Zpracování výběrů asymetrického rozdělení biochemických dat

Milan Meloun
Katedra analytické chemie, Univerzita Pardubice, 532 10 Pardubice
milan.meloun@upce.cz

Jiří Militký
Katedra textilních materiálů, Technická univerzita Liberec, 461 17 Liberec
jiri.militky@vslib.cz

Martin Hill,
Endokrinologický ústav, Národní 8, CZ-116 94 Praha 1,
mhill@endo.cz
a
Karel Kupka
Trilobyte Statistical Software Ltd., 530 02 Pardubice,
email: kupka@trilobyte.cz

Souhrn: *Průzkumová analýza dat provádí první kontakt s biochemickými daty a slouží k odhalení všech statistických zvláštností výběru, asymetrie rozdělení výběru a vybočujících hodnot. Když data nesplňují požadavky, kladené na výběr, nevykazují Gaussovo rozdělení a navíc obsahují vybočující hodnoty je uživatel vystaven problému jak vyčíslit odhad střední hodnoty. Datové výběry studovaných steroidů se vyznačují silně sešikmeným, asymetrickým rozdělením. Srovnáním středních hodnot dehydroepiandrosteronu DHEA, dehydroepiandrosteron sulfátu DHEAS, androstendionu ANDION, testosteronu TESTO, sexuálního hormonu vaziciglobulinu SHBG a konečně logaritmu indexu volného testosteronu (IFT=100•T/SHBG) u skupiny žen s mírnějším a výraznějším stupněm akné lze vyšetřit, zda existuje vztah mezi studovanými steroidy, SHBG a stupněm akné. Mocninná a Box-Coxova transformace slouží k nalezení objektivního odhadu střední hodnoty. Zaručuje uživateli spolehlivý odhad střední hodnoty i v případě silně sešikmeného rozdělení. Navržená metoda s doprovodným softwarem je dokumentována na úlohách vyčíslení bodového a intervalového odhadu střední hodnoty u vybraných steroidů.*

ÚVOD

Pokud v analýze jednorozměrných dat nemá být statistická analýza pouhým numerickým počítáním bez hlubšího smyslu, je třeba, aby byly ověřeny všechny předpoklady, které vedly k návrhu postupu analýzy [1, 2]. Je nezbytné vyšetřit platnost základních předpokladů, tj. nezávislosti, homogenity a normality prvků výběru. *Reprezentativní náhodný výběr* je popsán základními vlastnostmi: prvky výběru x_i jsou vzájemně nezávislé o dostatečné četnosti, výběr je homogenní a pochází z normálního rozdělení pravděpodobnosti, všechny prvky souboru mají stejnou pravděpodobnost, že budou zařazeny do výběru. Je třeba mít na paměti, že malé porušení předpokladu normality nemusí být katastrofické s ohledem na výsledek statistické analýzy. Na druhé straně je však špatné, když odhady i testy závisejí spíše na jiných faktorech než je chování většiny dat, na velikosti výběru, na uspořádání výsledků nesledovaných proměnných, atd.

Asymetrie rozdělení je častým jevem při měření kvantit blízkých mezi detekce přístroje, některých velmi malých veličin (stopové koncentrace, znečištění, velikosti malých částic). Při vyhodnocení nelze v takovém případě použít postupů založených na normálním rozdělení, jako aritmetický průměr, pravidlo 3 sigma nebo Shewhartovy regulační diagramy [1-4].

Poměrně jednoduchá technika nelineární transformace umožní i pro asymetricky rozdělená data užití klasických metod. Pokud data nesplňují předpoklad normality, je v řadě případů možné zlepšit jejich rozdělení vhodnou transformací. Cílem transformace je nalézt funkci $y = f(x)$ původních hodnot x, která zajistí minimální šikmost, případně maximální věrohodnost transformovaných dat vzhledem k normálnímu rozdělení.

Dehydroepiandrosteron (DHEA) je v endokrinologii běžně stanovovaný steroid. V metabolické řadě je nepřímým prekurzorem pohlavních hormonů. Srovnáním středních hodnot dehydroepiandrosteronu DHEA, dehydroepiandrosteron-sulfátu DHEAS, androstendionu ANDION, testosteronu TESTO, sexuálního hormonu vaziciglobulinu SHBG a konečně logaritmu indexu volného testosteronu ($IFT = 100 \cdot T/SHBG$) u skupiny žen s mírnějším a výraznějším stupněm akné lze zjistit, zda existuje vztah mezi studovanými steroidy, SHBG a stupněm akné. Zhodnocení rozdílů mezi dotyčnými středními hodnotami bylo jedním s kroků při hledání vztahů mezi androgeny a stupněm akné.

METODICKÁ ČÁST

1. Postup analýzy biochemických dat

Experimentální data se v analytické laboratoři často vyznačují asymetrickým rozdělením a porušením předpokladů, kladených na výběr. Při rutinním zpracování experimentálních dat se obyčejně provádí:

A. Popisná analýza, tj. nalezení odhadu parametrů polohy, rozptýlení a tvaru,

B. Určení intervalů spolehlivosti těchto odhadů,

C. Testování významnosti těchto parametrů.

V popisné statistické analýze dat se obyčejně užívá dvou technik:

(a) *Klasické techniky*, kdy se počítá aritmetický průměr, rozptyl, šikmost a špičatost.

(b) *Exploratorní techniky*, kdy se používá robustních či kvantilových charakteristik polohy (mediánu), rozptýlení, šikomosti a špičatosti.

Při *klasické popisné analýze dat* se předpokládá splnění *základních předpokladů*, kladených na výběr, jako jsou nezávislost prvků, homogenita výběru, dostatečný rozsah výběru a rozdělení výběru. Jsou-li tyto předpoklady splněny, následuje vyčíslení odhadů polohy a rozptýlení, tj. obyčejně aritmetického průměru a rozptylu. Dále se vyčíslí intervaly spolehlivosti následované testováním statistických hypotéz. V pesimistickém případě následuje pokus o úpravu dat.

V *průzkumové analýze dat EDA* se vyšetřují *statistické zvláštnosti*, jako je lokální koncentrace dat, tvarové zvláštnosti rozdělení dat a přítomnost podezřelých hodnot. Odhalí se také anomálie a odchylky rozdělení výběru od typického rozdělení, obyčejně Gaussova. Interaktivní statistická analýza na počítači tento postup ulehčuje, většina statistického software totiž nabízí řadu diagnostických grafů a diagramů. Pokud je rozdělení dat nevhodné pro standardní statistickou analýzu (tj. většinou asymetrické), provádí se často vhodná transformační úprava dat. Pokud bylo indikováno sešikmené rozdělení nebo rozdělení s dlouhými konci, vede ke zlepšení mocninná a Boxova-Coxova transformace. Transformace je vhodná především při asymetrii rozdělení původních dat, resp. nekonstantnosti rozptylu.

