeśli spełnione jest Założenie 8.2.5 i $\hat{\beta} = ENK(\beta)$ to $\hat{\beta}$ jest zmienną losową niezależną od $Y - \hat{Y}$. Ponadto mamy $\|Y - \hat{Y}\|^2 \sim \chi^2(n-p)$ i $\hat{\beta} \sim \mathrm{N}(\beta, \sigma^2(\mathsf{X}^\top\mathsf{X})^{-1})$.*

Dowód. Ponieważ Q jest macierzą ortogonalną, więc jest izometrią, stąd

$$\begin{aligned}\|Y - \hat{Y}\|^2 = \|\mathsf{Q}^\top Y - \mathsf{Q}^\top\hat{Y}\|^2 = \|\mathsf{Q}^\top(\mathsf{I} - \mathsf{H})\mathsf{Q}\underline{Y}\|^2 = \|(\mathsf{I} - \underline{\mathsf{H}})\underline{Y}\|^2 \\ = \underline{\varepsilon}_{p+1}^2 + \cdots + \underline{\varepsilon}_n^2 \sim \chi^2(n-p).\end{aligned}$$

Z kolei $\hat{Y} = \mathsf{Q}\underline{\hat{Y}} = \mathsf{Q}\underline{\mathsf{H}Y} = \mathsf{Q}(\underline{\varepsilon}_1, \ldots, \underline{\varepsilon}_p, 0, \ldots, 0)^\top$. Stąd widać, że $\hat{Y}$ jest zmienną niezależną od $Y - \hat{Y}$. Oczywiście, $\hat{\beta}$ jest funkcją $\hat{Y}$, a więc też jest zmienną niezależną od $Y - \hat{Y}$. Wreszcie, $\hat{\beta}$ jest to liniową funkcją wektora Y, a więc ma rozkład normalny. Wiemy, że $\mathbb{E}\hat{\beta} = \beta$ i $\mathrm{VAR}(\hat{\beta}) = \sigma^2(\mathsf{X}^\top\mathsf{X})^{-1}$, co kończy dowód. □

8.2.11 Wniosek. *Nieobciążonym estymatorem wariancji błędu, σ^2, jest*

$$S^2 = \frac{\|Y - \hat{Y}\|^2}{n-p} = \frac{\mathrm{SSE}}{n-p}.$$

Estymatory najmniejszych kwadratów $\hat{\beta}_j$ można „uzupełnić" konstrukcją przedziałów ufności.

8.2.12 Wniosek. *Przedział ufności dla parametru β_j jest określony wzorem*

$$\left[\hat{\beta}_j - Std_j, \hat{\beta}_j + Std_j\right],$$

gdzie $S = \sqrt{S^2}$, $t = \mathrm{t}_{1-\alpha/2}(n-p)$ jest kwantylem rozkładu t-Studenta z $n-p$ stopniami swobody, zaś $d_j = \sqrt{(\mathsf{X}^\top\mathsf{X})^{-1}_{jj}}$ (wskaźnik jj odpowiada j-temu elementowi na przekątnej macierzy).

Żeby ten wniosek uzasadnić, wystarczy zauważyć, że $\mathrm{Var}\beta_j = \sigma^2 d_j$, a zatem

$$\frac{\hat{\beta}_j - \beta_j}{d_j S} \sim \mathrm{t}(n-p),$$

na mocy twierdzenia Fishera.

Predykcja

Po co właściwie dopasowujemy funkcję do punktów doświadczalnych? Rzecz jasna, jest przyjemnie mieć prosty, liniowy model opisujący zależność. Ostatecznym sprawdzianem wartości poznawczej modelu jest możliwość *przewidywania wyników doświadczeń*. W przypadku modelu regresji, chodzi o przewidywanie wartości zmiennej Y dla *danej wartości* x. Tak jak dotąd, mamy dane punkty (x_i, Y_i) dla $i = 1, \ldots, n$. Dla ustalenia uwagi umówmy się, że wracamy do modelu regresji liniowej z wyrazem wolnym i do oznaczenia $\beta = (\beta_0, \beta_1 \ldots, \beta_r)$ na wektor współczynników, gdzie $p = r + 1$. Rozważamy „nowy" wektor zmiennych objaśniających, który oznaczymy $x_*^\top = (1, x_1^*, \ldots, x_r^*)$ i uważamy za znany. Jeśli przeprowadzimy *nowe* doświadczenie, to pojawi się odpowiednia wartość Y_*. Naszym zadaniem jest *predykcja* nieznanej wartości Y_* *przed* dokonaniem tego dodatkowego doświadczenia. Nasz model przewiduje, że

$$Y_* = x_*^\top \beta + \varepsilon_*,$$

gdzie współczynniki β są te same, co we wzorze $Y_i = x_i^\top \beta + \varepsilon_i$, zaś $\varepsilon_* \sim \mathrm{N}(0, \sigma^2)$ jest błędem losowym niezależnym od poprzednich błędów ε_i. Musimy zmierzyć się z dwiema trudnościami. Po pierwsze, nie znamy współczynników β. Po drugie, musimy się liczyć z nowym, losowym odchyleniem ε_* od prostej regresji. Niemniej, nasuwa się dość oczywiste rozwiązanie. Za *przewidywany* wynik doświadczenia możemy przyjąć

$$\hat{Y}_* = x_*^\top \hat{\beta},$$

gdzie $\hat{\beta}$ jest estymatorem obliczonym na podstawie *poprzednich* punktów doświadczalnych (x_i, Y_i). Spróbujemy teraz oszacować dokładność predykcji.

Mamy $\mathbb{E}\hat{Y}_* = \mathbb{E}Y_* = x_*^\top \beta$ i możemy powiedzieć, że *predyktor* $\hat{Y}_*$ jest nieobciążony [4]. Obliczmy jego wariancję:

$$\mathrm{Var}\hat{Y}_* = \sigma^2 x_*^\top (\mathsf{X}^\top \mathsf{X})^{-1} x_*.$$

Łatwo stąd wywnioskować ważny wzór na błąd średniokwadratowy predykcji:

$$\mathbb{E}(\hat{Y}_* - Y_*)^2 = \sigma^2 \left[1 + x_*^\top (\mathsf{X}^\top \mathsf{X})^{-1} x_*\right].$$

„Dodatkowa jedynka" w tym wzorze pochodzi stąd, że musimy uwzględnić wpływ błędu ε_*, czyli losowe odchylenie punktu Y_* od funkcji regresji.

Bardzo podobnie jak we Wniosku 8.2.12 konstruuje się przedziały ufności dla wartości funkcji regresji i predykcji.

8.2.13 Wniosek. *Przedział ufności dla wartości funkcji regresji w punkcie x_*, czyli dla $\beta^\top x_*$ jest określony wzorem*

$$\left[x_*^\top \hat{\beta} - Std_*, x_*^\top \hat{\beta} + Std_*\right],$$

gdzie $d_* = \sqrt{x_*^\top (\mathsf{X}^\top \mathsf{X})^{-1} x_*^\top}$ *i* $t = \mathrm{t}_{1-\alpha/2}(n-p)$.

Przejdźmy do przedziałów ufności dla predykcji. Ustalamy poziom ufności $1-\alpha$ i chcemy skonstruować takie statystyki $\underline{Y_*}$ i $\overline{Y_*}$, żeby, dla dowolnych β i σ,

$$\mathbb{P}(\underline{Y_*} \leq Y_* \leq \overline{Y_*}) = 1 - \alpha.$$

W powyższym wzorze występuje rozkład prawdopodobieństwa na przestrzeni próbkowej $\mathbb{R}^{n+1}$. Jest to łączny rozkład zmiennych losowych $Y_1, \ldots, Y_n$ *oraz* Y_*. Statystyki $\underline{Y_*}$ i $\overline{Y_*}$ są to funkcje *obserwacji*, czyli zmiennych losowych Y_i dla $i = 1, \ldots, n$. Poza tym mogą zależeć od znanych liczb x_i oraz x_*, ale nie mogą zależeć od Y_*. Przedział $[\underline{Y_*}, \overline{Y_*}]$ będziemy dla uproszczenia nazywać *przedziałem ufności dla predykcji*, ale nie jest to przedział ufności w rozumieniu Definicji **??**. Wartość, którą staramy się przewidzieć, Y_*, nie jest funkcją nieznanego parametru, tylko *nieobserwowaną zmienną losową*.

[4] Zauważmy , że przewidywana wielkośc Y_* jest zmienną losową, a zatem nieobciążoność predyktora wymaga osobnej definicji.

8.2.14 Wniosek. *Przedział ufności dla predykcji Y_* jest określony wzorem*

$$\left[x_*^\top\hat{\beta} - StD_*, x_*^\top\hat{\beta} + StD_*\right],$$

gdzie $D_* = \sqrt{1 + x_*^\top(\mathsf{X}^\top\mathsf{X})^{-1}x_*^\top}$ *i* $t = \mathrm{t}_{1-\alpha/2}(n-p)$.

Uzasadnienie Wniosków 8.2.13 i 8.2.14 jest analogiczne jak Wniosku 8.2.12. Wystarczy powołać się na twierdzenie Fishera i wykorzystać wzory $\mathrm{Var}\hat{Y}_* = \sigma^2 d_* = \sigma^2 x_*^\top(\mathsf{X}^\top\mathsf{X})^{-1}x_*$ i $\mathbb{E}(\hat{Y}_* - Y_*)^2 = \sigma^2 D_* = \sigma^2\left[1 + x_*^\top(\mathsf{X}^\top\mathsf{X})^{-1}x_*\right]$.

W szczególnym przypadku *prostej regresji liniowej z wyrazem wolnym* wzory na przedziały ufności mają wyjątkowo intuicyjną inerpretację i warto je przytoczyć. Wprowadźmy wygodne oznaczenie $\mathrm{SS_x} = \sum(x_i - \bar{x})^2$. Mamy

$$d_* = \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\mathrm{SS_x}},$$

$$D_* = 1 + d_* = 1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\mathrm{SS_x}}.$$

To można sprawdzić wykorzystując ogólne wzory (trzeba obliczyć macierz odwrotną $(\mathsf{X}^\top\mathsf{X})^{-1}$ wymiaru 2×2). Można też obliczyć bezpośrednio $\mathrm{Var}\hat{Y}_*$ i $\mathrm{Var}(Y_* - \hat{Y}_*)$ w rozważanym szczególnym przypadku (Zadanie 4). Tak czy inaczej, rachunki są łatwe. Zwrócmy uwagę, że liczba stopni swobody resztowej sumy kwadratów SSE jest równa $n-2$, ze względu na obecność wyrazu wolnego. Zatem $t = \mathrm{t}_{1-\alpha/2}(n-2)$.

Jeśli prawe strony wzorów na przedziały ufności,

$$\beta_0 + \beta_1 x_* = \hat{\beta}_0 + \hat{\beta}_1 x_* \pm tSd_*,$$
$$Y_* = \hat{\beta}_0 + \hat{\beta}_1 x_* \pm t\hat{S}D_*,$$

potraktujemy jako funkcje x_*, to otrzymamy krzywe wyznaczające „pasy ufności" odpowiednio, dla funkcji regresji i predykcji.

