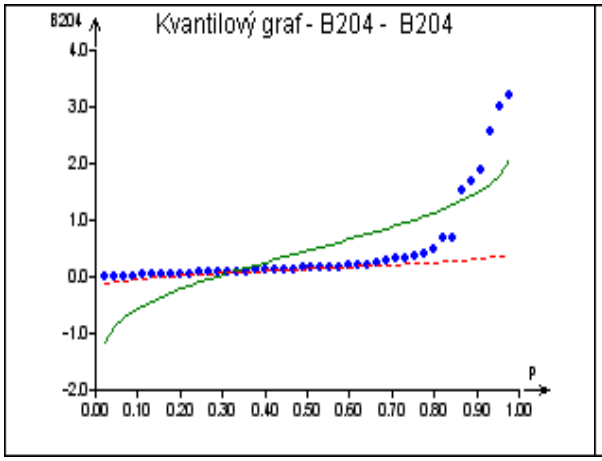
Vzorová úloha 2.2: *Průzkumová analýza velkého výběru*

Na úloze **B2.04** *Kontrola obsahu ergosterinu v calciferolu* ukážeme postup průzkumové analýzy dat. Při výrobě calciferolu se provádí kontrola meziproduktu 3,5 DNB esteru calciferolu metodou HPLC. Sleduje se také obsah přítomného ergosterinu jako nečistoty, jehož střední hodnota by neměla přesáhnout 0.4%. Metodou průzkumové analýzy dat vyšetřete, zda jsou splněny požadavky, kladené na náhodný výběr a zda je splněn i požadavek čistoty calciferolu. Určete typ rozdělení. Které diagnostiky shodně indikují vybočující hodnoty? Jak velké procento hodnot dosahuje obsahu 0.4%? Zkonstruujte barierově-číslicové schéma formou sedmi-písmenového zápisu výběru.

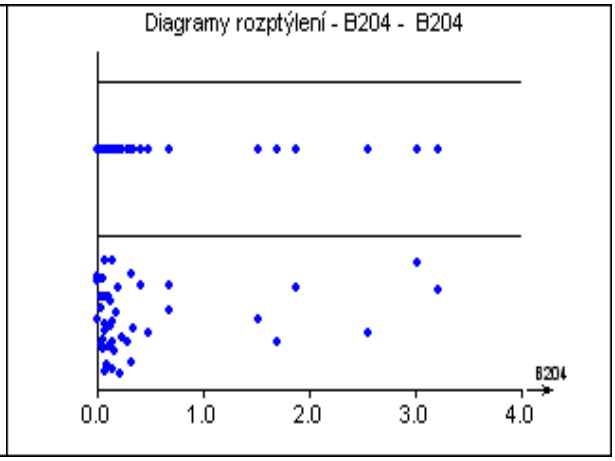
Řešení:

1. Zkoumání zvláštností dat: grafické diagnostiky indikují vedle stupně symetrie a špiča-tosti rozdělení také odlehlé body.

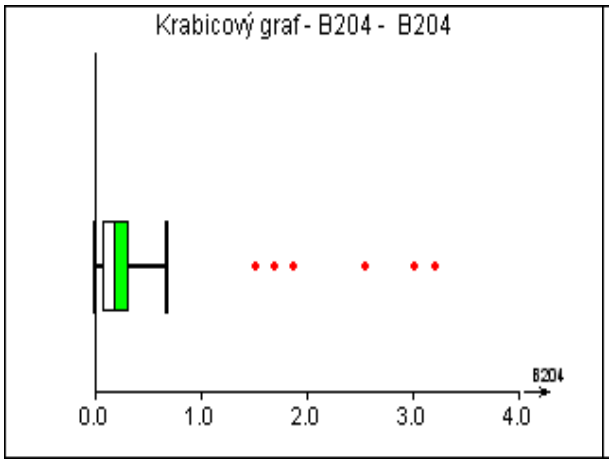
(a) *Odhalení stupně symetrie a špičatosti rozdělení:* celkem 12 grafických diagnostik indikuje symetrii a špičatost rozdělení s těmito závěry:



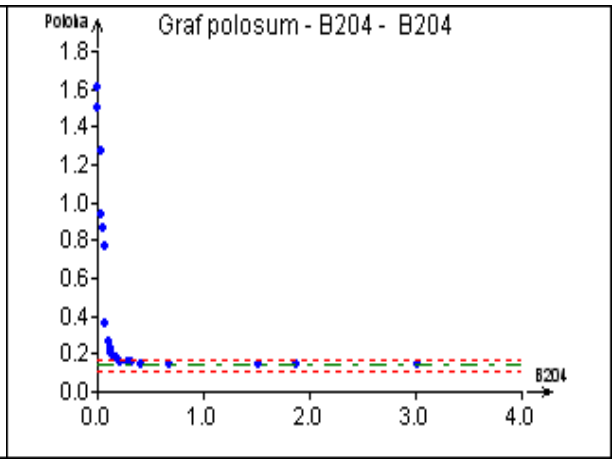
Obr. 2.20a Kvantilový graf



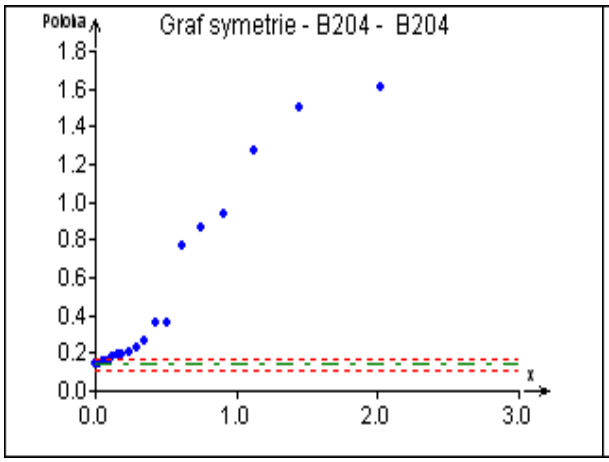
Obr. 2.20b Diagram rozptýlení a rozmítnutý diagram rozptýlení



