Pojęcie statystyki

Pojęcie statystyki w statystyce matematycznej jest odpowiednikiem pojęcia zmiennej losowej w rachunku prawdopodobieństwa. Niech $\boldsymbol{X}=(X_1,...,X_n)$ będzie próbą z pewnej populacji.

Definicja. Wektorową funkcję mierzalną $\boldsymbol{T}$: $\mathcal{X} \to \boldsymbol{T}(\boldsymbol{X})=(T_1(\boldsymbol{X}),...,T_k(X)) \in R^k$ próby $\boldsymbol{X}$ nazywamy k wymiarową statystyką.

Należy koniecznie podkreślić, że **statystyka nie może zależeć od parametru θ** indeksującego rodzinę rozkładów.

Niech $\boldsymbol{X}=(X_1,...,X_n)$ będzie próbą prostą z populacji o rozkładzie $N(m,\sigma^2)$.Niech

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

Wówczas $\boldsymbol{X}=(X_1,...,X_n)$, $\overline{X}$, S^2 są statystykami. Natomiast $\dfrac{\overline{X}-m}{S}$ i $\dfrac{\overline{X}-m}{\sigma}$ nie są statystykami.

Dystrybuanta empiryczna i jej podstawowe własności

Rozważmy przestrzeń statystyczną $(R,\mathcal{B},\{P_F,F\in\mathcal{F}\})^n$. Stawiamy pytanie:

Czy mając do dyspozycji ciąg niezależnych obserwacji $x_1,...,x_n$ zmiennej losowej o rozkładzie z nieznaną dystrybuantą F można choćby w przybliżeniu odtworzyć tą dystrybuantę ?

Aby odpowiedzieć na to pytanie definiujemy na R funkcję $\hat{F}_n(t)$ zwaną dystrybuantą empiryczną

$$\hat{F}_n(t) = \frac{\text{liczba obserwacji nie większych niż } t}{\text{liczba wszystkich obserwacji}} = \frac{\#\{1\le j\le n : x_j \le t\}}{n}$$

Ponieważ obserwacja $(x_1,...,x_n)$ jest realizacją wektora losowego $(X_1(\omega),...,X_n(\omega))$, to dla każdego ustalonego t wartość dystrybuanty empirycznej $\hat{F}_n(t)$ traktujemy jako zaobserwowaną wartość zmiennej losowej $\hat{F}_n(t,\omega)$ zwanej również dystrybuantą empiryczną określonej wzorem

$$\hat{F}_n(t,\omega) = \frac{\#\{1\le j\le n : X_j(\omega) \le t\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,t]}(X_i(\omega))$$

gdzie $$\mathbf{1}_{(-\infty,t)}(x) = \begin{cases} 1, \text{ dla } x \in (-\infty,t] \\ 0, \text{ dla } x \notin (-\infty,t] \end{cases}.$$

Z powyższego wzoru widać, że dla ustalonego t dystrybuanta empiryczna jest sumą niezależnych zmiennych losowych $\mathbf{1}_{(-\infty,t]}(X_i(\omega))$ o rozkładzie dwupunktowym. Ogólnie, dystrybuanta empiryczna jest procesem stochastycznym na R. Z powyższych uwag wynika.

Twierdzenie. Dla każdego ustalonego t dystrybuanta empiryczna ma następujące własności:

1. $E_F(\hat{F}_n(t,\omega)) = F(t)$

2. $P_F\{\omega : \lim_{n\to\infty} \hat{F}_n(t,\omega) = F(t)\}=1$

3. Rozkład zmiennej losowej $\sqrt{n}\frac{\hat{F}_n(t,\omega)-F(t)}{\sqrt{F(t)(1-F(t))}}$ dąży do rozkładu $N(0,1)$, gdy $n\to\infty$.