8.2.15 PRZYKŁAD (Produkt i katalizator, kontynuacja). Wróćmy do Przykładu 8.2.15. Przypomnijmy, że na podstawie $n = 15$ punktów doświadczalnych obliczyliśmy estymatory współczynników równe $\hat{\beta}_0 = 7.61$ i $\hat{\beta}_1 = 3.75$. Dopasowana prosta regresji jest więc taka:

$$\hat{Y} = 7.61 + 3.75x.$$

Przypuśćmy teraz, że chcemy przewidzieć, jaką uzyskamy ilość produktu w nowym doświadczeniu, przy stężeniu katalizatora $x_* = 10.5$. Oczywiście,

$$\hat{Y}_* = 7.61 + 3.75 * 10.5 = 46.98.$$

Szerokość przedziału ufności obliczymy według wzoru $\mathrm{t}_{0.975}(13) * \sqrt{\mathrm{SSE}/13} * \sqrt{1 + 1/15 + (10.5 - \bar{x})^2/\mathrm{SS_x}} = 24.39$, gdzie $\bar{x} = 12.27$, $\mathrm{SS_x} = 358.93$ i $\mathrm{SSE} = 1541.1$. Na poziomie ufności 0.95 możemy twierdzić, że doświadczenie da wynik 46.98 ± 24.39, czyli Y_* zmieści się w przedziale

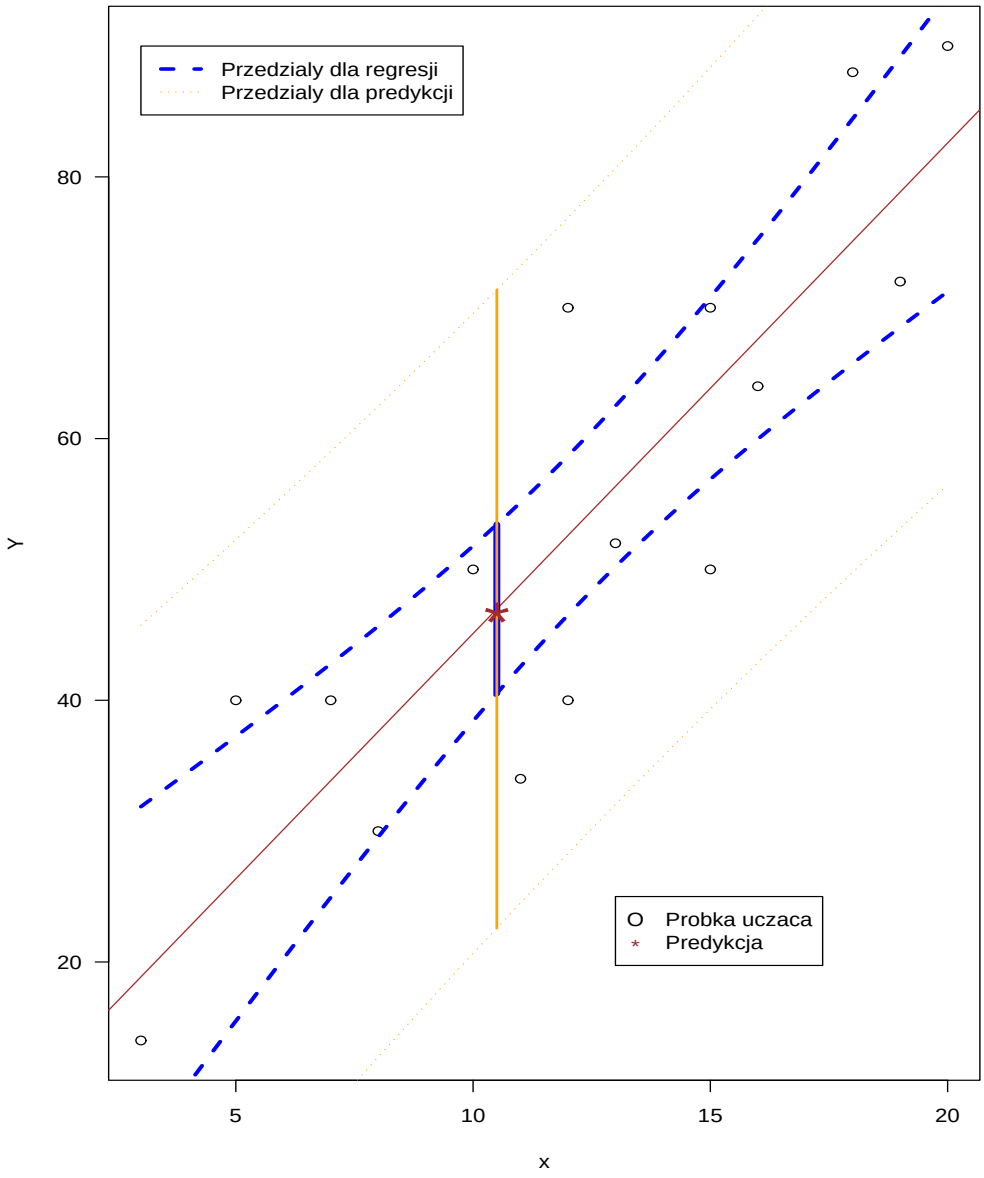
$$[22.58, 71.37]$$

Zatrzymajmy się jeszcze nad interpretacją przedzia łu ufności dla funkcji regresji, $\beta_0 + \beta_1 * 10.5$. Ten przedział w naszym przykładzie przybiera postać 46.98 ± 6.46, czyli

$$[40.52, 53.43].$$

Powiedzmy, że zdecydujemy się uruchomić produkcję na większą skalę i powtarzać wielokrotnie reakcję przy tym samym „roboczym" stężeniu $x_* = 10.5$. Wtedy *średnia* ilość otrzymywanego produktu będzie równa $\beta_0 + 10.5\beta_1$. Ponieważ parametry zależności β_0 i β_1 są nieznane (wartości $\hat{\beta}_0 = 7.61$ i $\hat{\beta}_1 = 3.75$ są tylko *estymatorami !*) to średnią ilość produktu możemy oszacować tylko „z dokładnością ± 6.46".

Przedziały ufności dla Y_*, dla wartości funkcji regresji w punkcie $x_* = 10.5$ oraz pasy ufności widać na następującym rysunku:


Przedzialy dla regresji
Przedzialy dla predykcji
Probka uczaca
Predykcja
Y
x
20
40
60
80
5
10
15
20

Testowanie hipotez

Najprostsze i najważniejsze zagadnienie testowania hipotez w modelu liniowym zmierza do odpowiedzi na pytanie: „czy wszystkie zmienne objaśniające mają istotny wpływ na zmienną objaśnianą?". Czy może pewien podzbiór zmiennych x, można pominąć? Formalnie, niech $\beta = (\beta_1, \ldots, \beta_q, \beta_{q+1}, \ldots, \beta_p)^\top$ dla $q < p$. Weryfikujemy hipotezę zerową

$$\mathrm{H}_0 : (\beta_{q+1}, \ldots, \beta_p)^\top = (0, \ldots, 0)^\top.$$

przeciwko alternatywie

$$\mathrm{H}_1 : (\beta_{q+1}, \ldots, \beta_p)^\top \neq (0, \ldots, 0)^\top.$$

Niech X_0 oznacza macierz planu X z pominiętymi kolumnami $q+1, \ldots, p$. Jest to więc macierz $n \times q$, która odpowiada modelowi regresji zbudowanemu przy założeniu prawdziwości H_0. Zauważmy, że geometryczne rozważania poprzedniego punktu przenoszą sie bez zmian, jesli zastąpimy X przez X_0. Co więcej, jeśli rozpatrzymy dekompozycję macierzy X podaną w Stwierdzeniu 8.2.9 to automatycznie otrzymujemy dekompozycję X_0. Wystarczy wybrać za R_0 podmacierz trójkątną o wymiarach $q \times q$ stojącą „w lewym górnym rogu R". Macierz Q pozostaje ta sama, czyli możemy pracować w tym samym wygodnym ortogonalnym układzie współrzędnych. Niech H_0 oznacza rzut na $\mathcal{R}(\mathsf{X}_0) \subset \mathcal{R}(\mathsf{X})$. Mamy $\mathsf{H}_0 = \mathsf{H}_0\mathsf{H}$: rzut rzutu jest rzutem. Najlepiej to widać w nowym układzie współrzędnych:

$$\underline{\mathsf{H}}_0 = \begin{pmatrix} \mathsf{I}_q & \vdots & \mathsf{O} & \vdots & \mathsf{O} \\ \cdots & & \cdots & & \cdots \\ \mathsf{O} & \vdots & \mathsf{O}_{p-q} & \vdots & \mathsf{O} \\ \cdots & & \cdots & & \cdots \\ \mathsf{O} & \vdots & \mathsf{O} & \vdots & \mathsf{O}_{n-p} \end{pmatrix}, \quad \underline{\mathsf{H}} = \begin{pmatrix} \mathsf{I}_q & \vdots & \mathsf{O} & \vdots & \mathsf{O} \\ \cdots & & \cdots & & \cdots \\ \mathsf{O} & \vdots & \mathsf{I}_{p-q} & \vdots & \mathsf{O} \\ \cdots & & \cdots & & \cdots \\ \mathsf{O} & \vdots & \mathsf{O} & \vdots & \mathsf{O}_{n-p} \end{pmatrix}.$$

W tym wzorze indeksy oznaczają wymiary kwadratowych bloków. Rzuty na $\mathcal{R}(\mathsf{X})$ i $\mathcal{R}(\mathsf{X}_0)$ opisują wzory

$$\begin{aligned} \underline{Y} &= (\underline{Y}_1, \ \ldots, \ \underline{Y}_q, \ \underline{Y}_{q+1}, \ \ldots, \ \underline{Y}_p, \ \underline{Y}_{p+1}, \ \ldots, \ \underline{Y}_n)^\top, \\ \underline{\hat{Y}} &= (\underline{Y}_1, \ \ldots, \ \underline{Y}_q, \ \underline{Y}_{q+1}, \ \ldots, \ \underline{Y}_p, \ 0, \ \ldots, \ 0)^\top, \\ \underline{\hat{Y}}_0 &= (\underline{Y}_1, \ \ldots, \ \underline{Y}_q, \ 0, \ \ldots, \ 0, \ 0, \ \ldots, \ 0)^\top. \end{aligned}$$

Wektory

$$\hat{Y}_0, \quad \hat{Y} - \hat{Y}_0, \quad Y - \hat{Y}$$

są wzajemnie prostopadłe. Z twierdzenia Pitagorasa wynikają tożsamości

$$\|\hat{Y}\|^2 + \|Y - \hat{Y}\|^2 = \|Y\|^2.$$

oraz

$$\|\hat{Y} - \hat{Y}_0\|^2 + \|Y - \hat{Y}\|^2 = \|Y - \hat{Y}_0\|^2.$$

Wprowadźmy oznaczenia

$$\mathrm{SSE} = \|Y - \hat{Y}\|^2, \qquad \mathrm{SSE}_0 = \|Y - \hat{Y}_0\|^2.$$