V *konfirmatorní analýze CDA* je nabízena paleta rozličných odhadů polohy, rozptýlení a tvaru. Základní jsou *klasické odhady* a *robustní odhady* (necitlivé na odlehlé prvky výběru, resp. další předpoklady o datech) nebo *adaptivní odhady*. Z nabídky odhadů parametrů vybírá uživatel ty, jež odpovídají závěrům průzkumové analýzy dat a ověření předpokladů o výběru.

A. Průzkumová (exploratorní) analýza experimentálních dat (EDA)

Odhalení stupně symetrie a špičatosti výběrového rozdělení;

Indikace lokální koncentrace dat;

Nalezení vybočujících a podezřelých prvků ve výběru;

Porovnání výběrového rozdělení dat s typickými rozděleními;

Mocninná transformace dat;

Box-Coxova transformace dat.

B. Ověření předpokladů o datech:

Ověření nezávislosti prvků dat;

Ověření homogenity rozdělení dat;

Určení minimálního rozsahu dat.

Ověření normality rozdělení dat.

C. Konfirmatorní analýza dat (CDA)

Odhady parametrů (polohy, rozptýlení a tvaru):

1. Klasické odhady (bodové a intervalové) parametrů;
2. Robustní odhady (bodové a intervalové) parametrů;

2. Transformace experimentálních dat

Pro statistickou analýzu experimentálních dat je ideální, pokud jsou prvky výběru *náhodné* vzájemně *nezávislé* veličiny se stejným *normálním* rozdělením. Reálné výběry se od tohoto stavu více či méně odlišují a vzniká problém jak potom data vůbec vyhodnotit. V jednodušším případě má rozdělení *delší konce* (vyšší špičatost) než odpovídá normálnímu rozdělení. Běžné statistické testy předpokládají symetrické rozdělení dat a jsou vůči vyšší špičatosti dat poměrně necitlivé. Robustní metody odhadu parametrů polohy a rozptýlení zde nefungují dobře, protože opět předpokládají, že symetricky rozdělená data obsahují kontaminaci jistým podílem vybočujících hodnot.

Komplikovanější je případ, kdy je rozdělení výběru *sešikmené* (obyčejně k vyšším hodnotám). Módus pak již není totožný s mediánem ani střední hodnotou a vlastní interpretace parametru polohy je ztížena. Efektivní odhad parametru polohy je možný jen při znalosti rozdělení pravděpodobnosti. Běžné statistické testy jsou vůči sešikmenému rozdělení dat nerobustní. Také základní robustní metody odhadu polohy a rozptýlení zde nefungují dobře. Je zřejmé, že již symetrizační transformace bude v analýze takových dat velmi užitečná. Část lze nalézt *vhodnou transformaci*, která vede ke stabilizaci rozptylu, zesymetričtění rozdělení a někdy i k normalitě. Vychází se z představy, že zpracovávaná data jsou nelineární transformací normálně rozdělené náhodné veličiny x a hledá se k nim inverzní transformace $g(x)$.

(a) Transformace stabilizující rozptyl: nekonstantnost rozptylu je původním jevem u řady měření v instrumentálních metodách. Indikuje buď neplatnost aditivního modelu měření $x_i = \mu + \varepsilon_i$, kde ε_i jsou náhodné chyby s nulovou střední hodnotou a konstantním rozptylem, nebo indikuje nenormalitu rozdělení výběru. Stabilizace rozptylu vyžaduje nalezení transformace $y = g(x)$, ve které je již rozptyl $\sigma^2(y)$ konstantní. Pokud je rozptyl původní proměnné x funkcí typu $\sigma^2(x) = f_1(x)$, lze rozptyl $\sigma^2(y)$ určit z Taylorova rozvoje funkce $g(x)$

$$\sigma^2(y) \approx \left[\frac{dg(x)}{dx}\right]^2 f_1(x) = C$$

kde C je konstanta. Hledaná transformace $g(x)$ je pak řešením diferenciální rovnice

$$g(x) \approx C \int \frac{dx}{\sqrt{f_1(x)}}$$

U řady instrumentálních metod je zajištěna konstantnost relativní chyby měření $\delta(x)$. To znamená, že rozptyl $\sigma^2(x)$ je dán funkcí $f_1(x) = \delta^2(x)\, x^2 = \text{konst}\, x^2$. Po dosazení vyjde $g(x) = \ln x$. Optimální je pro tento případ logaritmická transformace původních dat. Z toho vyplývá také vhodnost použití geometrického průměru. Pokud je závislost $\sigma^2(x) = f_1(x)$ mocninná, bude optimální transformace $g(x)$ také mocninná. Jelikož pro normální rozdělení je střední hodnota na rozptylu nezávislá, bude transformace stabilizující rozptyl také zajišťovat přiblížení k normalitě.

(b) Symetrizující tranformace: zesymetričtění rozdělení výběru se provede *jednoduchou mocninnou transformací*

$$y = g(x) = \begin{cases} x^{\lambda} & \lambda > 0 \\ \ln x \quad \text{pro} & \lambda = 0 \\ -x^{-\lambda} & \lambda < 0 \end{cases}$$

Tato transformace však nezachovává měřítko, není vzhledem k hodnotě λ všude spojitá, zachovává však pořadí dat ve výběru a hodí se pouze pro kladná data. Optimální odhad $\hat{\lambda}$ se hledá s ohledem na minimalizaci vhodných charakteristik asymetrie. Kromě šikmosti $g_1(y)$ je možné užít i robustní verzi šikmosti definovanou výrazem

$$\hat{g}_{1R}(y) = \frac{(\tilde{y}_{0.75} - \tilde{y}_{0.50}) - (\tilde{y}_{0.50} - \tilde{y}_{0.25})}{\tilde{y}_{0.75} - \tilde{y}_{0.25}},$$

kde y_P je P%ní kvantil transformovaného výběru. Stejně jednoduché je sledovat rozdíl mezi střední hodnotou $\bar{y}$ a mediánem $\tilde{y}_{0.5}$ pomocí statistiky šikmosti

$$g_P = \frac{\bar{y} - \tilde{y}_{0.5}}{\sqrt{\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n - 1}}}$$

Pro symetrická rozdělení je statistika $\hat{g}_p(y)$ rovna nule. Stejně tak jsou rovny nule i statistiky $\hat{g}_1(y)$ a $\hat{g}_{1R}(y)$. Hodnotu $\hat{\lambda}$ lze hledat pomocí rankitového grafu, kde pro optimální $\hat{\lambda}$ budou kvantily $y_{(i)}$ ležet přibližně na přímce.