Dowód: Własność 1 wynika z liniowości operatora wartości oczekiwanej. Istotnie

$$E_F(\hat{F}_n(t,\omega)) = E_F[\tfrac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,t]}(X_i(\omega))] = \tfrac{1}{n}\sum_{i=1}^{n} E_F[\mathbf{1}_{(-\infty,t]}(X_i(\omega))] =$$

$$\tfrac{1}{n}\sum_{i=1}^{n} P_F(X_i(\omega)\le t) = \tfrac{1}{n}\sum_{i=1}^{n} F(t) = F(t)$$

Własność 2 wynika bezpośrednio z mocnego prawa wielkich liczb Kołmogorowa a własność 3 z centralnego twierdzenia granicznego CTG dla ciągu prób Bernoulliego.

Własności 1, 2 i 3 mają charakter lokalny. Warto zaznaczyć, że zbiór $\{\omega: \lim_{n\to\infty} \hat{F}_n(t,\omega) \neq F(t)\}$ o zerowym P_F prawdopodobieństwie jest zależny od t. Istotnym wzmocnieniem własności 2 jest następujące twierdzenie Gliwienki-Cantellego zwane także **podstawowym twierdzeniem statystyki matematycznej**.

Twierdzenie Gliwienki-Cantellego. Jeżeli $X_1,\ldots, X_n$ jest ciągiem niezależnych zmiennych losowych o tym samym rozkładzie z dystrybuantą F, to

$$P_F\{\omega : \lim_{n\to\infty} D_n(\omega) = 0\}=1, \text{ gdzie } D_n(\omega) = \sup_{-\infty<t<\infty} |\hat{F}_n(t,\omega) - F(t)|.$$

Twierdzenie to orzeka, że dystrybuanta empiryczna zbiega z prawdopodobieństwem 1 jednostajnie na R do dystrybuanty teoretycznej. Zatem, jeżeli rozmiar próby jest dostatecznie duży, to dystrybuanta empiryczna dowolnie dobrze przybliża nieznaną dystrybuantę F.

Bardzo ważnym i użytecznym jest następujące:

Twierdzenie Kołmogorowa. Jeżeli $X_1,\ldots, X_n$ jest ciągiem niezależnych zmiennych losowych o tym samym rozkładzie z **ciągłą** dystrybuantą $F \in \mathcal{F}_c$, to

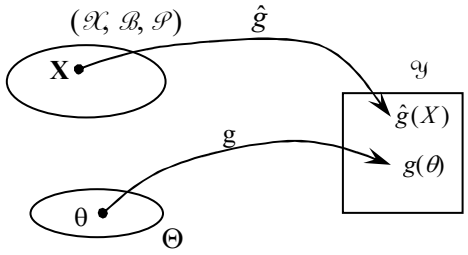
$$\lim_{n\to\infty} P_F(\sqrt{n}D_n \le x) = K(x) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2x^2} \text{ dla } x>0.$$

Dystrybuanta $K(x)$ jest stablicowana. Rozkład zmiennej losowej D_n przy założeniu $F \in \mathcal{F}_c$ nie zależy od dystrybuanty F. Wynika to z faktu, że jeśli X jest zmienną losową o dystrybuancie F, to $F(X)$ jest zmienną losową o rozkładzie jednostajnym $U_{(0,1)}$.

Podstawowe problemy statystyki matematycznej

Niech $(\mathcal{X}, \mathcal{B}, \mathcal{P}=\{P_\theta : \theta\in\Theta\})$, będzie przestrzenią statystyczną indukowaną przez zmienną losową X (zwykle wektorową) o wartościach w przestrzeni $\mathcal{X}$. Podstawowe problemy statystyki matematycznej to:

Problem estymacji punktowej. Na podstawie obserwacji zmiennej losowej X oszacować $g(\theta)\in \mathcal{Y}$, gdzie $g: \Theta\to \mathcal{Y}$ jest znaną funkcją parametru θ o wartościach w pewnej przestrzeni metrycznej (zwykle euklidesowej). Ponieważ parametr θ jest nieznany wartość $g(\theta)$ jest nieznana. Rozwiązaniem tego problemu będzie pewna funkcja (statystyka, element losowy) $\hat{g}:\mathcal{X}\to\mathcal{Y}$ zwana estymatorem. Estymator może być uznany za dobry estymator, jeżeli funkcja $\hat{g}$ przyjmuje wartości bliskie wartościom $g(\theta)$ $\forall\theta\in\Theta$. Wprowadzenie do sformułowania problemu funkcji g poszerza klasę jednolicie rozważanych zagadnień, gdyż oprócz szacowania samego parametru θ (g jest wtedy identycznością) i różnych jego funkcji (np. θ^2) obejmuje szacowanie wartości pewnych funkcjonałów w przypadkach nieparametrycznych – $(R,\mathcal{B},\{P_F,F\in\mathcal{F}^n$ np. $g(\theta)=g(F)=\int xdF = E_\theta X$.