Wiemy, że $\mathrm{SSE} = \underline{\varepsilon}_{p+1}^2 + \cdots + \underline{\varepsilon}_n^2$ (przy założeniu, że model jest poprawny). Jeśli że H_0 jest prawdziwa (czyli model w istocie zawiera tylko q zmiennych x) to mamy analogicznie $\mathrm{SSE}_0 = \underline{\varepsilon}_{q+1}^2 + \cdots + \underline{\varepsilon}_p^2 + \cdots + \underline{\varepsilon}_n^2$. Stąd wyciągamy wniosek, pozwalający na skonstruowanie testu H_0:

8.2.16 Wniosek. *Przy prawdziwości* H_0*, statystyka*

$$F = \frac{(\mathrm{SSE}_0 - \mathrm{SSE})/(p-q)}{\mathrm{SSE}/p}$$

ma rozkład $\mathrm{F}(p-q, p)$ *(rozkład Fishera-Snedecora z* $p-q$ *stopniami swobody w liczniku i* p *stopniami swobody w mianowniku).*

Warto zauważyć, że ten test jest niczym innym jak testem ilorazu wiarogodności dla hipotez złożonych. W istocie, wiarogodność jest dana wzorem

$$\begin{aligned}\mathcal{L}(\beta, \sigma) &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum (Y_i - x_i^\top \beta)^2\right) \\ &\propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\|Y - \mathsf{X}\beta\|^2\right),\end{aligned}$$

więc

$$\ell(\beta, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2}\|Y - \mathsf{X}\beta\|^2 + \mathrm{const.}$$

Ponieważ $\mathrm{ENW}(\sigma) = \hat{\sigma} = \mathrm{SSE}/n$, więc

$$\ell(\hat{\beta}, \hat{\sigma}) = -n \log \mathrm{SSE} + \mathrm{const.}$$

Analogicznie, dla „mniejszego modelu" otrzymujemy estymator z ograniczeniami $\mathrm{ENW}_0(\sigma) = \hat{\sigma}_0 = \mathrm{SSE}_0/n$. W rezultacie,

$$\ell(\hat{\beta}, \hat{\sigma}) - \ell(\hat{\beta}_0, \hat{\sigma}_0) = n \log \frac{\mathrm{SSE}_0}{\mathrm{SSE}}.$$

Statystyka F jest rosnącą funkcją obliczonej powyżej statystyki ilorazu wiarogodności.

8.2.17 Uwaga (Ogólne hipotezy liniowe). Skoncentrowaliśmy się na zagadnieniu testowania hipotezy $\mathrm{H}_0 : (\beta_{q+1}, \ldots, \beta_p)^\top = (0, \ldots, 0)^\top$ po pierwsze dlatego, że to jest ważne w zastosowaniach: chcemy wyeliminować „niepotrzebne zmienne" i uprościć model. Po drugie, macierz Q w Stwierdzeniu 8.2.9 daje ortogonalny układ współrzędnych idealnie pasujący do tej postaci hipotez. Oczywiście, należy wpierw ustawić zmienne objaśniające w odpowiedniej kolejności, zaczynając numerację od tych, które wydają się ważniejsze a kończąc na tych, które podejrzewamy o „bycie zbędnymi". Co więcej, cała teoria bez zmian stosuje się do ogólnych hipotez liniowych postaci

$$\mathrm{H}_0 : \mathsf{C}\beta = 0,$$

gdzie C jest macierzą $(p-q) \times p$ pełnego rzędu. Taka hipoteza stwierdza, że β należy do podprzestrzeni liniowej wymiaru q. Wektor $\hat{Y}_0$ jest rzutem ortogonalnym na $\{y : y = \mathsf{X}\beta, \mathsf{C}\beta = 0\}$ i definiujemy, tak jak poprzednio, $\mathrm{SSE}_0 = \|Y - \hat{Y}_0\|^2$. Wniosek 8.2.16 pozostaje prawdziwy.

8.2.18 Uwaga (Test t i test F). Rozumowanie uzasadniające Wniosek 8.2.12 można wykorzystać do konstrukcji testu hipotezy $\mathrm{H}_0 : \beta_j = 0$. Używamy statystyki testowej Studenta,

$$T = \frac{\hat{\beta}_j}{d_j S}, \quad d_j = \sqrt{(\mathsf{X}^\top \mathsf{X})^{-1}_{jj}}$$

i odrzucamy H_0 jeśli $|T| > t$. Jeśli H_0 jest prawdziwa, to $T \sim \mathrm{t}(n-p)$, więc próg odrzuceń jest odpowiednim kwantylem rozkładu t-Studenta, $t = \mathrm{t}_{1-\alpha/2}(n-p)$ (zamiast ustalać próg, możemy obliczać P-wartości testu).

Z drugiej strony, $\mathrm{H}_0 : \beta_p = 0$ jest szczególnym przypadkiem hipotezy rozpatrywanej we Wniosku 8.2.16 i może być użyty test F. Jeśli H_0 jest prawdziwa, to $F \sim \mathrm{F}(1, n-p)$. Zamiast p możemy wziąć dowolne j, zmieniając porządek współczynników. Oba testy, t i F, są *równoważne*, bo $T^2 = F$ (Zadanie 7).

Test t ma tę przewagę, że nadaje się do testowania H_0 przeciw alternatywie jednostronnej, powiedzmy $H_1 : \beta_j > 0$, podczas gdy F jest dostosowany do alternatywy dwustronnej $H_1 : \beta_j \neq 0$.

Analiza wariancji

Rozważmy zagadnienie porównywania kilku próbek. Chodzi o sprawdzenie, czy wszystkie pochodzą z tej samej populacji, czy też z populacji o różnych średnich. Najprostszy model zakłada, że mamy p niezależnych próbek z rozkładów normalnych:

$$\begin{array}{lll} \text{próbka 1:} & Y_{11}, \ldots, Y_{1n_1} \sim \mathrm{N}(\mu_1, \sigma^2); \\ \ldots \quad & \ldots \quad & \ldots \\ \text{próbka } j\text{:} & Y_{j1}, \ldots, Y_{jn_j} \sim \mathrm{N}(\mu_j, \sigma^2); \\ \ldots \quad & \ldots \quad & \ldots \\ \text{próbka } p\text{:} & Y_{p1}, \ldots, Y_{pn_p} \sim \mathrm{N}(\mu_p, \sigma^2). \end{array}$$

Zakłada się przy tym, że w poszczególnych próbkach wariancja σ^2 jest jednakowa, natomiast wartości średnie μ_j mogą być różne. Jest to szczególny przypadek modelu liniowego. W rzeczy samej, napiszmy

$$Y_{ji} = \mu_1 + \alpha_j + \varepsilon_{ji},$$

gdzie $\alpha_j = \mu_j - \mu_1$ dla $j = 1, \ldots, p$ oraz $\varepsilon_{ji} = Y_{ji} - \mu_j$. Oczywiście, $\varepsilon_{ji} \sim \mathrm{N}(0, \sigma^2)$ są niezależnymi zmiennymi losowymi. Wprowadźmy sztuczne, „*nieme*" zmienne objaśniające $x_1, \ldots, x_p$. Przyjmiemy *umownie*, że dla obserwacji z j-tej próbki mamy $x_1 = 1$, $x_j = 1$, zaś wszystkie inne zmienne x-owe są zerami. Obrazuje to taka tabelka:

próbka \ zmienne	x_1	x_2	$\ldots$	x_p
1	1	0	$\ldots$	0
2	1	1	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p	1	0	$\ldots$	1

Niech Y oznacza „długi" wektor o $n = \sum_{j=1}^{r} n_j$ współrzędnych, powstały przez ustawienie kolejnych próbek „jedna nad drugą". Podobnie określamy wektor błędów ε. „Nieme zmienne" umieścimy w macierzy X. Model możemy napisać w postaci macierzowej tak:

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ \vdots \\ Y_{p1} \\ \vdots \\ Y_{pn_p} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \vdots \\ \varepsilon_{p1} \\ \vdots \\ \varepsilon_{pn_p} \end{pmatrix},$$

a w skrócie

$$\underset{n\times 1}{Y} = \underset{n\times p}{\mathsf{X}} \cdot \underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon}$$

gdzie $\beta = (\mu_1, \alpha_2, \dots, \alpha_p)^\top$. Zauważmy, że w tym modelu μ_1 odgrywa rolę „wyrazu wolnego". Można sobie wyobrazić, że średnią μ_1 traktujemy jako „poziom bazowy" zaś pozostałe parametry uznajemy za „odchylenia od poziomu bazowego".

Hipoteza

$$\mathrm{H}_0 : \alpha_2 = \dots = \alpha_p = 0$$

sprowadza się do stwierdzenia, że wszystkie próbki pochodzą z tego samego rozkładu. Alternatywa jest bardzo ogólna: H_1 : nie jest prawdą, że $\alpha_2 = \dots = \alpha_p = 0$ (czyli nie wszystkie średnie μ_j są jednakowe).

Statystyka testowa F wyprowadzona w ogólnej sytuacji we Wniosku 8.2.16 przybiera dla modelu wielu próbek szczególnie prostą postać. Niech

$$\bar{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ji}$$

będzie średnią w j-tej grupie, zaś

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j} Y_{ji} = \frac{1}{n}\sum_{j=1}^{p} n_j \bar{Y}_j$$

oznacza średnią „globalną", obliczoną z połączonych próbek. Wprowadźmy oznaczenia

$$\mathrm{SST} = \sum_{j=1}^{r}\sum_{i=1}^{n_j}(Y_{ji} - \bar{Y})^2, \qquad \mathrm{SSB} = \sum_{j=1}^{r} n_j(\bar{Y}_j - \bar{Y})^2,$$

$$\mathrm{SSW} = \sum_{j=1}^{r}\sum_{i=1}^{n_j}(Y_{ji} - \bar{Y}_j)^2.$$

Te skróty są związane ze specyfiką modelu kilku próbek: SST jest *całkowitą* sumą kwadratów (ang. „Sum of Squares, Total") SSB jest sumą kwadratów *pomiędzy* próbkami (ang. „Between"), zaś SSW jest sumą kwadratów *wewnątrz* próbek (ang. „Within").

Rozpatrujemy tylko szczególny przypadek ogólnego modelu liniowego. Łatwo zauważyć związek naszych nowych oznaczeń z używanymi poprzednio. Mamy

$$\hat{Y}_0 = (\underbrace{\bar{Y}, \ldots, \bar{Y}}_{n}),$$

$$\hat{Y} = (\underbrace{\bar{Y}_1, \ldots, \bar{Y}_1}_{n_1}, \ldots, \underbrace{\bar{Y}_p, \ldots, \bar{Y}_p}_{n_p})^\top.$$

Stąd

$$\mathrm{SST} = \|Y - \hat{Y}_0\|^2, \quad \mathrm{SSB} = \|\hat{Y} - \hat{Y}_0\|^2, \quad \mathrm{SSW} = \mathrm{SSE} = \|Y - \hat{Y}\|^2$$

Otrzymujemy podstawową tożsamość analizy wariancji:

$$\mathrm{SST} = \mathrm{SSB} + \mathrm{SSW}.$$

Wiemy również, że $\mathrm{SSW} \sim \chi^2(n-p)$. Przy założeniu prawdziwości H_0 mamy $\mathrm{SSB} \sim \chi^2(p-1)$. Statystyka testowa przyjmuje postać

$$F = \frac{\mathrm{MSB}}{\mathrm{MSW}} = \frac{\mathrm{SSB}/(p-1)}{\mathrm{SSW}/(n-p)}.$$

Hipotezę H_0 odrzucamy, jeśli $F > \mathrm{F}_{1-\alpha}(p-1, n-p)$. W praktyce, zamiast ustalać próg odrzuceń, podaje się P-wartość testu.