Hines - Hinesův selekční graf (osa x: $\bar{x}_{0.5}/x_{1-P_i}$, osa y: $\tilde{x}_{Pi}/\tilde{x}_{0.5}$): diagnostickou pomůckou pro odhad optimálního parametru λ je selekční graf dle Hinese a Hinesové, obr. 1. Vychází z požadavků symetrie jednotlivých kvantilů kolem mediánu

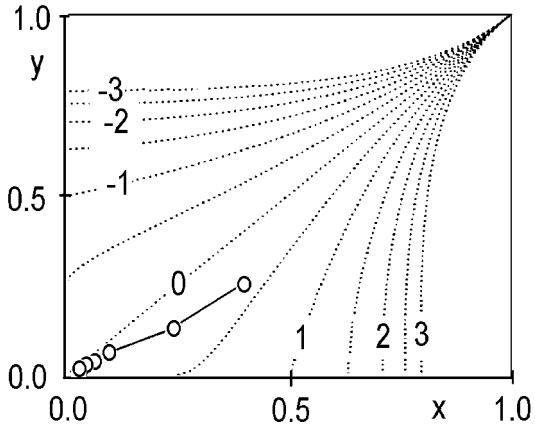
$$\left(\frac{\tilde{x}_{P_i}}{\tilde{x}_{0.5}}\right)^{\lambda} + \left(\frac{\tilde{x}_{0.5}}{\tilde{x}_{1-P_i}}\right)^{-\lambda} = 2$$

kde jako pořadové pravděpodobnosti jsou obvykle voleny hodnoty, $P_i = 2^{-i}$, $i = 2, 3$. K porovnání průběhu experimentálního bodů s ideálním (teoretickým) pro zvolené λ se do grafu zakreslují i řešení rovnice $y^{\lambda} + x^{-\lambda} = 2$ pro $0 \le x \le 1$ a $0 \le y \le 1$:

a) pro $\lambda = 0$ je řešením přímka $y = x$,

b) pro $\lambda < 0$ je řešením vztah $y = (2 - x^{-\lambda})^{1/\lambda}$,

c) pro $\lambda > 0$ je řešením vztah $x = (2 - y^{\lambda})^{-1/\lambda}$.



Obr. 1 Ukázka selekčního grafu pro výběr, vykazující téměř lognormální rozdělení

Podle umístění experimentálních bodů na teoretických křivkách selekčního grafu lze odhadovat velikost λ a posuzovat kvalitu transformace v různých vzdálenostech od mediánu.

(c) Normalizační transformace: pro přiblížení rozdělení výběru k rozdělení normálnímu vzhledem k šikmosti a špičatosti se užívá rodiny *Boxovy-Coxovy transformace*

$$y = g(x) = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & (\lambda \neq 0) \\ \ln x & (\lambda = 0) \end{cases}$$

Boxova-Coxova transformace má tyto vlastnosti:

1. Transformace *g(x)* jsou vzhledem k veličině λ spojité, protože v okolí nuly platí

$$\lim_{\lambda \to 0} \frac{x^{\lambda} - 1}{\lambda} = \lim_{\lambda \to 0} x^{\lambda} \,.\, \ln x = \ln x$$

2. Všechny transformace procházejí bodem [$y = 0$; $x = 1$] a mají v tomto bodě společnou směrnici, jsou zde co do průběhu totožné.

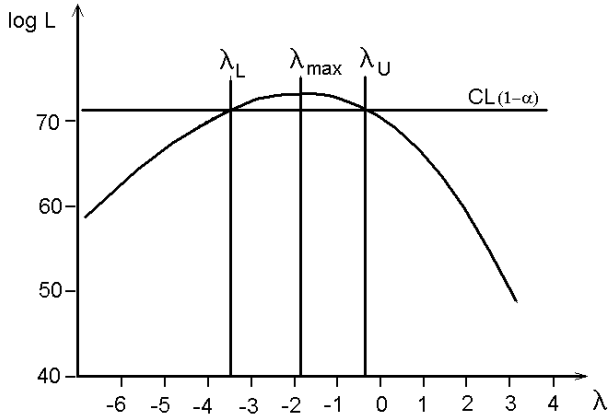
3. Mocninné transformace s exponenty -2; -3/2; -1; -1/2; 0; 1/2; 1; 3/2; 2 jsou co do křivosti rovnoměrně rozmístěné.

Boxova-Coxova transformace je použitelná pouze pro kladná data. Rozšíření této transformace na oblast, kdy rozdělení dat začíná od prahové hodnoty x_0, spočívá v náhradě x rozdílem $(x - x_0)$, který je vždy kladný.

Graf logaritmu věrohodnostní funkce (osa x: λ, osa y: ln L): pro odhad parametru λ v Boxově-Coxově transformaci lze užít metodu maximální věrohodnosti s tím, že pro $\lambda = \hat{\lambda}$ je rozdělení transformované veličiny y normální, $N(\mu_y, \sigma^2(y))$. Po úpravách bude logaritmus věrohodnostní funkce ve tvaru

$$\ln L(\lambda) = -\frac{n}{2} \ln s^2(y) + (\lambda - 1) \sum_{i=1}^{n} \ln x_i$$

kde $s^2(y)$ je výběrový rozptyl transformovaných dat y. Průběh věrohodnostní funkce ln L = $f(\lambda)$ lze znázornit ve zvoleném intervalu např. $-3 \le \lambda \le 3$ a identifikovat i maximum $\hat{\lambda}$, obr. 2.



Obr. 2 Graf logaritmu věrohodnostní funkce pro výběr z lognormálního rozdělení

Pro asymptotický 100(1 - α)%ní interval spolehlivosti parametru λ platí

$$2\left[\ln L(\hat{\lambda}) - \ln L(\lambda)\right] \le \chi^2_{1-\alpha}(1)$$

kde $\chi^2_{1-\alpha}(1)$ je kvantil χ^2-rozdělení s jedním stupněm volnosti. V tomto intervalu spolehlivosti leží všechna λ, pro která je ln $L(\lambda)$ větší nebo roven $\ln L(\hat{\lambda}) - 0.5\chi^2_{1-\alpha}(1)$. Výhodně lze do grafu logaritmu věrohodnostní funkce
ln $L(\lambda)$ na λ zakreslit obyčejně 95% interval spolehlivosti. Z tohoto grafu lze snadno odhadnout jak kvalitu transformace, odhad exponentu $\hat{\lambda}$, tak i posoudit, v jakých mezích se může hodnota λ pohybovat. Platí totiž, že čím je interval spolehlivosti exponentu $\bar{\lambda}$ tj. $\langle L_D, L_H \rangle$ širší, tím je transformace méně výhodná. Pokus tento interval obsahuje i hodnotu $\lambda = 1$, není transformace ze statistického hlediska přínosem.