Problem estymacji przedziałowej. Dla przedstawionego powyżej problemu estymacji możemy konstruować inne rozwiązanie w postaci zbioru $\hat{g}(X)\subset\mathcal{Y}$ np. przedziału $(\hat{g}_1(X), \hat{g}_2(X))\subset\mathcal{Y}$ w przypadku szacowania parametru liczbowego $g(\theta)$ tak, aby $\forall\theta\in\Theta$ $P_\theta(\hat{g}_1(X)\le g(\theta)\le\hat{g}_2(X))\ge 1-\alpha$. Przedział $(\hat{g}_1(X), \hat{g}_2(X))$ jest oczywiście przedziałem losowym. Liczbę $1-\alpha$ (zwykle bliską jedności) nazywamy poziomem ufności. Zagadnienie estymacji przedziałowej można uogólnić zastępując przedział innym **zbiorem ufności** (np. kulą jeśli $\mathcal{Y}$ jest przestrzenią metryczną). Zazwyczaj przyjmuje się pewne założenia regularności (np. spójność zbioru ufności) i założenia dotyczące kształtu zbioru ufności.

Problem testowania hipotez. Niech $\Theta=\Theta_0\cup\Theta_1$ i $\Theta_0\cap\Theta_1=\varnothing$. Na podstawie obserwacji zmiennej losowej X zweryfikować hipotezę H_0: $\theta\in\Theta_0$ wobec alternatywy H_1: $\theta\in\Theta_1$.

Rozwiązaniem problemu będzie pewna funkcja φ: $\mathcal{X}\to[0,1]$ zwana zrandomizowanym testem statystycznym. Dysponując testem statystycznym φ i obserwacją x zmiennej losowej X odrzucamy hipotezę H_0 z prawdopodobieństwem $\varphi(x)$. Aby podjąć konkretną decyzję należy więc użyć pewnego

mechanizmu losowego, który produkuje dwa wyniki z prawdopodobieństwami $\varphi(x)$ i $1-\varphi(x)$ i na tej podstawie podjąć (wylosować) decyzję. Jeżeli zbiór wartości funkcji φ jest zbiorem dwuelementowym $\{0,1\}$, to test φ nazywamy niezrandomizowanym. Test taki dzieli przestrzeń prób na dwa rozłączne zbiory :

$\varphi^{-1}(\{1\})=\{x\in\mathcal{X}: \varphi(x)=1\}$ zwany zbiorem odrzucenia hipotezy H_0

i

$\varphi^{-1}(\{0\})=\{x\in\mathcal{X}: \varphi(x)=0\}$ zwany zbiorem akceptacji H_0.

Konstrukcja testu niezrandomizownego jest więc równoważna rozbiciu przestrzeni prób na dwa rozłączne podzbiory.

Problem klasyfikacji (zwany również problemem dyskryminacji.) jest uogólnieniem problemu testowania hipotez. Uogólnienie to polega na rozbiciu zbioru parametrów Θ (lub równoważnie rodziny $\mathcal{P}$ rozkładów prawdopodobieństwa na przestrzeni prób) na $k \geq 2$ rozłącznych podzbiorów tzn. $\Theta=\bigcup_{i=1}^{k}\Theta_i$, $i\neq j \Rightarrow \Theta_i\cap\Theta_j = \varnothing$. Dla dowolnej obserwacji x zmiennej losowej X należy zadecydować z której z k rozłącznych populacji ona pochodzi. Rozwiązaniem tego problemu będzie pewna funkcja wektorowa $\varphi : \mathcal{X}\rightarrow\varphi(x)=(\varphi_1(x),..., \varphi_k(x))$ zwana zrandomizowaną funkcją dyskryminacyjną, gdzie $\varphi_i(x)$ jest prawdopodobieństwem zakwalifikowania obserwacji x do i-tej populacji i $\sum_{i=1}^{k}\varphi_i(x)=1 \quad (\forall x)$. Zagadnienie testowania hipotez jest szczególnym przypadkiem zagadnienia klasyfikacji dla k=2. Warunek $\varphi_1(x)+\varphi_2(x)=1$ umożliwia używanie tylko jednej składowej funkcji wektorowej (φ_1,φ_2).