Zwięzłą formą podsumowania wyników „analizy wariancji" jest taka tabelka:

Źródło zmienności	Sumy kwadratów	Stopnie swobody	Średnie kwadraty	F
między próbkami	SSB	$p-1$	$\mathrm{MSB} = \frac{\mathrm{SSB}}{p-1}$	$F = \frac{\mathrm{MSB}}{\mathrm{MSW}}$
wewnątrz próbek	SSW	$n-p$	$\mathrm{MSW} = \frac{\mathrm{SSW}}{n-p}$	
razem	SST	$n-1$	$\mathrm{MST} = \frac{\mathrm{SST}}{n-1}$	

8.2.19 Uwaga. Opisany powyżej sposób „zakodowania" modelu kilku próbek nie jest najbardziej naturalny. Możnaby zdefiniować „nieme" zmienne objaśniające $x_1, \ldots, x_p$ inaczej, według takiej tabelki:

próbka \ zmienne	x_1	x_2	$\ldots$	x_p
1	1	0	$\ldots$	0
2	0	1	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p	0	0	$\ldots$	1

Odpowiada to przyjęciu, że dla obserwacji z j-tej próbki mamy, $x_j = 1$ i wszystkie inne zmienne x-owe są zerami. Wtedy w roli wektora współczynników mielibyśmy po prostu wektor średnich w próbkach: $\beta = (\mu_1, \ldots, \mu_p)$. Interesującą nas hipotezę napisalibyśmy w postaci $\mathrm{H}_0 : \mu_1 = \cdots = \mu_p$. To byłoby bardziej symetryczne i eleganckie.

Nietrudno zauważyć, że oba sposoby kodowania są całkowicie równoważne. Są dwa powody, dla których wybraliśmy „mniej elegancki". Po pierwsze, hipoteza $\mathrm{H}_0 : \alpha_2 = \cdots = \alpha_p = 0$ mieści sie w schemacie rozważanym poprzednio. Po drugie, taki sposób kodowania jest użyty w R. Zobaczymy to na przykładzie.

8.2.20 PRZYKŁAD (Czy portfel ryzyk jest jednorodny?). Rozważmy trzech klientów towarzystwa ubezpieczeniowego. Powiedzmy, że są to firmy wynajmujące samochody. Wyobraźmy sobie, że roczne sumy szkód są takie:

Lata:	1	2	3	4	5	6	7	8	9	10	średnie
1 firma	10	15	16	14	8	17	11	13	14	12	13
2 firma	8	17	11	9	10	11	9	13			11
3 firma	9	18	14	15	11	17	16	12	14		14

„Średnia globalna" jest równa $(10/27) * 13 + (8/27) * 11 + (9/27) * 14 = 12.44$. Chcemy sprawdzić, czy wysokości szkód w trzech firmach są „istotnie różne". Formalnie, testujemy hipotezę zerową: „wartości oczekiwane wszystkich trzech próbek są równe". Oto fragment kodu w R:

```
> Y1 <- c(10,15,16,14,8,17,11,13,14,12)
> Y2 <- c(8,17,11,9,10,11,9,13)
> Y3 <- c(9,18,14,15,11,17,16,12,14)
> mean(Y1) [1] 13
> mean(Y2) [1] 11
> mean(Y3) [1] 14
> Y=c(Y1,Y2,Y3)
> grupa=c(rep(1,length(Y1)),rep(2,length(Y2)),rep(3,length(Y3)))
> grupa=factor(grupa)
```

Zwróćmy uwagę na sposób przygotowania danych do użycia funkcji `lm()`. „Zlepiamy" wszystkie próbki w „długi wektor" `Y`. Wektor `grupa` jest tej samej długości co `Y`, to znaczy `length(Y1)+length(Y2)+length(Y3)=10+8+9=27` i zawiera numery grup (próbek), z których pochodzą odpowiednie elementy `Y`. Należy przy tym koniecznie dać R-owi znać, że chcemy potraktować wektor `grupa` jako obiekt typu *czynnik*. To jest rola funkcji `factor()`. Wydruk „czynnika" zawiera nie tylko wartości, ale i listę „poziomów":

```
> grupa [1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3
Levels:  1 2 3
```

Teraz możemy zastosować funkcję `lm()` i otrzymujemy następujący wynik:

```
> fit=lm(Y~grupa)
> fit
Call:  lm(formula = Y ~ grupa)

Coefficients:
 (Intercept)  grupa2  grupa3
          13      -2       1
```

Dzięki temu, że `grupa` jest czynnikiem, R tworzy wartości „niemych" zmiennych, kodując „poziomy czynnika" w sposób opisany poprzednio. Grupa 1 jest potraktowana jako „bazowa" i odpowiada jej wyraz wolny. Współczynnki (α_2, α_3) związane z grupami 2 i 3 oznaczają odchylenia od poziomu bazowego.

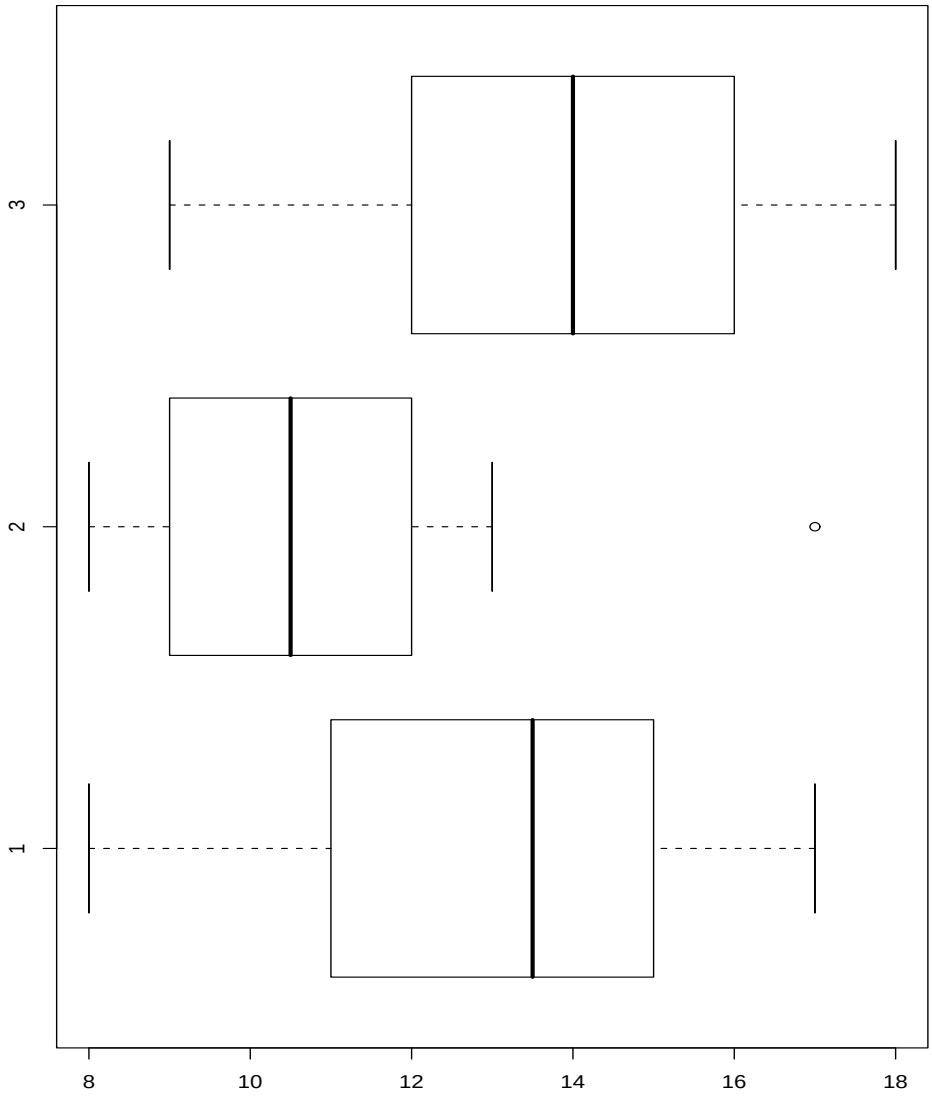
Tabelka analizy wariancji wygląda następująco:

```
> anova(fit)
Analysis of Variance Table
Response:  Y
            Df   Sum Sq  Mean Sq  F value  Pr(>F)
 grupa       2   39.185   19.593   2.3991  0.1122
 Residuals  24  196.000    8.167
```

Test F na poziomie istotności $\alpha = 0.05$ nie odrzuca hipotezy zerowej, bo odpowiednia P-wartość (obliczona z rozkładu F-Snedecora) jest równa 0.1122. Możemy przypuszczać, że wysokość zgłaszanych w przyszłości szkód będzie podobna dla wszystkich trzech firm. W każdym razie, nasze dane nie dają dostatecznych podstaw, by zwątpić w to przypuszczenie.

Dodajmy jeszcze komentarz na temat założeń, które są wymagane przy stosowaniu testu analizy wariancji. Możemy dość spokojnie przyjąć, że sumaryczne (lub średnie) szkody w kolejnych latach są zmiennymi losowymi o rozkładzie zbliżonym do normalnego. To jest pierwsze z podstawowych założeń modelu. Gorzej jest z drugim założeniem: o równości wariancji poszczególnych próbek. Jest ono uzasadnione właściwie tylko wtedy, gdy liczba ubezpieczonych samochodów dla trzech firm (i dla kolejnych lat) jest w przybliżeniu równa.

Możemy, całkiem heurystycznie, wizualnie porównać nasze trzy próbki przy pomocy wykresów pudełkowych. Zastosowanie funkcji `boxplot(Y1,Y2,Y3)` daje rezultat widoczny na rysunku.



Wydaje się, że różnice pomiędzy rozkładami szkód (średnimi) dla trzech firm są dość wyraźne. Formalny test prowadzi do przeciwnego wniosku.