3. Zpětná transformace

Pokud se podaří nalézt vhodnou transformaci, která vede k přibližné normalitě, lze určit $\bar{y}$, $s^2(y)$, interval spolehlivosti $\bar{y} \pm t_{1-\alpha/2}(n-1).\, s(y)/\sqrt{n}$ a provádět i statistické testování. Problém však spočívá v tom, že všechny statistické charakteristiky a jejich intervaly spolehlivosti je třeba určit pro původní proměnné.

1. **Nekorektní (naivní) přístup** spočívá v pouhé zpětné transformaci $\bar{x}_R = g^{-1}(\bar{y})$. Pro jednoduchou mocninnou transformaci vede zpětná transformace na obecný průměr definovaný vztahem

$$\bar{x}_R = \bar{x}_\lambda = \left[\frac{\sum_{i=1}^{n} x_i^\lambda}{n}\right]^{1/\lambda}$$

Pro $\lambda = 0$ se místo x^{λ} používá ln x a místo $x^{1/\lambda}$ pak e^{x}. Hodnota $x_R = \bar{x}_{-1}$ představuje *harmonický průměr*, $\bar{x}_R = \bar{x}_0$ *geometrický průměr*, $\bar{x}_R = \bar{x}_1$ *aritmetický průměr* a $\bar{x}_R = \bar{x}_2$ *kvadratický průměr*. Tento způsob zpětné transformace nebere v úvahu variabilitu střední hodnoty.

2. **Správnější přístup** zpětné transformace vychází z Taylorova rozvoje funkce $y = g(x)$ v okolí $\bar{y}$. Pro retransformovaný průměr $\bar{x}_R$ lze pak odvodit přibližný vztah

$$\bar{x}_R \approx g^{-1}\left[\bar{y} - \frac{1}{2}\frac{d^2 g(x)}{dx^2}\left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)\right]$$

Pro rozptyl vyjde $$s^2(x_R) \approx \left(\frac{dg(x)}{dx}\right)^{-2} s^2(y) \ .$$

Zde jednotlivé derivace jsou vyčísleny v bodě $x = \bar{x}_R$. Pro 100(1 - α)%ní interval spolehlivosti střední hodnoty původního souboru dat x platí

$$I_D \leq \mu \leq I_H$$

kde

$$I_D = g^{-1}\left(\bar{y} + G - t_{1-\alpha/2}(n-1)\frac{s(y)}{\sqrt{n}}\right)$$

$$I_H = g^{-1}\left(\bar{y} + G + t_{1-\alpha/2}(n-1)\frac{s(y)}{\sqrt{n}}\right)$$

$$G = -0.5\frac{d^2 g(x)}{dx^2}\left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)$$

Symbolem $t_{1-\alpha/2}(n - 1)$ je označen 100(1 - α/2)%ní kvantil Studentova rozdělení s (n - 1) stupni volnosti. Při znalosti hodnot konkrétní transformace $y = g(x)$ a odhadů $\bar{y}$, $s^2(y)$ je snadné vyčíslit hodnoty $\bar{x}_R$ a $s^2(x_R)$:

a) Pro speciální případ $\lambda = 0$, tzn. logaritmickou transformaci typu $g(x) = \ln x$, bude $\bar{x}_R \approx \exp\left[\bar{y} + 0.5\, s^2(y)\right]$. Rozptyl se určí vztahem $s^2(x_R) \approx \bar{x}_R^2\, s^2(y)$.

b) Pro případ $\lambda \neq 0$ a Boxovy-Coxovy transformace bude $\bar{x}_R$ jedním z kořenů kvadratické rovnice, pro které platí

$$\bar{x}_{R,1,2} = \left[0.5(1 + \lambda\bar{y}) \pm 0.5\sqrt{1 + 2\lambda(\bar{y} + s^2(y)) + \lambda^2(\bar{y}^2 - 2\, s^2(y))}\right]^{1/\lambda}$$

Jako odhad x_R se pak bere kořen $\bar{x}_{R,i}$, který je nejbližší mediánu $\tilde{x}_{0.5} = g^{-1}(\tilde{y}_{0.5})$. Při znalosti retransformovaného průměru $\bar{x}_R$ lze z vyčíslit i odpovídající rozptyl

$$s^2(x) = \bar{x}_R^{-2\lambda+2}\, s^2(y) \ .$$

EXPERIMENTÁLNÍ ČÁST

Dehydroepiandrosteron (DHEA) je běžně stanovovaný steroid. Jeho koncentrace se obvykle udávají v nmol/l. V metabolické řadě je nepřímým prekurzorem pohlavních hormonů. Je produkován nadledvinami (převážně v *zona reticularis*) a mužskými i ženskými gonádami. Jeho nadpodukce u žen jej jedním z markerů hyperandrogenismu. Zvýšené hladiny androgenů u žen s akné byly potvrzeny v mnoha studiích [5,7,8,9,10,12,14,15].

Terapie antiandrogeny navíc bývá u aknózních žen velmi účinná [5,9]. Je třeba vyšetřit, zda existuje jednoduchý vztah mezi stupněm akné a hladinou androgenů. Odpověď byla hledána porovnáním hladin steroidů u dvou výběrů žen s různým stupněm akné. Srovnáním středních hodnot DHEA (nmol/l), DHEAS (µmol/l), ANDION (nmol/l), TESTO (nmol/l) SHBG (nmol/l) a IFT=*100•T/SHBG* u skupiny žen s mírnějším stupněm akné a u skupiny žen s výraznějším stupněm akné lze snadno zjistit, zda existuje nějaký vztah mezi studovanými steroidy, SHBG a stupněm akné, nebo zda je stupeň akné ovlivňován spíše jinými faktory. Zhodnocení rozdílů mezi skupinovými středními hodnotami u skupin s různým stupněn akné bylo jedním s kroků při zkoumání vztahů mezi androgeny a stupněm akné.

VÝSLEDKY A JEJICH DISKUSE

U steroidů dehydroepiandrosteronu DHEA (nmol/l), dehydroepiandrosteron- sulfátu DHEAS (µmmol/l), androstendionu ANDION (nmol/l), testosteronu TESTO (nmol/l), sexuálního hormonu vaziciglobulinu SHBG (nmol/l) a konečně logaritmu indexu volného testosteronu IFT u skupiny žen s mírnějším stupněm akné (index 0) a u skupiny žen s výraznějším stupněm akné (index 1) bylo třeba nalézt spolehlivou střední hodnotu obsahu. Ze statistického hlediska to znamená předem vyšetřit rozdělení každého výběru, určit počet odlehlých hodnot ve výběru. Ke statistickému vyhodnocení dat je třeba užít průzkumovou analýzu dat, ověření předpokladů o náhodném výběru, a případně i transformaci dat. Na výběru jednoho steroidu DHEA bude ukázán celý postup statistického zpracování experimentálních dat.