Każdy z przedstawionych problemów ma swój specyficzny aspekt, ale można też wyróżnić pewne wspólne cechy pozwalające traktować jednolicie te problemy jako tzw. **statystyczny problemem decyzyjny**, czyli grę dwóch osób: statystyka i natury. To podejście będzie precyzyjnie omówione w kursie statystyki matematycznej

Uzupełnienia

Wybrane rozkłady prawdopodobieństwa użyteczne w statystyce

- Rozkład χ^2_n chi-kwadrat o n stopniach swobody - to rozkład sumy kwadratów n niezależnych zmiennych losowych o standaryzowanym rozkładzie normalnym $N(0,1)$ tzn.

  $X_1,...,X_n$ i.i.d. $N(0,1) \Rightarrow Y=\sum_{i=1}^{n} X_i^2$ ma rozkład χ^2_n o funkcji gęstości $f_Y(y)=\frac{1}{2^{n/2}\Gamma(n/2)} y^{n/2-1}e^{-y/2}$,

  $y>0$. Widać, że rozkład χ^2_n jest szczególnym przypadkiem rozkładu gamma.

  Uzasadnienie. Niech zmienna losowa X ma rozkład normalny $N(0,1)$. Wyznaczymy rozkład zmiennej losowej $Y=X^2$.

  Dla $y>0$; $F_Y(y)=P(Y<y)=P(X^2<y)=P(-\sqrt{y}<X<\sqrt{y})=F_{N(0,1)}(\sqrt{y})-F_{N(0,1)}(-\sqrt{y})$

  Stąd
  $$f_Y(y)=\frac{d}{dy}F_Y(y)=f_{N(0,1)}(\sqrt{y})\frac{1}{2\sqrt{y}}-f_{N(0,1)}(-\sqrt{y})(-\frac{1}{2\sqrt{y}})=$$
  $$=f_{N(0,1)}(\sqrt{y})\frac{1}{\sqrt{y}}=\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{y}}e^{-\frac{y}{2}}=\frac{(\frac{1}{2})^{\frac{1}{2}}}{\Gamma(\frac{1}{2})}y^{\frac{1}{2}-1}e^{-\frac{y}{2}}$$
  , czyli $Y\sim\mathcal{G}(\frac{1}{2},\frac{1}{2})$

  Z twierdzenia o dodawaniu dla rozkładu Gamma wynika, że $\chi^2_n=\mathcal{G}(\frac{1}{2},\frac{n}{2})$.

- Rozkład t-Studenta o n stopniach swobody – niech X będzie zmienną losową o rozkładzie $N(0,1)$ a Y zmienną losową o rozkładzie χ^2_n. Jeżeli zmienne X i Y są niezależne, to zmienna losowa $T=\frac{X}{\sqrt{\frac{Y}{n}}}$ ma rozkład t Studenta o n stopniach swobody i funkcji gęstości

  $f_T(t)=\frac{\Gamma(n+1/2)}{\sqrt{n\pi}\,\Gamma(n/2)}(1+\frac{t^2}{n})^{-\frac{n+1}{2}}$

Uzasadnienie. Rozważmy niezależne zmienne $X\sim N(0,1)$ i $Y\sim\chi^2_n$. Znajdziemy rozkład zmiennej losowej $T=\frac{X}{\sqrt{\frac{Y}{n}}}$.