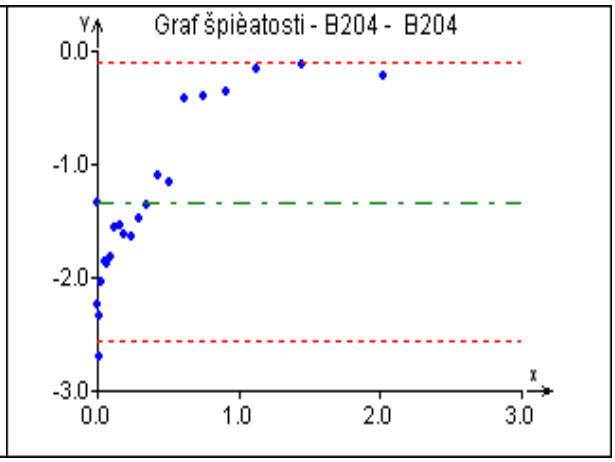
Obr. 2.20c Vrubový krabicový graf



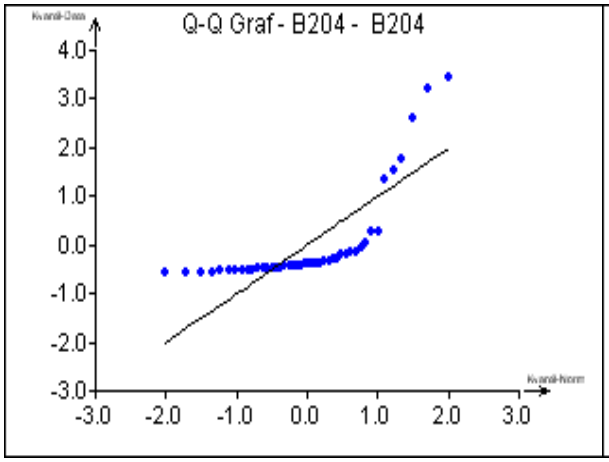
Obr. 2.20d Graf polosum



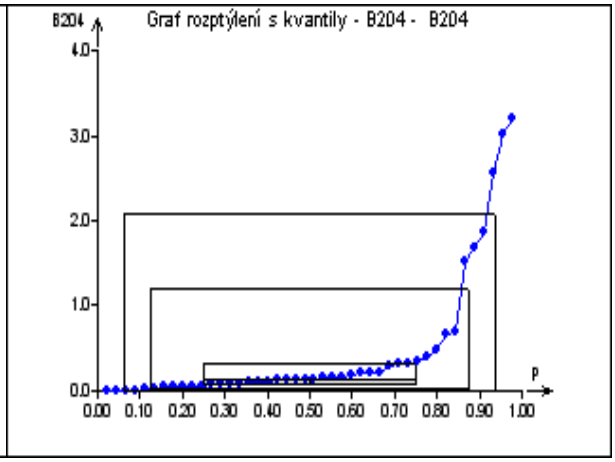
Obr. 2.20e Graf symetrie



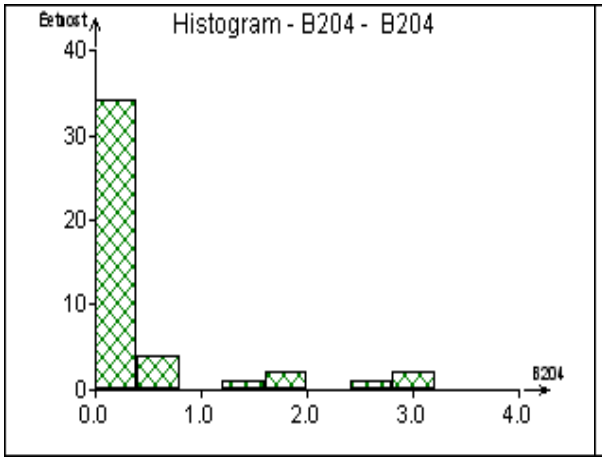
Obr. 2.20f Graf špičatosti



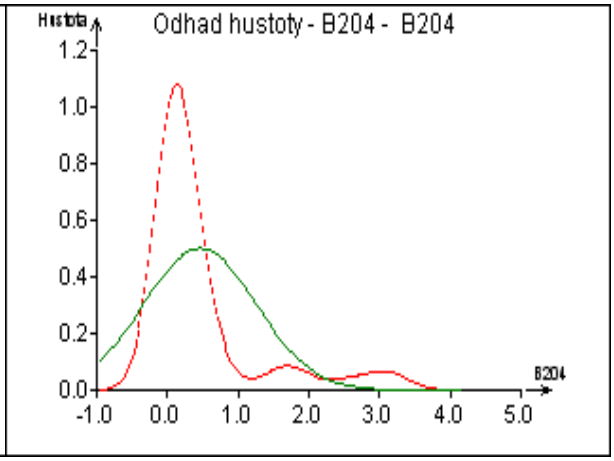
Obr. 2.20g Kvantil-kvantilový (Q-Q) graf