1. Analýza výběru dehydroepiandrosteronu (DHEA)

Na příkladu dehydroepiandrosteronu (DHEA) je ukázána cela metodologie statistického zpracování biochemických dat. DHEA [nmol/l], $n = 43$:

5.0	8.66	7.67	10.7	6.4	5.3	3.0	3.7	7.6	47.3	4.0
16.8	10.8	12.4	4.5	6.4	7.4	12.2	5.5	6.5	6.7	6.0
5.3	6.7	3.9	6.1	2.9	7.1	7.8	7.1	3.1	8.5	13.1
9.2	3.3	3.3	4.5	6.9	9.8	9.0	8.69	7.2	5.4	

(a) Přehled popisných statistik: software NCSS2000 vyčíslil parametry polohy, rozptýlení a tvaru, z nichž nejdůležitější jsou uvedeny:

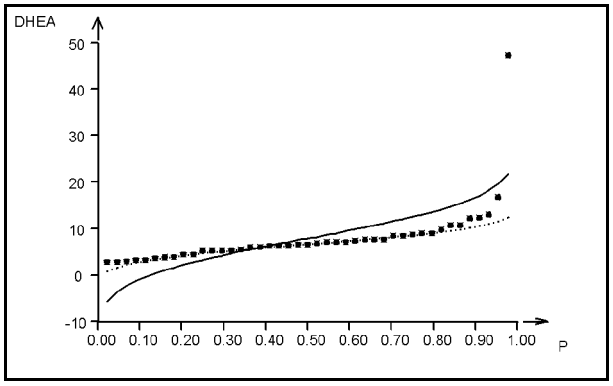
Tabulka 1. Přehled odhadů parametrů polohy a rozptýlení (NCSS2000 a ADSTAT):

Střední hodnota	Bodový odhad	Dolní mez	Horní mez	Užito
Aritmetický průměr	7.99	5.89	10.10	43
Geometrický průměr	6.77	-	-	43
Harmonický průměr	6.04	-	-	43
Medián	6.70	5.84	7.56	43
Módus	5.30	-	-	43
5%ní uřezaný průměr	7.01	6.02	8.00	39
10%ní uřezaný průměr	6.89	5.85	7.94	34
40%ní uřezaný průměr	6.78	6.11	7.44	6
M-odhad	6.77	5.83	7.70	43
Hoggův *M*-odhad	6.77	6.04	7.50	43
Směrodatná odchylka	6.83	-	-	43
Rozpětí	44.4	-	-	43
Interkvartilové rozpětí	3.39	-	-	43

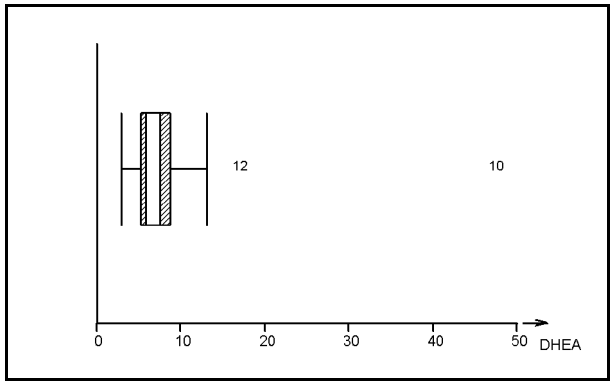
Z těchto vyčíslených odhadů si má uživatel vybrat správný. Pro n = 43 bylo vyčísleno minimum 2.90 a maximum 47.3. Z parametrů polohy pak aritmetický průměr $\bar{x}$ = 7.99 s 95%ním intervalovým odhadem L_L = 5.89 a L_U = 10.10, medián $\hat{x}_{0.5}$ = 6.70 s 95%ním intervalovým odhadem L_L = 5.84 a L_U = 7.56. Dále je geometrický průměr $\bar{x}_g$ = 6.77, harmonický průměr $\bar{x}_h$ = 6.04, modus x_M = 5.30, a následující uřezané průměry $\bar{x}(5\%)$ = 7.01 s $s(5\%)$ = 2.52 a pro $n(5\%)$ = 39, $\bar{x}(10\%)$ = 6.89 s $s(10\%)$ = 2.05 a pro $n(10\%)$ = 34, $\bar{x}(40\%)$ = 6.78 s $s(40\%)$ = 0.21 (1.36) pro $n(10\%)$ = 6. Robustní *M*-odhad polohy je $\hat{\mu}_M$ = 6.77 a rozptýlení σ_M = 2.86 s intervalovým odhadem L_L = 5.83 a L_U = 7.70 a dále robustní Hoogův *M*-odhad polohy je $\hat{\mu}_M$ = 6.77 a rozptýlení σ_M = 2.38 s intervalovým odhadem L_L = 6.04 a L_U = 7.51. Z parametrů rozptýlení jsou to směrodatná odchylka s = 6.83, rozpětí R = 44.4, interkvartilové rozpětí R_F = 3.39 a z parametrů tvaru je to šikmost g_1 = 4.59 (test ukazuje, že odchylka od 0 je statisticky významná a jde o nenormální rozdělení) a špičatost g_2 = 26.87 (test ukazuje, že odchylka od 3 je statisticky významná a jde o nenormální rozdělení).

(b) Základní diagnostické grafy EDA jsou užity ke grafickému znázornění datového výběru: *kvantilový graf* (obr. 3) vykazuje odlehlé hodnoty a asymetrické rozdělení, klasická a empirická křivka se totiž od sebe výrazně liší. *Krabicový graf* (obr. 4) indikuje asymetrické rozdělení a odlehlé hodnoty v horní části. *Graf polosum* (obr. 5) a *graf symetrie* (obr. 6) vykazují asymetrické rozdělení, protože značné množství bodů leží vně konfidenčního intervalu mediánové přímky. *Graf rozptýlení s kvantily* (obr. 7) ukazuje na

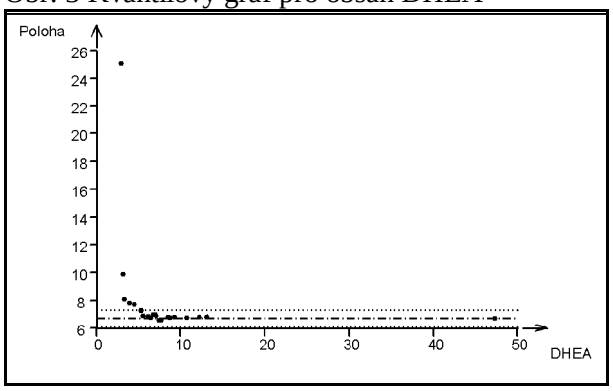
řadu odlehlých bodů, které leží vně sedecilového obdélníku. Poloha mediánu *M* je vyznačena krátkou mediánovou úsečkou ve střední části kvartilového grafu pro $P_1 = 0.5$. V *kruhovém grafu* (obr. 8) se liší obě kruhové křivky, teoretická elipsa pro normální rozdělení a empirická zborcená elipsa pro výběrové rozdělení.