Hipoteza o braku zależności

Wróćmy do liniowej regresji wielorakiej z wyrazem wolnym. Wektor współczynników zapisujemy w tej sytuacji jako $(\beta_0, \beta_1, \ldots, \beta_r)$, gdzie $r+1 = p$. Jeśli znikają wszystkie współczynniki funkcji regresji *z wyjątkiem wyrazu wolnego*, to wartości zmiennej Y nie są powiązane z wartościami zmiennych objaśniających x. Ważną kwestią jest więc weryfikacja hipotezy

$$\mathrm{H}_0 : \beta_1 = \cdots = \beta_r = 0 \text{ przeciw } \mathrm{H}_1 : \beta_1 \neq 0 \text{ lub } \cdots \text{ lub } \beta_r \neq 0.$$

W istocie, gdy nie ma podstaw do odrzucenia H_0, to model traci swoją użyteczność. Jest to tylko szczególny przypadek zagadnienia testowania hipotezy liniowej i możemy skorzystać z ogólnych wyników. Trzeba tylko uważnie liczyć stopnie swobody. Oprócz używanego już oznaczenia

$$\mathrm{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2,$$

wprowadźmy nowe nazwy sum kwadratów:

$$\mathrm{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2, \qquad \mathrm{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2.$$

SST nazywamy całkowitą sumą kwadratów, zaś SSR sumą kwadratów związana z regresją (ang. „Sum of Squares, Regression"). Zauważmy, że $(\bar{Y}, \ldots, \bar{Y})$ jest predykcją przy założeniu H_0, a $(\hat{Y}_1, \ldots, \hat{Y}_i)$ jest predykcją w „dużym modelu" z $r+1$ współczynnikami. Stąd natychmiast wynika tożsamość analizy wariancji,

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}.$$

Tę równość interpretuje się w taki sposób:

$$\begin{aligned} \text{całkowita zmienność } Y = &\ \text{zmienność „wyjaśniona regresją"} \\ &+ \ \text{zmienność „resztowa".} \end{aligned}$$

Pozostawimy Czytelnikowi wytłumaczenie intuicji stojących za tą sugestywną terminologią. *Współczynnikiem dopasowania* nazywamy

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}.$$

Zgodnie z przytoczoną wyżej interpretacją, R^2 jest „częścią zmienności, wyjaśnioną przez regresję". Zazwyczaj współczynnik dopasowania wyraża się w procentach. Im większe R^2, tym lepiej (estymowana) prosta regresji „pasuje" do punktów doświadczalnych – stąd nazwa.

Z ogólnej teorii wynika, że $\mathrm{SSE} \sim \chi^2(n-r-1)$. Przy założeniu prawdziwości H_0 mamy $\mathrm{SSR} \sim \chi^2(r)$. Statystyka testu Snedecora jest następująca

$$F = \frac{\mathrm{MSR}}{\mathrm{MSE}} = \frac{\mathrm{SSR}/r}{\mathrm{SSE}/(n-r-1)}.$$

Hipotezę H_0 odrzucamy, jeśli $F > \mathrm{F}_{1-\alpha}(r, n-r-1)$ lub, równoważnie, jeśli P-wartość testu jest poniżej α.

Tabelka „analizy wariancji" przybiera postać:

Źródło zmienności	Sumy kwadratów	Stopnie swobody	Średnie kwadraty	F
regresja	SSR	r	$\mathrm{MSR} = \frac{\mathrm{SSR}}{r}$	$F = \frac{\mathrm{MSR}}{\mathrm{MSE}}$
błąd	SSE	$n-r-1$	$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-r-1}$	
razem	SST	$n-1$	$\mathrm{MST} = \frac{\mathrm{SST}}{n-1}$	

Wartości statystyki F interpretuje się jako wskaźnik „istotnej zależności" zmiennej Y od zmiennych $x_1, \ldots, x_r$. Mówi się w żargonie statystycznym, że zależność „jest istotna na poziomie α", jeśli test F na tym poziomie istotności odrzuca hipotezę o braku zależności.

8.3 Losowa zmienna objaśniająca

Przedstawimy *inny* model statystyczny, który też wiążemy z nazwą *regresji liniowej*. Tak jak na początku tego rozdziału, chcemy opisać zależność między dwiema zmiennymi (cechami), mierzonymi dla wielu obiektów (przypadków). Różnica polega na tym, że teraz *obie* zmienne potraktujemy jako *losowe*. Oznaczymy je dużymi literami X i Y, zgodnie z ogólną konwencją przyjętą w tych wykładach. Dla prostoty ograniczymy się do przypadku, gdy zmienne X i Y są jednowymiarowe. Uogólnienie na przypadek wielowymiarowej zmiennej objaśniającej X nie jest trudne, ale zajęłoby zbyt dużo miejsca.

Prosta regresja liniowa

Dane mają postać par

$$(X_1, Y_1), \ldots, (X_n, Y_n).$$

Przyjmujemy, że spełniony jest jeden z następujących warunków. Bardziej ograniczający z nich jest taki.

8.3.1 Założenie. $(X_1, Y_1), \ldots, (X_n, Y_n)$ *są niezależnymi wektorami losowymi o jednakowym rozkładzie normalnym* $\mathrm{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \varrho)$.

Słabszy warunek jest następujący.

8.3.2 Założenie. $(X_1, Y_1), \ldots, (X_n, Y_n)$ *są niezależnymi wektorami losowymi o jednakowym rozkładzie,* $\mathbb{E}X_i = \mu_X$, $\mathbb{E}Y_i = \mu_Y$, $\mathrm{Var}X_i = \sigma_X^2$, $\mathrm{Var}Y_i = \sigma_Y^2$, $\mathrm{corr}(X_i, Y_i) = \varrho$.

W pewnym sensie, Założenie 8.3.1 jest odpowiednikiem 8.2.5 podczas gdy Założenie 8.3.2 pełni podobną rolę co 8.2.6. W każdym razie zakładamy, że pary są prostą próbką losową z dwuwymiarowego rozkładu wektora losowego (X, Y).

Na początek rozważmy zależność pomiędzy dwiema zmiennymi losowymi na poziomie probabilistycznym, to znaczy zakładając znajomość łącznego rozkładu prawdopodobieństwa (X, Y).

8.3.3 Stwierdzenie. *Jeżeli* $(X, Y) \sim \mathrm{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \varrho)$ *to*

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X,$$

gdzie

$$\beta_1 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}X} = \frac{\sigma_Y}{\sigma_X}\varrho, \qquad \beta_0 = \mathbb{E}Y - \beta_1 \mathbb{E}X = \mu_Y - \beta_1 \mu_X.$$

Ponadto, zmienne losowe X *i* $\varepsilon = Y - (\beta_0 + \beta_1 X)$ *są niezależne,* $\varepsilon \sim \mathrm{N}(0, \sigma_Y^2(1 - \varrho^2))$ *i* $\mathrm{Var}(Y|X) = (1 - \varrho^2)\sigma_Y$.

Dowód. Można wyprowadzić wzór na $\mathbb{E}(Y|X)$ posługując się łączną gęstością rozkładu (X,Y) i wyliczając gęstość warunkową (Zadanie 8). Jeśli jednak wyjdziemy od wzorów na współczynniki β_0 i β_1 to sprawdzenie wszystkich faktów wymienionych w stwierdzeniu jest bardzo proste. Wystarczy zauważyć, że X, ε mają łączny rozkład normalny (dlaczego?). Sprawdzamy, że $\mathbb{E}\varepsilon = 0$, $\mathrm{Cov}(X,\varepsilon) = 0$, stąd wnioskujemy, że X i ε są niezależne. Równanie $\mathrm{Var}Y = \beta_1^2 \mathrm{Var}X + \mathrm{Var}\varepsilon$ pozwala obliczyć $\mathrm{Var}\varepsilon$. Wreszcie, $\mathrm{Var}(Y|X) = \mathrm{Var}(\varepsilon|X) = \mathrm{Var}\varepsilon$ ze względu na niezależność. □

Wiadomo, że wśród dowolnych funkcji $h(X)$ traktowanych jako predyktory zmiennej losowej Y, funkcja $\mu(X) = \mathbb{E}(Y|X)$ minimalizuje błąd średniokwadratowy:

$$\mathbb{E}(Y - \mu(x))^2 \leqslant \mathbb{E}(Y - h(x))^2.$$

Jeśli (X,Y) mają rozkład normalny, to $\mu(X)$ jest funkcją liniową.

Zawsze możemy zapisać w podobnej postaci jak w

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

gdzie

$$\beta_1 = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}X} = \frac{\sigma_Y}{\sigma_X}\varrho, \qquad \beta_0 = \mathbb{E}Y - \beta_1 \mathbb{E}X = \mu_Y - \beta_1 \mu_X.$$

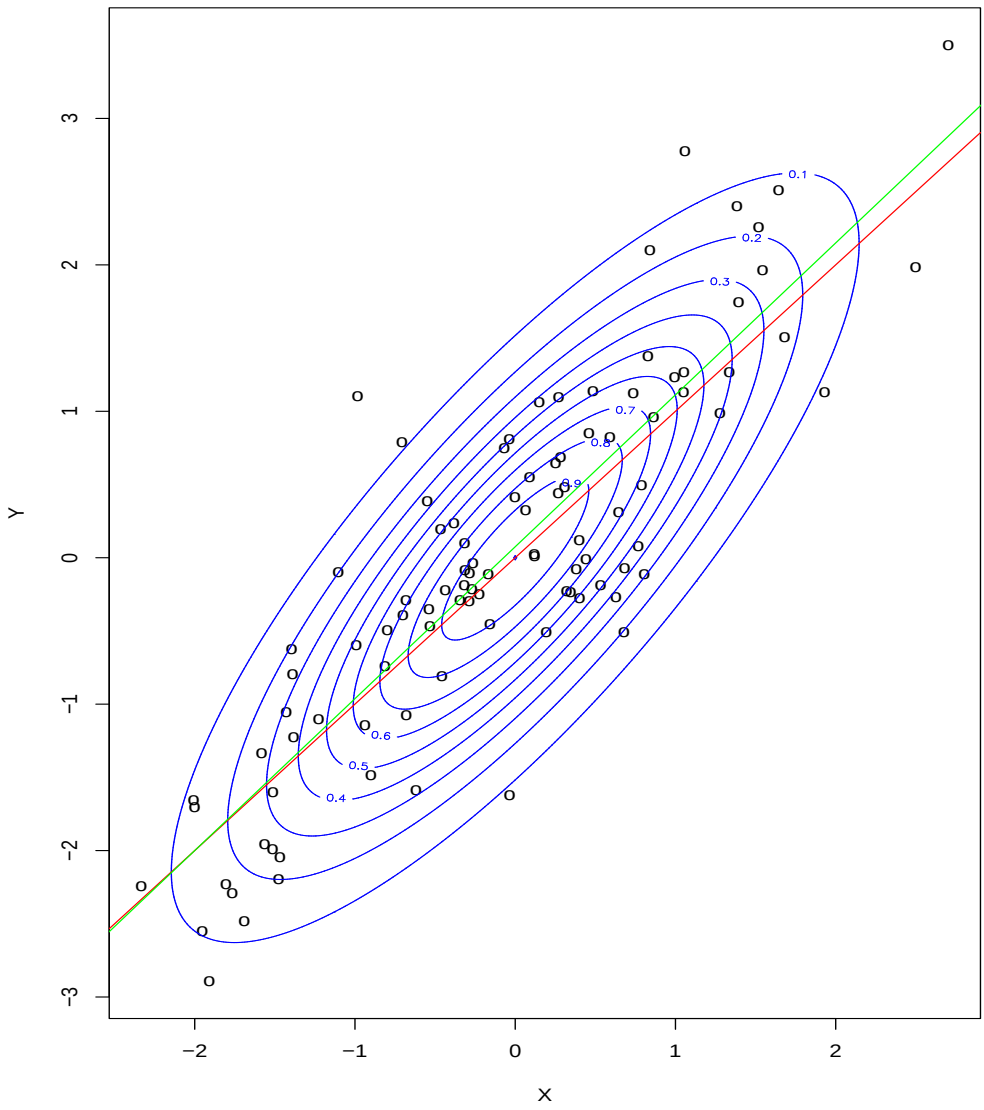
Współczynniki β_0 i β_1 w powyższym wzorze są wybrane tak, że liniowa funkcja $\beta_0 + \beta_1 X$ *najlepiej przybliża* zmienną Y w sensie błędu średniokwadratowego: jak wiemy,

$$\mathbb{E}\varepsilon^2 = \mathbb{E}\big(Y - \beta_0 - \beta_1 X\big)^2 \leq \mathbb{E}\big(Y - b_0 - b_1 X\big)^2$$

dla dowolnych b_0 i b_1. Jeśli $(X,Y) \sim \mathrm{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \varrho)$, to można wzmocnić to stwierdzenie. Mamy wtedy $\beta_0 + \beta_1 X = \mathbb{E}(Y|X)$. Funkcja $\mu(X) = \mathbb{E}(Y|X)$ minimalizuje błąd średniokwadratowy wśród dowolnych (niekoniecznie liniowych) funkcji $h(X)$:

$$\mathbb{E}\varepsilon^2 = \mathbb{E}\big((Y - \mu(x))^2\big) \leq \mathbb{E}\big((Y - h(x))^2\big).$$

Jeśli β_0 i β_1 są określone jak wyżej, to $\mathbb{E}\varepsilon = 0$, $\mathrm{Var}\varepsilon = \sigma_Y^2(1 - \varrho^2)$ i $\mathrm{Cov}(X,\varepsilon) = 0$ (dla $\varepsilon = Y - \beta_0 - \beta_1 X$). Jeżeli założymy dodatkowo, że łączny rozkład zmiennych X i Y jest normalny, to X i ε są niezależne, $\varepsilon \sim \mathrm{N}(0, \sigma_Y^2(1 - \varrho^2))$.