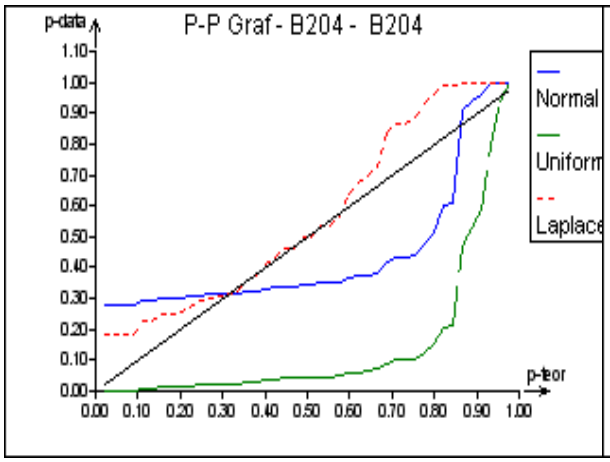
Obr. 2.20h Graf rozptýlení s kvantily



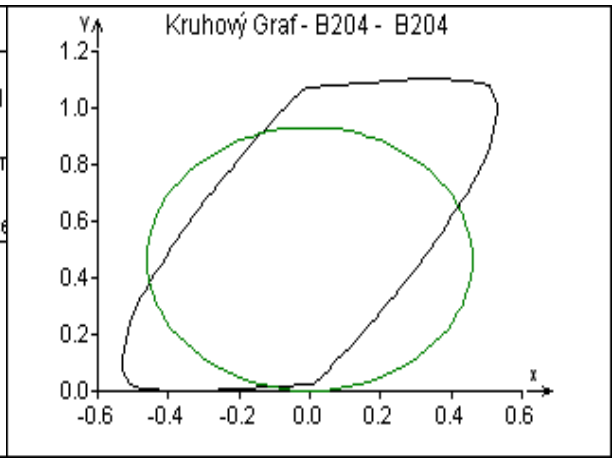
Obr. 2.20i Histogram



Obr. 2.20j Graf hustoty pravděpodobnosti



Obr. 2.20k P-P graf



Obr. 2.20l Kruhový graf

1. *Kvantilový graf (obr. 2.20a):* je patrný velký rozdíl mezi symetrickým Gaussovým a empirickým rozdělením. Tvar křivky je charakteristický pro asymetrické rozdělení, silně sešikmené k vyšším hodnotám.
2. *Diagram rozptýlení a rozmítnutý diagram rozptýlení (obr. 2.20b):* ukazuje na 4 odlehlé body v horní části diagramu a 1 odlehlý bod v dolní části diagramu.
3. *Vrubový krabicový graf (obr. 2.20c):* v horní části je detekováno 6 odlehlých bodů. Krabice je rozdělena na dvě části mediánem.
4. *Graf polosum (obr. 2.20d):* indikuje velkou část bodů jako vybočujících ze symetrického rozdělení. Body, ležící na mediánové rovnoběžce s *x*-ovou osou jsou ze symetrického rozdělení, ostatní nikoliv.
5. *Graf symetrie (obr. 2.20e):* indikuje valnou část bodů jako vybočujících ze symetrického rozdělení nebo patřících do asymetrického rozdělení.
6. *Graf špičatosti (obr. 2.20f):* většina bodů neleží na rovnoběžce s *x*-ovou osou pro symetrické rozdělení, a proto jde o rozdělení asymetrické.
7. *Kvantil-kvantilový (Q-Q) graf (obr. 2.20g):* jelikož většina bodů neleží na přímce jde o asymetrické rozdělení.
8. *Graf rozptýlení s kvantily (obr. 2.20h):* asymetrie kvantilových obdélníků obdélníků dokazuje silně asymetrické rozdělení. Body ležící vně sedecilového obdélníka indikuje tato pomůcka jako odlehlé.
9. *Histogram (obr. 2.20i):* zřetelně ukazuje na asymetrické rozdělení sešikmené k vyšším hodnotám.
10. *Jádrový odhad hustoty pravděpodobnosti (obr. 2.20j):* ve srovnání s Gaussovým rozdělením je patrné silné sešikmení k vyšším hodnotám. Empirickou křivku nelze aproximovat symetrickým Gaussovým rozdělením.
11. *Pravděpodobnostní P-P graf (obr. 2.20k):* empirická křivka nesouhlasí s žádnou křivkou symetrického rozdělení (norm., rovnoměrn. a Laplaceova). Rozdělení je asymetrické.
12. *Kruhový graf (obr. 2.20l):* tvar elipsy dokazuje silně asymetrické rozdělení sešikmené k vyšším hodnotám.