$$F_T(t)=P(T<t)=P(\frac{X}{\sqrt{\frac{1}{n}Y}}<t)=P(X<t\sqrt{\tfrac{1}{n}Y})=\iint\limits_{x<t\sqrt{\frac{1}{n}y}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\frac{1}{2^{n/2}\Gamma(n/2)}y^{\frac{n}{2}-1}e^{-\frac{y}{2}}dxdy=$$

$$\left\{\begin{cases}u=y\\ v=\frac{x}{\sqrt{\frac{1}{n}y}}\end{cases}\Leftrightarrow\begin{cases}x=v\frac{\sqrt{u}}{\sqrt{n}}\\ y=u\end{cases},\left|\frac{\partial(x,y)}{\partial(u,v)}\right|=\begin{Vmatrix}\frac{v}{2\sqrt{u}\sqrt{n}} & \frac{\sqrt{u}}{\sqrt{n}}\\ 1 & 0\end{Vmatrix}=\left|-\frac{\sqrt{u}}{\sqrt{n}}\right|\right\}$$

$$=\int\limits_{-\infty}^{t}[\int\limits_{0}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{uv^2}{2n}}\frac{1}{2^{n/2}\Gamma(n/2)}u^{\frac{n}{2}-1}e^{-\frac{u}{2}}\frac{\sqrt{u}}{\sqrt{n}}du]dv=\int\limits_{-\infty}^{t}[\frac{1}{\sqrt{n}2^{\frac{n+1}{2}}\Gamma(1/2)\Gamma(n/2)}\int\limits_{0}^{\infty}u^{\frac{n+1}{2}-1}e^{-\frac{1}{2}(1+\frac{v^2}{n})u}du]dv=$$

$$=\int\limits_{-\infty}^{t}\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n}2^{\frac{n+1}{2}}\Gamma(1/2)\Gamma(n/2)(\frac{1}{2}(1+\frac{v^2}{n}))^{\frac{n+1}{2}}}[\int\limits_{0}^{\infty}\frac{(\frac{1}{2}(1+\frac{v^2}{n}))^{\frac{n+1}{2}}}{\Gamma(\frac{n+1}{2})}u^{\frac{n+1}{2}-1}e^{-\frac{1}{2}(1+\frac{v^2}{n})u}du]dv=\int\limits_{-\infty}^{t}\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n}2^{\frac{n+1}{2}}\Gamma(1/2)\Gamma(n/2)(\frac{1}{2}(1+\frac{v^2}{n}))^{\frac{n+1}{2}}}dv=$$

$$=\int\limits_{-\infty}^{t}\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n}\Gamma(1/2)\Gamma(n/2)(1+\frac{v^2}{n})^{\frac{n+1}{2}}}dv$$

Stąd otrzymujemy funkcję gęstości $f_T(t)=\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(n/2)(1+\frac{t^2}{n})^{\frac{n+1}{2}}}$.

Uwaga : $\int\limits_{0}^{\infty}\frac{(\frac{1}{2}(1+\frac{v^2}{n}))^{\frac{n+1}{2}}}{\Gamma(\frac{n+1}{2})}u^{\frac{n+1}{2}-1}e^{-\frac{1}{2}(1+\frac{v^2}{n})u}du=1$ - całka z funkcji gęstości rozkładu $\mathcal{G}\left(\frac{1}{2}(1+\frac{v^2}{n}),\frac{n+1}{2}\right)$

- Rozkład $F_{n,m}$ Snedecora-Fishera- niech X i Y będą niezależnymi zmiennymi losowymi o rozkładach χ^2_n χ^2_m wówczas zmienna losowa $Z=\frac{X/n}{Y/m}$ ma rozkład $F_{n,m}$ Snedecora-Fishera z funkcją gęstości $f_Z(z)=\frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}(\frac{m}{n})^{\frac{m}{2}}\frac{z^{n/2-1}}{(z+m/n)^{(n+m)/2}}$, $z>0$

Niecentralne rozkłady chi-kwadrat $\chi^2_{n,\delta}$, Studenta $t_{n,\delta}$ i Snedecora $F_{n_1,n_2,\delta}$

Niecentralny rozkład chi-kwadrat

Niech $X_1,...,X_n$ będą niezależnymi zmiennymi losowymi o rozkładach $N(m_i,1)$ $i=1,...,n$. Wówczas zmienna losowa $U=\sum_{i=1}^{n}X_i^2$ ma rozkład niecentralny chi-kwadrat o n stopniach swobody i parametrze niecentralności $\delta=\sqrt{\sum_{i=1}^{n}m_i^2}$ oznaczany $\chi^2_{n,\delta}$.