Rysunek przedstawia dwuwymiarowy rozkład normalny $\mathrm{N}(0, 0, \sigma_X^2 = 1, \sigma_Y^2 = 1.25, \rho = 0.8944272)$, poziomice gęstości, funkcję regresji ($y = x$) oraz próbkę z tego rozkładu i estymowaną (metodą najmniejszych kwadratów) funkcję regresji.

Interpretacja rozpatrywanego przez nas modelu jest bardzo zbliżona do tej opisanej w poprzednich podrozdziałach. Różnica jest tylko w tym, że „nie kontrolujemy zmiennej objaśniającej"; wartości X są *wylosowane* zgodnie z pewnym rozkładem prawdopodobieństwa, a nie dowolnie wybierane przez eksperymentatora. Zaraz przekonamy się, że w naszym nowym modelu zasadnicze estymatory *wyglądają* zupełnie tak samo, jak w modelu z deterministycznymi wartościami zmiennej objaśniającej. Jeśli nie będziemy pamiętać, że *matematyczna* treść wzorów jest nieco inna, to może powstać pewne zamieszanie.

Napiszemy „oczywiste" estymatory parametrów μ_X, μ_Y, σ_X^2, σ_Y^2 i ϱ:

$$\hat{\mu}_X = \bar{X} = \frac{1}{n}\sum X_i, \qquad \hat{\mu}_Y = \bar{Y} = \frac{1}{n}\sum Y_i,$$
$$\hat{\sigma}_X^2 = \frac{1}{n}\sum (X_i - \bar{X})^2, \qquad \hat{\sigma}_Y^2 = \frac{1}{n}\sum (Y_i - \bar{Y})^2,$$
$$\hat{\varrho} = R = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}.$$

Są to po prostu próbkowe odpowiedniki estymowanych wielkości lub, inaczej mówiąc, estymatory otrzymane *metodą momentów*. Jeśli spełnione jest założenie (N') to są to estymatory *największej wiarogodności*. Można to pokazać bez trudu, ale wyprowadzenie zajmuje sporo miejsca, więc je pominiemy. Ponieważ $\beta_1 = \varrho\sigma_Y/\sigma_X$ i $\beta_0 = \mu_Y - \beta_1\mu_X$, więc za estymatory współczynników regresji liniowej przyjmiemy

$$\hat{\beta}_1 = \hat{\varrho}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2},$$
$$\hat{\beta}_0 = \hat{\mu}_Y - \hat{\beta}_1\hat{\mu}_X = \bar{Y} - \hat{\beta}_1\bar{X}.$$

Są to nasi starzy znajomi: estymatory najmniejszych kwadratów, wyprowadzone w inny sposób i przy innych założeniach w 9.2! Być może, obecnie rozpatrywany model z losowymi X-ami pozwala łatwiej zinterpretować ENK. Niestety, w tym modelu nie możemy już mówić, że ENK są estymatorami liniowymi (bo zależą nieliniowo od obserwacji X-ów).

Zwróćmy uwagę, że statystyka R jest, w rozpatrywanym teraz modelu, po prostu *próbkowym współczynnikiem korelacji*. Uzasadnia to terminologię

wprowadzoną w Zadaniu 6. Hipoteza o braku zależności liniowej przybiera postać

$$\mathrm{H}_0 : \varrho = 0.$$

Jest to, formalnie, inna hipoteza niż rozpatrywana w Podrozdziale 8.2. Jej intuicyjny sens jest jednak taki sam. Zauważmy, że $\varrho = 0$ wtedy i tylko wtedy, gdy $\beta_1 = 0$. Jeśli łączny rozkład zmiennych X i Y jest normalny, to H_0 stwierdza *niezależność* tych zmiennych.

8.3.4 Stwierdzenie. *Zakładamy, że spełniony jest warunek* **??**. *Jeśli* H_0 *jest prawdziwa, to statystyka*

$$\frac{R}{\sqrt{1-R^2}}\sqrt{n-2}$$

ma rozkład Studenta $\mathrm{t}(n-2)$.

Dowód. Będziemy stosowali te same oznaczenia sum kwadratów SST, SSR, SSE i $\mathrm{SS_X}$, co w Podrozdziale 8.2 (z oczywistą zamianą małych literek x na duże X). Mamy $\beta_1(X_i - \bar{X}) = (\hat{Y}_i - \bar{Y})$, gdzie $\hat{\beta}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. Stąd otrzymujemy $\beta_1^2 = \mathrm{SSR}/\mathrm{SS_X}$. Ponieważ $\beta_1^2 = R^2\mathrm{SST}/\mathrm{SS_X}$, więc $R^2 = \mathrm{SSR}/\mathrm{SST}$ i $1 - R^2 = \mathrm{SSE}/\mathrm{SST}$. Zatem

$$\frac{R^2}{1-R^2}(n-2) = \frac{\mathrm{SSR}}{\mathrm{SSE}/(n-2)}.$$

Dla *ustalonych* wartości $X_i = x_i$, zmienna losowa po prawej stronie ($F = \mathrm{MSR}/\mathrm{MSE}$) ma, jak wiemy, rozkład (warunkowy) Snedecora $\mathrm{F}(1, n-2)$. Jeśli rozkład warunkowy nie zależy od wartości x_i, to jest on równy rozkładowi brzegowemu (bezwarunkowemu). Pokazaliśmy w ten sposób, że *kwadrat* rozpatrywanej przez nas statystyki ma rozkład $\mathrm{F}(1, n-2)$. Ponieważ sama statystyka ma rozkład symetryczny względem zera, musi to być rozkład $\mathrm{t}(n-2)$. □

Tak więc, do weryfikacji $\mathrm{H}_0 : \varrho = 0$ przeciw $\mathrm{H}_1 : \varrho \neq 0$ możemy stosować test oparty na statystyce $(n-2)R^2/(1-R^2) \sim \mathrm{F}(1, n-2)$. Jest to faktycznie *ten sam test* F, który pojawił się w Podrozdziale 8.2. Rzecz jasna, Stwierdzenie pozwala też zbudować testy *jednostronne*, powiedzmy $\mathrm{H}_0 : \varrho = 0$ przeciw $\mathrm{H}_1 : \varrho > 0$. Korzystamy wtedy ze statystyki o rozkładzie $t(n-2)$.

Rozumowanie w dowodzie ostatniego stwierdzenia jest dość charakterystyczne: korzystamy z gotowego wyniku dla *ustalonych* X*-ów*, przy tym obliczony *w innym modelu* rozkład prawdopodobieństwa interpretujemy jako rozkład *warunkowy*.

Na zakończenie wspomnijmy, że model regresji *wielorakiej* z losowymi zmiennymi objaśniającymi: $Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon = \beta_0 + \beta^\top X + \varepsilon$ nie wymaga istotnie nowych pojęć. Należy rozpatrzyć łączny rozkład $p+1$-wymiarowego wektora losowego $(X_1, \ldots, X_p, Y)$. Współczynniki regresji β_0 i β dobieramy tak, by $\mathbb{E}(Y - \beta_0 - \beta^\top X) = \min$. Poprzednie rozważania ulegają modyfikacjom na tyle oczywistym, że szkoda na nie czasu i miejsca.

8.4 Zadania

Poniższe zadania dotyczą modelu z jedną zmienną objaśniającą i wyrazem wolnym (prosta regresja liniowa).

1. Wyprowadzić wzory (8.2.1) na $\hat{\beta}_0$ i $\hat{\beta}_1$.
2. Wyprowadzić bezpośrednio wzory na $\mathrm{Var}\hat{\beta}_1$ i $\mathrm{Var}\hat{\beta}_0$.
3. Pokazać, że zmienne losowe $\bar{Y}$ i $\hat{\beta}_1$ są niezależne.
4. Wyprowadzić bezpośrednio wzory na $\mathrm{Var}\hat{Y}^*$ i $\mathrm{Var}(Y^* - \hat{Y}^*)$.
   Wskazówka: Skorzystać z poprzednich zadań.
5. Udowodnić bezpośredno (nie korzystając z geometrycznych rozważań w przestrzeni $\mathbb{R}^n$) podstawową tożsamość analizy wariancji:
$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2.$$
6. *Współczynnik korelacji* [5] R określamy wzorem
$$R = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (Y_i - \bar{Y})^2}}.$$

[5] Związek R z pojęciem korelacji zmiennych losowych staje się jasny, gdy rozpatrujemy model z *losową* zmienną objaśniającą. W modelu z deterministycznym x, przyjmijmy po prostu, że „tak się mówi".

Pokazać, że kwadrat współczynnika korelacji jest współczynnikiem dopasowania.

7. Udowodnić fakt sformułowany w Uwadze 8.2.18: $T^2 = F$.

   Pokazać, że test F odrzuca $\mathrm{H}_0 : \beta_1 = 0$ (na poziomie istotności α) wtedy i tylko wtedy, gdy przedział ufności dla β_1 (na poziomie $1 - \alpha$) nie zawiera zera.

   Następujące zadania dotyczą modelu z losową zmienną objaśniającą (Podrozdział 8.3).

8. Wyprowadzić wzór na $\mathbb{E}(Y|X)$ w Stwierdzeniu 8.3.3 przez obliczenie gęstości warunkowej $f(y|x)$.

Część V

Podejście bayesowskie

Rozdział 9

Bayesowskie modele statystyczne

9.1 Wstęp

Podejście bayesowskie polega na tym, że nieznany parametr θ traktujemy jako realizację *zmiennej losowej* ϑ. Rozkład prawdopodobieństwa tej zmiennej losowej przyjęto nazywać rozkładem *a priori*, ponieważ wyraża on naszą wiedzę (lub przekonania) o parametrze przed zaobserwowaniem zmiennej losowej X (bez brania pod uwagę danych).