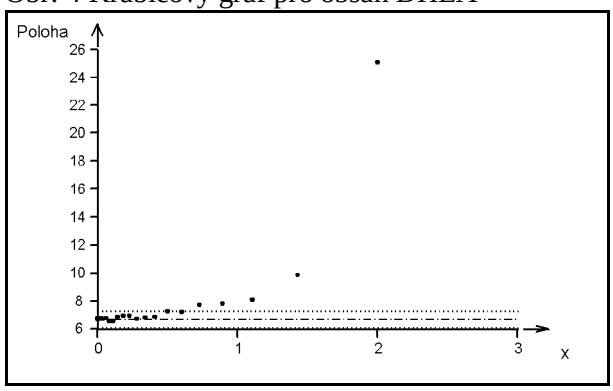
Obr. 3 Kvantilový graf pro obsah DHEA



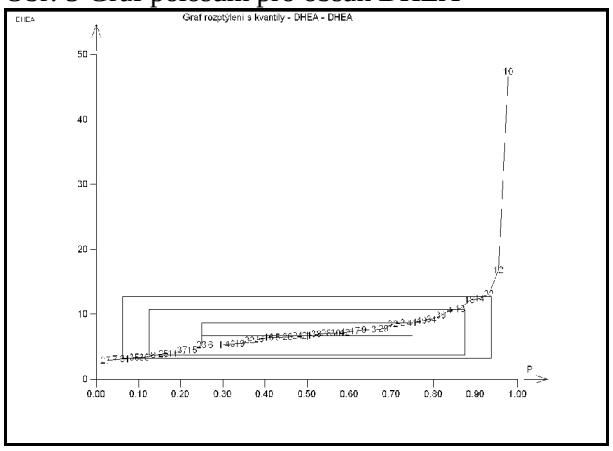
Obr. 4 Krabicový graf pro obsah DHEA



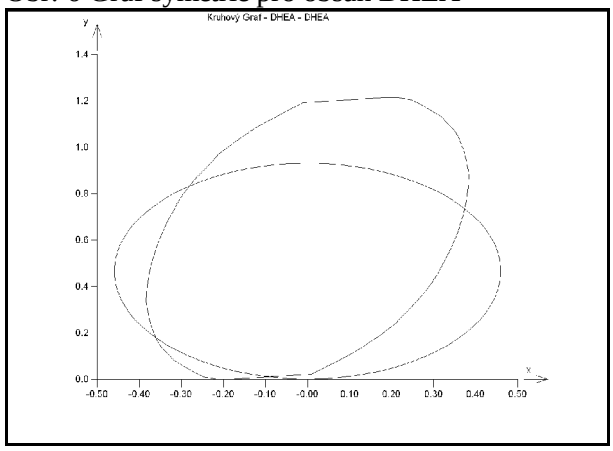
Obr. 5 Graf polosum pro obsah DHEA



Obr. 6 Graf symetrie pro obsah DHEA

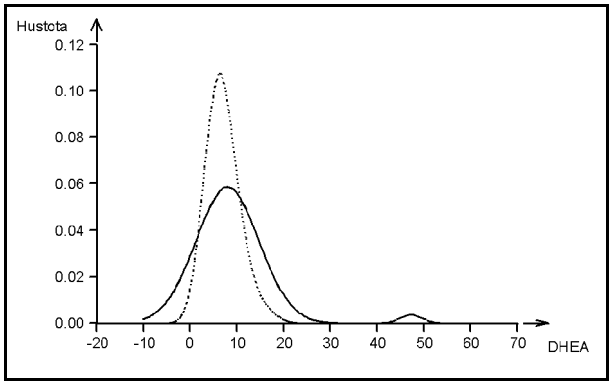


Obr. 7 Graf rozptýlení s kvantily pro obsah DHEA

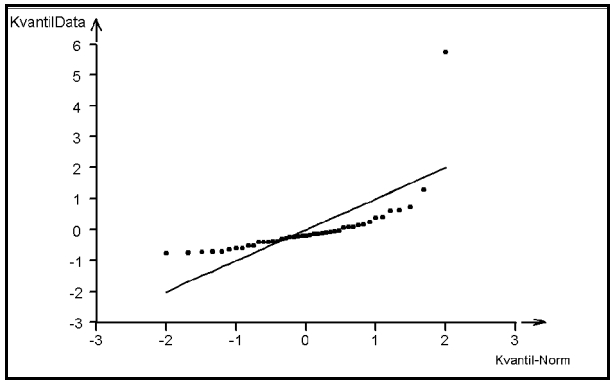


Obr. 8 Kruhový graf pro obsah DHEA

(c) Určení výběrového rozdělení (EDA): výběrové rozdělení je definováno svou symetrií, šikmostí a špičatostí a lze ho indikovat pomocí čtyř grafů: *Jádrový odhad hustoty pravděpodobnosti* (obr. 9) ukazuje nenormální rozdělení, protože obě křivky, teoretická aproximující normální rozdělení a empirická pro výběrové rozdělení, se významně odlišují. V *rankitovém Q-Q grafu* (obr. 10) většina bodů neleží na přímce normálního rozdělení, což je důkaz, že výběrové rozdělení není normálního charakteru. Korelační koeficient *Q-Q* grafu $r_{xy} = 0.9180$ ukazuje na silně sešikmené log.-normální rozdělení.



Obr. 9 Jádrový odhad hustoty pravděpodobnosti pro obsah DHEA



Obr. 10 Rankitový Q-Q graf pro obsah DHEA

Tabulka 2. Kvantilové míry polohy, rozptýlení a tvaru pro obsah DHEA[nmol/l]

Kvantil	P	Dolní kvantil Q_D	Horní kvantil Q_H	Rozsah R_Q	Polosuma Z_Q	Šikmost S_Q	Délka konců T_Q
Median	0.5	6.7	6.7	-	-	-	-
Kvartil	0.25	5.3	8.68	3.375	6.99	1.24	0
Oktil	0.125	3.75	10.78	7.025	7.26	0.52	0.73
Sedecil	0.0625	3.23	12.66	9.438	7.94	0.38	1.03

Délka oktilových konců T_E = 0.733 se liší od tabulované hodnoty pro normální rozdělení T_E = 0.534 a také sedecilových konců T_D = 1.028 se liší od tabulované hodnoty pro normální rozdělení T_D = 0.822. Bodový odhad šikmosti g_1 = 4.59 a bodový odhad špičatosti g_2 = 26.87 ukazují, že výběrové rozdělení je sešikmené a nedá se aproximovat normálním.