Niecentralny rozkład *t*- Studenta

Niech $X\sim N(\delta,1)$, $Y\sim\chi^2_n$, X i Y niezależne. Wówczas $T=\frac{X}{\sqrt{\frac{Y}{n}}}\sim t_{n,\delta}$

Niecentralny rozkład *F* Snedecora Fishera

Niech $X\sim\chi^2_{n_1,\delta}$ $Y\sim\chi^2_{n_2}$, X i Y niezależne. Wówczas $F=\frac{\frac{X}{n_1}}{\frac{Y}{n_2}}\sim F_{n_1,n_2,\delta}$

Uzupełnienie. (tw. Fishera) Niech $X_1,...,X_n$ będzie próbą prostą z rozkładu $N(m,\sigma^2)$.

Przekształćmy ortogonalnie wektor $\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \sim N\left(\begin{bmatrix} m \\ \vdots \\ m \end{bmatrix}, \sigma^2 I\right)$ na wektor $\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$ według wzoru

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \vdots & \cdots & \vdots \\ \cdot & \cdots & \cdot \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

, gdzie pozostałe wiersze macierzy przekształcenia są aby macierz była ortonormalna (można to zrobić traktując pierwszy wiersz jako wektor w R^n, uzupełnić ten wektor n-1 wektorami tak, aby uzyskać bazę w R^n i zastosować procedurę ortogonalizacji Grama –Schmidta).

Przekształcony ortogonalnie wektor ma rozkład $\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \sim N\left(\begin{bmatrix} \sqrt{n}\, m \\ \vdots \\ 0 \end{bmatrix}, \sigma^2 I\right)$, więc jego składowe są niezależne. Zauważmy, że $Y_1 = \sqrt{n}\,\overline{X}$, natomiast

$$\sum_{i=1}^{n}(X_i-\overline{X})^2 = \sum_{i=1}^{n}X_i^2 - n\overline{X}^2 = \sum_{i=1}^{n}Y_i^2 - Y_1^2 = \sum_{i=2}^{n}Y_i^2 .$$

Stąd przy powyższych założeniach otrzymujemy (Fisher)

- statystyki $\sqrt{n}\overline{X}$ i $\sum_{i=1}^{n}(X_i-\overline{X})^2$ są niezależne więc także
- statystyki $\overline{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$ i $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$ są niezależne.
- zmienna losowa $\frac{\sqrt{n}(\overline{X}-m)}{\sigma} = \frac{Y_1-\sqrt{n}m}{\sigma}$ ma rozkład N(0,1) natomiast $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 = \frac{1}{\sigma^2}\sum_{i=2}^{n}Y_i^2$ ma rozkład χ^2_{n-1}.
- $Q(X_1,...,X_n,m) = \frac{\sqrt{n}(\overline{X}-m)}{S} = \frac{\frac{\sqrt{n}(\overline{X}-m)}{\sigma}}{\frac{1}{\sigma}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}}$ ma rozkład t-Studenta z n-1 stopniami swobody, natomiast funkcja
- $Q(X_1,...,X_n,\sigma^2) = \frac{(n-1)S^2}{\sigma^2}$ ma rozkład χ^2_{n-1}.