9.1.1 PRZYKŁAD. Rozpatrzmy sytuację nieco sztuczną, ale dobrze ilustrującą nowy punkt widzenia. Załóżmy, że są dwa typy kierowców: „Ostrożni" i „Ryzykanci". Kierowca „O" w ciągu roku powoduje szkodę z prawdopodobieństwem 0.1 a „R" z prawdopodobieństwem 0.4. Niech $X = 1$ oznacza szkodę, a $X = 0$ jej brak. Napiszmy

$$\mathbb{P}(X = 1|O) = 0.1, \qquad \mathbb{P}(X = 1|R) = 0.4.$$

Przypuśćmy, że populacja kierowców składa się w 80% z typu „O" i w 20% z typu „R". Towarzystwo ubezpieczeniowe podpisując nową umowę nie wie, jakiego typu jest klient. Szanse natrafienia na „Ostrożnego" i „Ryzykanta" oceniamy przed zawarciem umowy na

$$\mathbb{P}(O) = 0.80, \qquad \mathbb{P}(R) = 0.20. \qquad (*)$$

Prawdopodobieństwo zgłoszenia szkody obliczamy ze wzoru na prawdopodobieństwo całkowite:

$$\begin{aligned}\mathbb{P}(X=1) &= \mathbb{P}(X=1|O)\mathbb{P}(O) + \mathbb{P}(X=1|R)\mathbb{P}(R) \\ &= 0.1 * 0.80 + 0.4 * 0.20 = 0.16.\end{aligned}$$

Po upływie roku, dysponujemy obserwacją X. Jeśli klient zgłosił szkodę, to elementarny wzór Bayesa pozwala obliczyć, że

$$\mathbb{P}(R|X=1) = \frac{\mathbb{P}(X=1|R)\mathbb{P}(R)}{\mathbb{P}(X=1)} = 0.5. \qquad (**)$$

Zbudowaliśmy model *dwuetapowego* doświadczenia losowego. Pierwszy etap polega na wylosowaniu klienta zgodnie z rozkładem *a priori* (*). Wynik pierwszego etapu jest nieznany. Obserwujemy wynik drugiego etapu, wystąpienie lub brak szkody, czyli zmienną X i na tej podstawie obliczamy prawdopodobieństwo *a posteriori* (**).

Pierwszy etap interpretujemy jako wylosowanie parametru rozkładu prawdopodobieństwa rządzącego drugim etapem. Aby to podkreślić, użyjmy nieco innych oznaczeń: $\mathbb{P}(X=1|\vartheta=\theta)=\theta$ dla $\theta \in \{0.1, 0.4\}$ oraz $\mathbb{P}(\vartheta=0.1) = \mathbb{P}(O)=0.8$ i $\mathbb{P}(\vartheta=0.4)=\mathbb{P}(R)=0.8$.

9.2 Rozkłady *a priori* i *a posteriori*

9.2.1 DEFINICJA. ***Model bayesowski*** *jest modelem statystycznym, w którym dodatkowo dany jest rozkład prawdopodobieństwa* Π *na przestrzeni parametrów* Θ*, zwany rozkładem* ***a priori****.*

W dalszych rozważaniach utożsamiamy rozkłady prawdopodobieństwa z ich gęstościami. W statystyce bayesowskiej określamy, oprócz rodziny gęstości $\{f_\theta : \theta \in \Theta\}$ na przestrzeni $\mathcal{X}$, również gęstośc π na przestrzeni Θ, która odpowiada rozkładowi Π. Zwykle jest to albo gęstość względem miary Lebesgue'a albo względem miary liczącej, w zależności od kontekstu. Prześledźmy najpierw konstrukcję modelu formalnie, z matematycznego punktu widzenia.

Rozpatrujemy parę zmiennych losowych ϑ i X. Łączną gęstość tych zmiennych *definiujemy* wzorem

$$f(\theta, x) = \pi(\theta) f_\theta(x), \qquad (\theta \in \Theta, x \in \mathcal{X}).$$

W ten sposób określamy *jeden* rozkład prawdopodobieństwa, nazwijmy go $\mathbb{P}$, *na nowej przestrzeni probabilistycznej* $\Theta \times \mathcal{X}$. Jeśli zmienne X i ϑ są dyskretne, to przez łączną gęstość rozumiemy $f(\theta, x) = \mathbb{P}(\vartheta = \theta, X = x)$. Będziemy też rozważali modele w których jedna ze zmiennych X i ϑ jest dyskretna a druga ma rozkład absolutnie ciągły. Symbole $\mathbb{E}$, Var, f (bez wskaźnika θ) będą odtąd oznaczały wartość oczekiwaną, wariancję i gęstość względem rozkładu $\mathbb{P}$. Zauważmy, że teraz f_θ staje się *warunkową* gęstością zmiennej losowej X dla $\vartheta = \theta$:

$$f_\theta(x) = \frac{f(\theta, x)}{\pi(\theta)} = f(x|\theta).$$

Jeśli dana jest wartość naszej obserwacji i wiemy, że $X = x$, to możemy przy pomocy znanego *wzoru Bayesa* policzyć rozkład *warunkowy* losowego parametru ϑ. Jest to tak zwany rozkład *a posteriori*. Gęstość tego rozkładu oznacza się czasami przez π_x.

9.2.2 TWIERDZENIE (Wzór Bayesa). *Rozkład* ***a posteriori*** *parametru ϑ, dla danej obserwacji $X = x$, ma gęstość*

$$\pi_x(\theta) = f(\theta|x) = \frac{\pi(\theta) f_\theta(x)}{f(x)},$$

gdzie

$$f(x) = \begin{cases} \int_\Theta \pi(\theta) f_\theta(x) \mathrm{d}\theta & \text{w przypadku zmiennej ciągłej;} \\ \sum_{\theta \in \Theta} \pi(\theta) f_\theta(x) & \text{w przypadku zmiennej ciągłej.} \end{cases}$$

Gęstość f opisuje *rozkład brzegowy* zmiennej losowej X w modelu bayesowskim. W tym kontekście mówi się f jest *mieszanką* wyjściowych gęstości f_θ. W istocie, możemy traktować f jako „średnią ważoną" funkcji f_θ, z „funkcją wagową" π.

W typowych prostych modelach bayesowskich rozkład *a priori* jest tak dobrany do rozkładu obserwacji, aby wyliczenie rozkładu *a posteriori* było bardzo łatwe. Są to tak zwane sprzężone rodziny rozkładów. Mianownik $f(x)$ we wzorze Bayesa jest dość skomplikowany, ale nie zawsze musimy go obliczać. Interesuje nas zależność gęstości a posteriori od θ, zatem pomijając czynniki nie zależące od θ (choć być może zawierające x) przepiszemy wzór Bayesa w bardziej przyjaznej postaci:

$$\pi_x(\theta) f(\theta|x) \propto \pi(\theta) f_\theta(x).$$

Symbol $\propto$ oznacza, że lewa i prawa strona są proporcjonalne jako funkcje θ. Podobnie postąpimy w następnych przykładach.

Podamy teraz bayesowskie wersje modeli z Przykładów 2.1.5, 2.1.8 i 2.1.10.

9.2.3 PRZYKŁAD (Rozkład dwumianowy i rozkład *a priori* beta). Załóżmy, że obserwujemy wyniki n prób w schemacie Bernoulliego z nieznanym prawdopodobieństwem sukcesu θ. Rozkład prawdopodobieństwa zero-jedynkowych zmiennych $X_1, \ldots, X_n$ jest taki jak w Przykładzie 2.1.5. Liczba sukcesów $S = \sum X_i$ m rozkład dwumianowy $\mathrm{Bin}(n, \theta)$. Wygodnie przyjąć, że rozkład *a priori* jest rozkładem $\mathrm{Beta}(\alpha, \beta)$:

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad (0 < \theta < 1).$$

Rozkład *a posteriori* ma gęstość

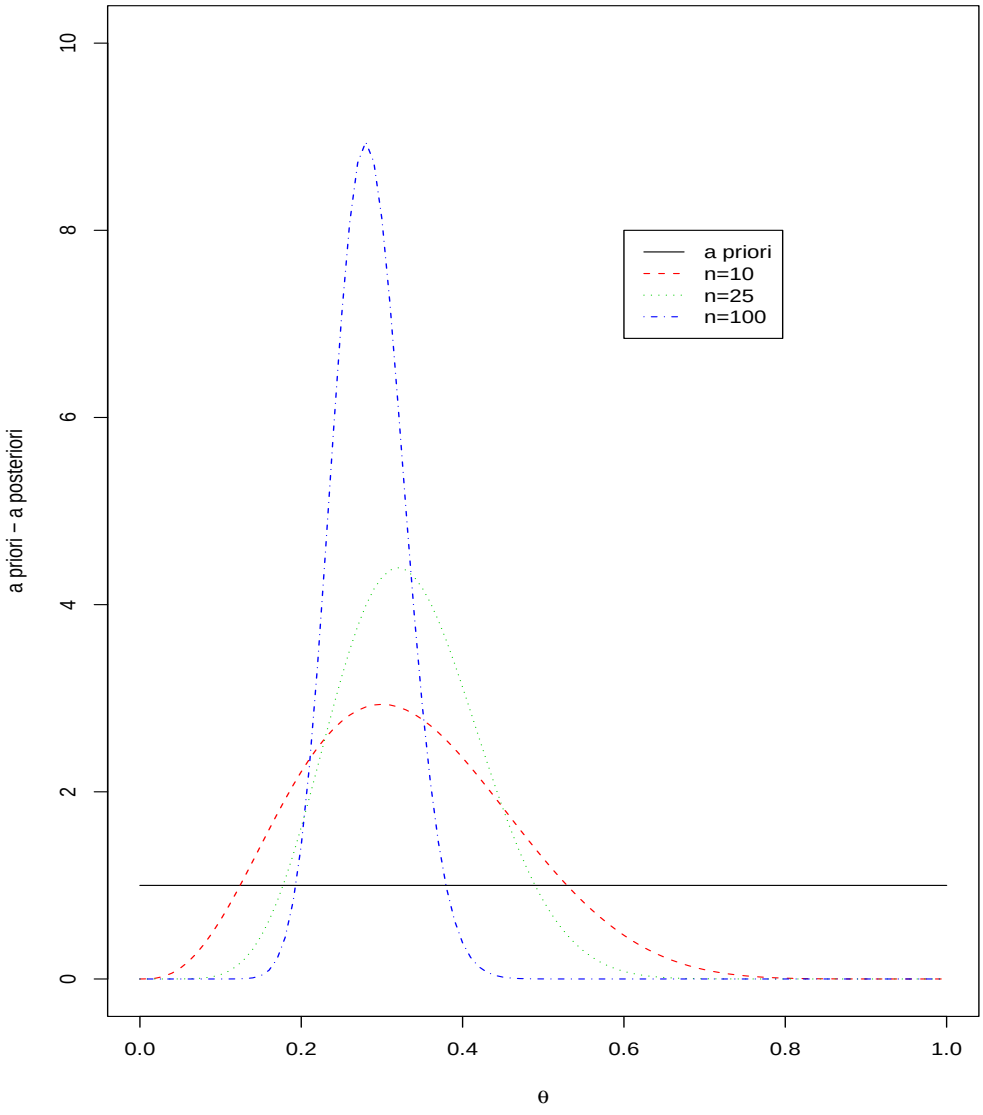
$$\pi_{x_1,\ldots x_n}(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^s(1-\theta)^{n-s} = \theta^{\alpha+s-1}(1-\theta)^{\beta+n-s-1},$$

gdzie $s = \sum x_i$. Rozkład *a posteriori* jest więc rozkładem $\mathrm{Beta}(\alpha+s, \beta+n-s)$ i zależy tylko od wartości zmiennej losowej $S = \sum X_i$.