Kvantily a písmenové hodnoty **Úlohy B2.04** (ADSTAT)

Kvantily a písmenové hodnoty:

Procento	**Kvantil**	**Procento**	**Kvantil**
5	1.0000E-04	10	3.2300E-02
15	4.3900E-02	20	5.4800E-02

25	6.8250E-02	30	7.6700E-02
35	9.4300E-02	40	1.1520E-01
45	1.3000E-01	50	1.4000E-01
55	1.5000E-01	60	1.5820E-01
65	1.8230E-01	70	2.0100E-01
75	2.3175E-01	80	2.9440E-01
85	3.1910E-01	90	3.3430E-01
95	4.7800E-01		

Písmenové hodnoty:

Kvantil	**Písmeno**	**Pravděpodobnost**	**Spodní mez L_D**	**Horní mez L_H**
Sedecil	D	0.0625	1.0000E-04	4.3500E-01
Oktil	E	0.1250	3.6750E-02	3.2062E-01
Kvartil	F	0.2500	6.8250E-02	2.3175E-01
Medián	M	0.5000	1.4000E-01	1.4000E-01

Kvantilové míry:

Kvantil	***F (0.25)***	***E (0.125)***	***D(0.0625)***
Rozpětí R_L	1.6350E-01	2.8387E-01	4.3490E-01
Polosuma Z_L	1.5000E-01	1.7869E-01	2.1755E-01
Délka konců T_L	0.0000E+00	5.5172E-01	9.7830E-01
Šikmost S_L	1.1443E-01	5.8392E-02	-4.8843E-03
PseudoSigma G_L	1.2129E-01	1.2342E-01	1.4212E-01

Kvantily a písmenové hodnoty umožňují posoudit jednak symetrii výběrového rozdělení a jednak procento prvků ve výběru. Pro *procento* 45 je *kvantil* 0.1300, což znamená, že pod hodnotou 0.1300 leží 45% a nad ní 55% prvků výběru. *Písmenové hodnoty* a *kvantilové míry* umožňují sestrojení *graficko-tabelárního schéma písmenově-číslicového zápisu výběru* či sumarizace dat. Lišící se hodnoty *kvantilových polosum* (kvartilové, oktilové, sedecilové) indikují asymetrické rozdělení, v případě symetrického rozdělení by totiž všechny polosumy dosahovaly stejné hodnoty.

Sedmi-písmenový zápis výběru:

	Dolní kvantil L_D	(Polosuma)	Horní kvantil L_H	
Median M		0.1400		Rozpětí R_L
Kvartil F	0.06825	(0.1500)	0.23175	0.16350
Oktil E	0.03675	(0.17869)	0.32062	0.28387
Sedecil D	0.0001	(0.21755)	0.43500	0.43490

(b) *Indikace lokální koncentrace dat a rozdělení výběru:* rozdělení je asymetrické, s dlouhým horním koncem s větší koncentrací bodů ve spodní části hodnot. Z analýzy kvantil-kvantilového Q-Q grafu vyplývá, že nejvyšší hodnoty korelačního koeficientu je dosaženo především pro exponenciální rozdělení. Výběrové rozdělení je zde aproximováno exponenciálním.

Linearita v kvantil-kvantilovém grafu **Úlohy B2.04** (ADSTAT)

Linearita kvantil-kvantilovém (Q-Q) grafu $y = \beta_0 + \beta_1 x$:

Rozdělení	**Směrnice β_0**	**Úsek β_1**	**Korelační koeficient r_{xy}**
Laplaceovo	0.11077	0.17671	0.92263
Normální	0.14976	0.17671	0.92054
Exponenciální	0.16708	0.01259	0.98963
Rovnoměrné	0.48940	-0.06799	0.89212
Lognormální	0.09055	0.03413	0.96956

(c) *Nalezení vybočujících prvků ve výběru:* z grafických diagnostik průzkumové (explo-ratorní) analýzy výběru byly nalezeny 3 až 6 podezřelých bodů, které by mohly být chápany v symetrickém rozdělení jako odlehlé. Jelikož však jde o asymetrické rozdělení exponenciální, nemá smyslu zde indikovat odlehlé body.

2. Ověření předpokladů o datech: *Reprezentativní náhodný výběr* je charakterizován třemi důležitými předpoklady, které je třeba před vlastní analýzou ověřit. Jsou to nezávislost, homogenita a případná normalita výběru.

(a) *Ověření normality rozdělení:* Na předpokladu normality je založena celá statistická analýza dat. Test kombinace výběrové šikmosti a špičatosti ukázal, že normalita výběrového rozdělení je zamítnuta.