(d) Ověření základních předpokladů o reprezentativním náhodném výběru: vyšetřením základních předpokladů, kladených na reprezentativní, náhodný výběr bylo dosaženo těchto závěrů:

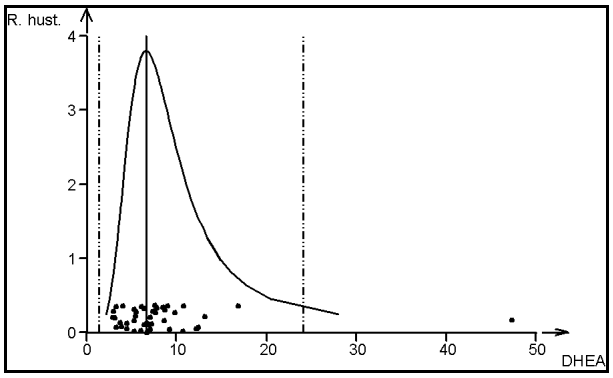
Vyšetření nezávislosti prvků výběru: von Neumannův test nezávislosti prvků ve výběru dospěl k hodnotě testačního kritéria t_{17} = 0.1996 < $t_{0.975}$(43+1) = 2.015, a proto je nezávislost přijata.

Vyšetření normality výběrového rozdělení: Jarque-Berrův test kombinované šikmosti a špičatosti vede k testační statistice C_1 = 1629.3 > χ^2(0.95, 2) = 5.992, což dokazuje, že předpoklad normality je zamítnut.

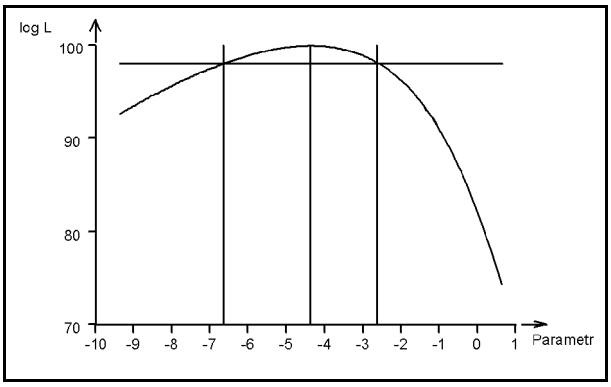
Vyšetření homogenity výběru: vně intervalu Hoaglinových mezí [B^*_L = -2.01; B^*_U = 15.98] leží 2 odlehlé hodnoty, 47.3 a 16.8.

(e) Transformace dat: asymetrické rozdělení výběru původních dat vyžaduje transformaci dat. Z grafu logaritmu maximální věrohodnosti plyne, že Box-Coxova transformace je statisticky významná, protože pod segmentem v tomto grafu neleží hodnota +1. Klasický odhad parametru polohy pro původní data aritmetický průměr $\bar{x}$ = 7.99 je nepoužitelný, protože není splněn předpoklad symetrického a normálního rozdělení. Symetrizující mocninná transformace (ADSTAT 1.25, $\hat{\lambda}$

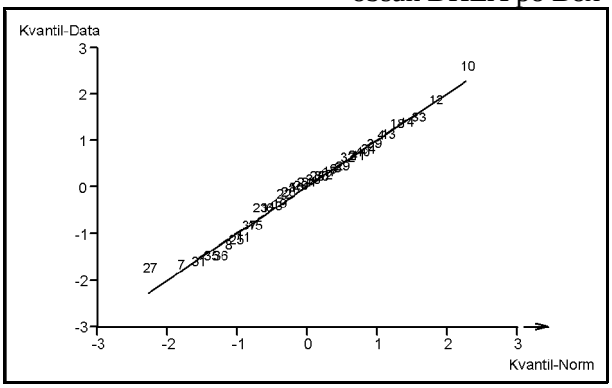
= -0.40 čili číslo blízké nule indikuje log.-normální rozdělení) vede na opravený průměr $\bar{x}_R$ = 6.45 se směrodatnou odchylkou s = 0.70 s intervalem spolehlivosti L_D = 5.58 a L_H = 7.51. Normalizační Box-Coxova transformace (ADSTAT 1.25, $\hat{\lambda}$ = -0.40) vede na stejný opravený průměr $\bar{x}_R$ = 6.45 se stejným intervalem spolehlivosti jako mocninná transformace.



Obr. 11 Hustota pravděpodobnosti pro obsah DHEA po Box-Coxově transformaci



Obr. 12 Graf logaritmu maximální věrohodnosti pro obsah DHEA po Box-Coxově transformaci



Obr. 13 Rankitový Q-Q graf pro obsah DHEA

(f) Závěry: diagnostiky průzkumové analýzy dat vedou k závěru, že 43 hodnot původních dat vykazuje asymetrické, silně sešikmené rozdělení log.-normální. Nelze proto použít klasické odhady parametrů polohy a rozptýlení $\bar{x}$ = 7.99 ± 2.11, platící pouze pro symetrické rozdělení. Data je třeba nejprve transformovat mocninnou nebo Box-Coxovou transformací. Re-transformovaný průměr pak představuje nejlepší odhad parametru polohy $\bar{x}_R$ = 6.45 ± 0.97. Rozdělení dat lze považovat za logaritmicko-normální. Retransformovanému průměru poměrně blízké odhady střední hodnoty přináší také robustní 40%ní uřezaný průměr $\bar{x}$(40%) = 6.78 ± 0.66, *M*-odhad $\hat{\mu}$ = 6.77 ± 0.94 nebo Hoggův *M*-odhad $\hat{\mu}$ = 6.77 ± 0.73. Je třeba si však uvědomit, že u biochemických a klinických dat nelze pro ztrátu informace vypouštět odlehlé hodnoty nebo užívat necitlivé robustní odhady.

2. Porovnání steroidů

Z výsledků je zřejmé, že mezi intenzitou akné a hladinami androgenů a jejich prekurzorů není přímý vztah. Lze uzavřít, že intenzita akné pravděpodobně není prediktorem intenzity hyperandrogenémie a intenzitu akné tedy ovlivňují spíše jiné faktory nezávislé na hladinách androgenů. U SHBG jako vazebného globulinu pohlavních hormonů byly dokonce nalezeny vyšší hladiny u žen s výraznějším akné (Tabulka 3).

ZÁVĚR

Symetrizující mocninná transformace a normalizující Boxova-Coxova transformace dat slouží k určení parametrů polohy pro případ nesymetrického rozdělení dat. Vlastní výpočet má postup:

1. Pro mocninnou transformaci se vypočtou různé míry symetrie a výběrová špičatost, a to v rozmezí $-3 \le \lambda \le 3$. Graficky je možno užít i Hinesův-Hinesové selekční graf k určení optimální hodnoty λ. Pro Box-Coxovu transformaci se vyčíslí také $\ln L(\lambda)$, různé míry symetrie a výběrová špičatost v rozmezí $-3 \le \lambda \le 3$. V transformaci se pak vyčíslí $\bar{y}$, $s^2(y)$, šikmost $g_1(y)$ a špičatost $g_2(y)$.