Uzupełnienie. Wielowymiarowy rozkład normalny

Wielowymiarowy rozkład normalny może być zdefiniowany poprzez funkcję gęstości względem odpowiedniej wielowymiarowej miary Lebesgue'a , lecz sposób ten prowadzi często do uciążliwych rachunków. Wygodniej jest wprowadzić n wymiarowy rozkład normalny w oparciu o **twierdzenie Cramera-Wolda** głoszące, że rozkład n wymiarowej zmiennej losowej $\boldsymbol{X}$ jest całkowicie określony poprzez podanie jednowymiarowych rozkładów funkcji liniowych $\boldsymbol{l}^T\boldsymbol{X}$ dla wszystkich $\boldsymbol{l} \in R^n$ (symbol ^{T} oznacza transpozycję wektora traktowanego jako macierz kolumnową). To podejście może być także użyte do zdefiniowania rozkładu normalnego w nieskończenie wymiarowych przestrzeniach funkcyjnych Banacha czy Hilberta: wystarczy postulować aby dowolny funkcjonał liniowy miał jednowymiarowy rozkład normalny (Frechet 1951).

Niech $\boldsymbol{X}=\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$ będzie n wymiarową zmienną losową, $\boldsymbol{m}=E(\boldsymbol{X})$ wektorem wartości oczekiwanych a $\boldsymbol{V}=V(\mathbf{X})$ macierzą kowariancyjną.

Def. Zmienna losowa $\boldsymbol{X}$ ma n wymiarowy rozkład normalny $N_n(\boldsymbol{m},\boldsymbol{V})$ wtedy i tylko wtedy, gdy dla każdego wektora $\boldsymbol{l} \in R^n$ jednowymiarowa zmienna losowa $\boldsymbol{l}^T\boldsymbol{X}$ ma rozkład normalny $N(\boldsymbol{l}^T\boldsymbol{m}, \boldsymbol{l}^T\boldsymbol{V}\boldsymbol{l})$.

Z powyższej definicji łatwo wynikają następujące własności:

- jeżeli $\boldsymbol{A}$ jest macierzą typu (q,n) a n wymiarowa zmienna losowa $\boldsymbol{X}$ ma rozkład $N_n(\boldsymbol{m},\boldsymbol{V})$, to q wymiarowa zmienna losowa $\boldsymbol{Y}=\boldsymbol{A}\boldsymbol{X}$ ma rozkład normalny $N_q(\boldsymbol{A}\boldsymbol{m}, \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T)$
- rozważmy $p+q$ wymiarową zmienną losową $\boldsymbol{X}=\begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}$ przy czym p pierwszych składowych tworzy p wymiarową zmienną $\boldsymbol{X}_1$ a pozostałe q wymiarową zmienną losową $\boldsymbol{X}_2$. Dokonując odpowiedniego podziału na bloki możemy przedstawić wektor wartości oczekiwanych $\boldsymbol{m}$ i macierz kowariancyjną $\boldsymbol{V}$ w postaci blokowej $\boldsymbol{m}=\begin{bmatrix} \boldsymbol{m}_1 \\ \boldsymbol{m}_2 \end{bmatrix}$ $\boldsymbol{V}=\begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix}$. Jeżeli $\begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \sim N_{p+q}\left(\begin{bmatrix} \boldsymbol{m}_1 \\ \boldsymbol{m}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix}\right)$, to $\boldsymbol{X}_1 \sim N_p(\boldsymbol{m}_1,\boldsymbol{V}_{11})$, $\boldsymbol{X}_2 \sim N_q(\boldsymbol{m}_2,\boldsymbol{V}_{22})$

  $$\boldsymbol{X}_1 | \boldsymbol{X}_2 \sim N_p\left(\boldsymbol{m}_1 + \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{m}_2), \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}\right)$$

  $$\boldsymbol{X}_2 | \boldsymbol{X}_1 \sim N_q\left(\boldsymbol{m}_2 + \boldsymbol{V}_{21}\boldsymbol{V}_{11}^{-1}(\boldsymbol{X}_1 - \boldsymbol{m}_1), \boldsymbol{V}_{22} - \boldsymbol{V}_{21}\boldsymbol{V}_{11}^{-1}\boldsymbol{V}_{12}\right)$$

- jeżeli macierz kowariancyjna $\boldsymbol{V}$ n wymiarowej zmiennej losowej $\boldsymbol{X}$ o rozkładzie $N_n(\boldsymbol{m},\boldsymbol{V})$ jest nieosobliwa, to istnieje gęstość tego rozkładu względem n wymiarowej miary Lebesgue'a i wyraża się wzorem $f_n(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{V}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{m})}$