Zauważmy, że jednostajny rozkład *a priori* należy do rodziny rozkładów beta: $U(0,1) = \mathrm{Beta}(\alpha, \beta)$, dla $\alpha = \beta = 1$. Ten rozkład wydaje się dobrze modelować „całkowitą niewiedzę na temat parametru θ", ale jest to interpretacja kontrowersyjna[1].

[1] Jeśli „nic nie wiemy o parametrze θ" to nic nie wiemy o θ^2. A więc $\vartheta \sim U(0,1)$ czy $\vartheta^2 \sim U(0,1)$?

Poniższy rysunek przedstawia gęstości jednostajnego rozkładu *a priori* i rozkładów *a posteriori* w Przykładzie 9.2.3 dla kilku rozmiarów próbki.



9.2.4 PRZYKŁAD (Rozkład Poissona i rozkład *a priori* gamma). Załóżmy, że $X_1, \ldots, X_n$ jest próbką z rozkładu Poiss(θ), tak jak w Przykładzie 2.1.8. Rozważmy rozkład a priori Gamma(α, λ):

$$\pi(\theta) \propto \theta^{\alpha-1}\mathrm{e}^{-\lambda\theta}, \qquad (\theta > 0).$$

Rozkład *a posteriori* ma gęstość

$$\pi_{x_1,\ldots x_n}(\theta) \propto \theta^{\alpha-1}\mathrm{e}^{-\lambda\theta}\mathrm{e}^{-n\theta}\theta^s = \theta^{\alpha+s-1}\mathrm{e}^{-(\lambda+n)\theta},$$

gdzie $s = \sum x_i$. Rozkład *a posteriori* jest więc rozkładem Gamma($\alpha+s, \lambda+n$) i zależy tylko od wartości zmiennej losowej $S = \sum X_i$.

Rozważany przez nas model znajduje zastosowanie w matematyce ubezpieczeniowej i tak zwanej „teorii zaufania". Interpretacja jest podobna jak w Przykładzie 9.1.1. Wyobraźmy sobie, że zmienne losowe X_i oznaczają liczby szkód dla konkretnego klienta w kolejnych latach. Parametr θ jest średnią liczbą roszczeń przypadających na rok. Każdemu klientowi odpowiada inna wartość parametru θ. Rozkład *a priori* opisuje „rozrzut" tego parametru w populacji klientów. Zgłaszanie się klientów uznajemy za zjawisko przypadkowe. Każda nowa umowa jest *dwuetapowym* doświadczeniem losowym. W *pierwszym etapie* pojawia się realizacja θ zmiennej losowej ϑ. W *drugim etapie* obserwujemy liczby szkód, czyli realizacje x_i zmiennych losowych X_i. Parametr θ jest już ustalony, ale nieznany. Nasz stan wiedzy (lub raczej stopień niewiedzy) o zmiennej θ po zaobserwowaniu liczb $x_1, \ldots, x_n$ opisuje rozkład *a posteriori*.

9.2.5 PRZYKŁAD (Rozkład normalny i rozkład *a priori* normalny). Niech zmienne $X_1, \ldots, X_n$ będą próbką z rozkładu normalnego N(μ, σ^2). Załóżmy, że σ^2 jest znane, zaś nieznany parametr μ ma rozkład *a priori* normalny N(m, v^2), czyli

$$\pi(\mu) \propto \exp\left[-\frac{1}{2v^2}(\mu - m)^2\right].$$

Możemy napisać gęstość *a posteriori*:

$$\pi_{x_1,\ldots,x_n}(\mu) \propto \exp\left[-\frac{1}{2v^2}(\mu - m)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right],$$

czyli

$$\pi_{x_1,\ldots,x_n}(\mu) \propto \exp\left[-\frac{nv^2 + \sigma^2}{2\sigma^2 v^2}\Big(\mu - \frac{nv^2\bar{x} + \sigma^2 m}{nv^2 + \sigma^2}\Big)^2\right]$$

Oczywiście, $\bar{x}$ oznacza $\sum x_i/n$. Wykonaliśmy tu znaną ze szkoły średniej operację sprowadzania trójmianu kwadratowego do postaci kanonicznej. Wyrazy wolne zostają „pochłonięte" przez symbol $\propto$. Dodanie stałej do argumentu funkcji wykładniczej jest tym samym, co pomnożenie tej funkcji przez stałą. Widać już, że rozkład *a posteriori* jest normalny,

$$\mathrm{N}\Big(\frac{nv^2\bar{x}+\sigma^2 m}{nv^2+\sigma^2}, \frac{\sigma^2 v^2}{nv^2+\sigma^2}\Big).$$

Zreasumujemy nasze rozważania i dokładniej omówimy *interpretację* modelu bayesowskiego. Jeśli rozkład *a priori* wyraża tylko przekonania statystyka i jest wybrany arbitralnie, to obliczone na jego podstawie prawdopodobieństwo ma charakter *subiektywnej* oceny szans. To jest klasyczny punkt widzenia teorii bayesowskiej. Wspomnieliśmy w Przykładzie 9.2.3 o związanych z tym trudnościach. W wielu zastosowaniach rozkład *a priori* ma inną, bardziej obiektywną interpretację, tak jak w Przykładzie 9.2.4. Łączny rozkład prawdopodobieństwa zmiennych losowych ϑ i X jest probabilistycznym modelem dwuetapowego doświadczenia losowego, w którym *obserwowujemy tylko wynik drugiego etapu.*

9.3 Warunkowa niezależność i dostateczność

W modelach bayesowskich ważną rolę gra pojęcie warunkowej niezależności. Rozważmy 3 zmienne losowe X, Y i Z, określone na tej samej przestrzeni probabilistycznej (przestrzenie wartości być być różne). Dla uproszczenia załóżmy, że istnieje łączna gęstość $f(x,y,z)$, albo „w zwykłym sensie" albo dyskretna. Zmienne X i Y są warunkowo niezależne pod warunkiem Z jeśli

$$f(x,y|z) = f(x|z)f(y|z)$$

dla (prawie)[2] wszystkich x, y i z. Łatwo sprawdzić, że warunkowa niezależność jest równoważna każdemu z dwóch następujących warunków, spełnionemu dla (prawie) wszystkich x, y i z:

$$f(x|y,z) = f(x|z), \qquad f(y|x,z) = f(y|z).$$

[2]Gęstości względem miary Lebesgue'a są jednoznacznie zdefiniowane prawie wszędzie. W przypadku zmiennych dyskretnych można pominąć zastrzeżenie „prawie".

Uogólnienie na przypadek większej liczby zmiennych nie przedstawia trudności.

9.3.1 PRZYKŁAD. W Przykładach 9.2.3, 9.2.4 i 9.2.5 zakładaliśmy warunkową niezależność obserwacji $X_1, \ldots, X_n$ przy danej wartości $\vartheta = \theta$:

$$f(\theta, x_1, \ldots, x_n) = \pi(\theta) f(x_1|\theta) \cdots f(x_n|\theta),$$

co jest równoważne

$$f(x_1, \ldots, x_n|\theta) = f(x_1|\theta) \cdots f(x_n|\theta).$$

Zauważmy, że obserwacje $X_1, \ldots, X_n$ w świecie bayesowskim *nie są* niezależne, bo brzegowa gęstość $f(x_1, \ldots, x_n)$ *nie jest* iloczynem $f(x_1) \cdots f(x_n)$. Zmienna X_j zależy od X_j „poprzez zmienną" ϑ.

W modelu bayesowskim dostateczność jest (prawie) równoważna warunkowej niezależności parametru i obserwacji pod warunkiem statystyki.

9.3.2 Stwierdzenie (Bayesowska dostateczność). *W modelu bayesowskim następujące warunki są równoważne:*

(i) zmienne losowe ϑ i X są warunkowo niezależne pod warunkiem $T(X)$,

(ii) zachodzi własność faktoryzacji $f_\theta(x) = g_\theta(T(x))h(x)$, z wyjątkiem być może zbioru parametrów θ o zerowym prawdopodobieństwie a priori[3]*.*

Szkic dowodu. Idea jest niezwykle prosta. Poniższe rozumowanie jest ścisłym dowodem w przypadku dyskretnych przestrzeni $\mathcal{X}$ i Θ, kiedy gęstości warunkowe są elementarnie zdefiniowanymi prawdopodobieństwami warunkowymi. W przypadku ogólniejszym są kłopoty techniczne związane z definicją rozkładów warunkowych, o czym już mówiliśmy w Uwadze 2.3.3. Stwierdzenie 2.3.2, przełożone na język bayesowski, mówi że dostateczność statystyki T jest równoważna warunkowi

$$f(x|t, \theta) = f(x|t),$$

[3] Zastrzeżenie o „wyjątkowym zbiorze parametrów" jest niestety konieczne, bo $f(x|\theta)$ jest określone tylko dla „prawie wszystkich" θ.

dla $T(x) = t$. To jest równoważne warunkowej niezależności, czyli $f(\theta, x|t) = f(\theta|t)f(x|t)$. Korzystając z faktu, że t jest wyznaczone przez x oraz z „symetrycznej" formy warunkowej niezależności, mamy kolejną równoważną tożsamość:

$$f(\theta|x) = f(\theta|x,t) = f(\theta|t).$$

□

Bayesowska interpretacja dostateczności jest bardzo intuicyjna: rozkład *a posteriori* (czyli wszystko co wie statystyk po zaobserwowaniu x) zależy tylko od $t = T(x)$.

9.4 Zadania

1. Rozpatrzmy sytuację opisaną w Przykładzie 9.1.1. Przypuśćmy, że kierowca zgłosił szkodę w 1-szym roku ubezpieczenia. Obliczyć prawdopodobieństwo, że zgłosi szkodę i w 2-gim roku. Zakładamy, że „typ" kierowcy się nie zmienia.

2. (Przykład Laplace'a). Mamy $r + 1$ urn. W każdej urnie znajduje się r kul, przy tym urna numer i zawiera i kul białych oraz $r - i$ kul czarnych. Losowanie przebiega w następujący spoób.

   - Wybieramy jedną z urn z jednakowym prawdopodobieństwem $\frac{1}{r+1}$.
   - Losujemy $n + 1$ razy ze zwracaniem z wybranej uprzednio urny.

   Wiadomo, że w n losowaniach wybraliśmy same kule białe. Obliczyć prawdopodobieństwo, że w kolejnym $(n + 1)$-szym losowaniu też wyjdzie biała?

3. Obserwacje $X_1, \ldots, X_n$ są próbką z rozkładu normalnego $\mathrm{N}(\mu, 1/\kappa)$, gdzie μ jest znane, zaś nieznany parametr κ (odwrotność wariancji) ma rozkład *a priori* $\mathrm{Gamma}(\alpha, \lambda)$. Obliczyć rozkład *a posteriori*.