2. Z hodnot $\bar{y}$, $s^2(y)$, $g_1(y)$ a $g_2(y)$ se vyčíslí retransformované hodnoty $\bar{x}_R$ a 95%ní interval spolehlivosti střední hodnoty μ.

3. Mezi intenzitou akné a hladinami androgenů a jejich prekurzorů není přímý vztah. Intenzita akné není prediktorem intenzity hyperandrogenémie a intenzitu akné ovlivňují spíše faktory nezávislé na hladinách androgenů.

Poděkování

Autoři děkují za finační podporu Grantové agentury ČR, č. 303/00/1559.

Literatura:

[1] Meloun M., Militký J.: *Statistické zpracování experimentálních dat*, PLUS Praha 1994, ISBN 80-85297-56-6.

[2] Meloun M., Militký J.: *Statistické zpracování experimentálních dat - Sbírka úloh s disketou*, Univerzita Pardubice 1997, ISBN 80-7194-075-5.

[3] Kupka K.: *Statistické řízení jakosti*, Trilobyte Pardubice 1998, ISBN 80-238-1818-X.

[4] Militký J.: *Moderní statistické metody pro životní prostředí*, PHARE, Svazek 15, Vysoká škola báňská, Ostrava 1996, ISBN 80-7078-360-5.

[5] Cunliffe W. J., Shuster, S., Pathogenesis of acne. Lancet 1969, i, 65-7.

[6] Cunliffe W. J., Acne, hormones, and treatment. Br Med J 1982, 285, 912-3.

[7] Darley C. R., Moore J.W., Besser G.M., et al. Androgen status in women with late onset or persistent acne vulgaris. Clin Exp Dermatol 1984, 9, 28-35.

[8] Henze Ch., Hinney, B., Wuttke, W. Incidence of increased androgen levels in patients suffering from acne. Dermatology 1998, 196, 53-4.

[9] Lucky, A. W., McGuire J., Rosenfield R. L., Lucky P. A., Rich B. H., Plasma androgens in women with acne vulgaris. J Invest Dermatol, 1983, 81, 70-74.
[10] Schiavone F. E., Rietschel R. L., Sgoutas D., Harris R. Elevated free testosterone levels in women with acne. Arch Dermatol 1983,119, 799-802.
[11] Schmidt J. B., Lindmaier A., Spona J., Endocrine parameters in acne vulgaris. Endocrinol Exp 1990, 24, 457-64.
[12] Scholl G. M., Wu Ch., Leyden J., Androgen excess in women with acne. Obstet Gynecol 1984,64,683-88.
[13] Strauss J. S, Kligman A. M., Pochi P. E., The effect of androgens and estrogens on human sebaceous glands. J Invest Dermatol, 1962, 39, 139.
[14] Timpanapong P., Rojanasakul A., Hormonal profiles and prevalence of polycystic ovary syndrome in women with acne. J Dermatol 1997, 24, 223-9.
[15] Vexiau P., Husson C., Chivot M., et al., Androgen excess in women with acne alone compared with women with acne and / or hirsutism.
[16] Walton S., Cunliffe W. J., Keczkes K. et al., Clinical, ultrasound and hormonal markers of androgenicity in acne vulgaris. Br J Dermatol 1995, 133, 249-53.
[17] Palatsi R., Hirvensalo E., Liukko P., Malmiharju T., Mattila L., Riihiluoma P et al., Serum total and unbound testosterone and sex hormone binding globulin (SHBG) in female acne patients treated with two different oral contraceptives. Acta Derm Venereol 1984 64 517-523.

Tabulka 3. Porovnání středních hodnot vybraných steroidů u pacientů bez akne (-0) a s akne (-1)

Steroid	n	Průměr (Dolní; Horní mez)	Směr. odch.	Medián	Re-transform. Průměr (Dolní; Horní mez)	Re-transf. směr. odch.	Šikmost	Špičatost	Normalita	Test H_0: shodné rozptyly	Test H_0: shodné průměry
TESTO-0	43	2.10 (1.78; 2.42)	1.05	2.00 (1.66; 2.34)	1.98 (1.67; 2.31)	1.04	0.57	2.85	Přijata	H_0 přijata	H_0 přijata
TESTO-1	46	1.91 (1.61; 2.21)	1.01	1.65 (1.37; 1.93)	1.71 (1.46; 1.99)	0.89	**1.12**	4.02	Zamítnuta		
SHBG-0	42	59.4 (48.7; 70.0)	34.19	55.4 (44.1; 66.6)	54.7 (44.7; 65.6)	33.4	**1.08**	**4.38**	Zamítnuta	H_0 zamítnuta	H_0 zamítnuta
SHBG-1	45	84.3 (68.2; 100.4)	53.4	70.4 (55.8; 85.0)	72.9 (59.5; 88.3)	47.9	**1**	3.12	Přijata		
IFT-0	42	9.59 (0.13; 19.05)	30.4	4.24 (2.71; 5.77)	3.65 (2.64; 5.13)	2.76	**6.08**	**38.65**	Zamítnuta	H_0 zamítnuta	H_0 přijata
IFT-1	44	4.35 (2.34; 6.37)	6.63	2.33 (1.31; 3.35)	2.24 (1.66; 3.07)	1.83	**3.52**	**16.18**	Zamítnuta		
ADION-0	42	9.22 (7.66; 10.79)	5.03	8.23 (6.98; 9.47)	8.04 (6.91; 9.38)	1.71	**1.71**	**6.52**	Zamítnuta	H_0 přijata	H_0 přijata
ADION-1	46	9.66 (8.45; 10.86)	4.06	9.71 (8.25; 11.16)	9.25 (8.08; 10.49)	4.06	0.37	2.48	Přijata		
DHEAS-0	43	6.30 (5.45; 7.15)	2.76	5.70 (4.35; 7.05)	5.88 (5.11; 6.73)	2.62	0.67	2.69	Přijata	H_0 přijata	H_0 přijata
DHEAS-1	47	6.55 (5.64; 7.46)	3.09	5.75 (4.97; 6.53)	5.83 (5.10; 6.68)	1.67	**0.93**	3.04	Zamítnuta		
DHEA-0	43	7.99 (5.89; 10.10)	6.83	6.70 (5.84; 7.56)	6.45 (5.58; 7.51)	0.71	**4.59**	**26.87**	Zamítnuta	H_0 přijata	H_0 přijata
DHEA-1	46	7.34 (6.12; 8.56)	4.09	6.63 (5.67; 7.58)	6.43 (5.42; 7.60)	3.66	**0.93**	3.01	Zamítnuta		