(b) *Ověření nezávislosti prvků výběru:* K identifikaci časové závislosti prvků výběru nebo závislosti související s pořadím jednotlivých měření se testuje významnost autokorelačního koeficientu prvního řádu podle von Neumannova testovacího kritéria. U analyzovaného výběru byla nezávislost prvků ve výběru prokázána.

(c) *Ověření homogenity rozdělení výběru:* Homogenní výběr znamená, že všechny jeho prvky pocházejí ze stejného rozdělení s konstantním rozptylem. Vybočující měření silně zkreslují odhady polohy a zejména rozptylu s^2, takže zcela znehodnocují další statistickou analýzu. Testování vybočujících měření bez doplňkových informací průzkumové analýzy dat je málo spolehlivé. Kritérium vnitřních mezí určilo 2 odlehlé body, a to bod č. 14 a 23. Doplněním informací z průzkumové analýzy lze identifikovat exponenciální rozdělení výběru, které již odlehlé body mít nebude, takže toto vyloučení dvou bodů nemá statistický smysl.

(d) *Určení minimálního rozsahu výběru:* Uvažujeme-li 25% relativní chybu směro-datné odchylky, bude minimální rozsah výběru n = 18, pro 10% relativní chybu směrodatné odchylky pak n = 110 a pro 5% relativní chybu směrodatné odchylky bude n = 437.

Základní předpoklady výběru **Úlohy B2.04** (ADSTAT)

(a) Odhady klasických parametrů:

Odhad aritmetického průměru $\bar{x}$:	0.177
Odhad rozptylu s^2:	2.536E-02
Odhad směrodatné odchylky s:	0.159
Odhad šikmosti $\hat{g}_1$:	1.54
Odhad špičatosti $\hat{g}_2$:	5.36
(b) Test normality: tabulkový kvantil $\chi^2_{1-\alpha}(2)$:	5.992
Odhad χ^2_{exp} statistiky:	35.87

Závěr: Předpoklad normality zamítnut na spočtené hladině významnosti α = 1.6254E-08

(c) Test nezávislosti: tabulkový kvantil $t_{1-\alpha/2}(n+1)$:	2.0141

Odhad von Neumannovy statistiky t_n: 0.4049

Závěr: Předpoklad nezávislosti přijat na spočtené hladině významnosti $\alpha = 0.3437$

(d) Detekce odlehlých hodnotami: metodou modifikované vnitřní hradby

Dolní vnitřní hradba B_D: -0.3054

Horní vnitřní hradba B_H: 0.6800

Závěr: Ve výběru jsou 2 odlehlé hodnoty.

Bod číslo 14 (horní): 0.6700

Bod číslo 23 (horní): 0.6800

Odhady parametrů s vynechanými odlehlými hodnotami:

Odhad aritmetického průměru $\bar{x}$: 0.153

Odhad rozptylu s^2: 1.391E-02

Odhad směrodatné odchylky s: 0.118

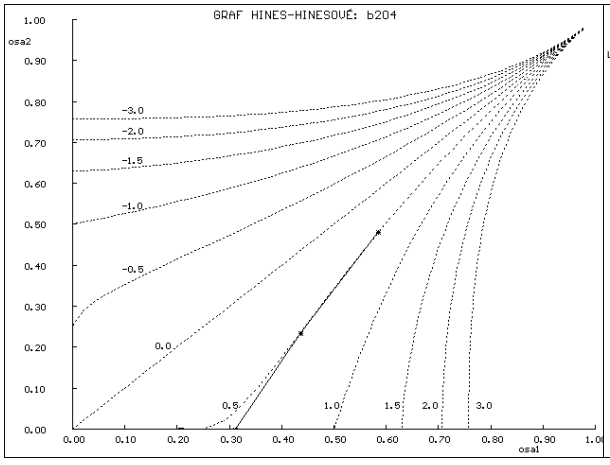
Odhad šikmosti $\hat{g}_1$: 0.89

Odhad špičatosti $\hat{g}_2$: 3.40

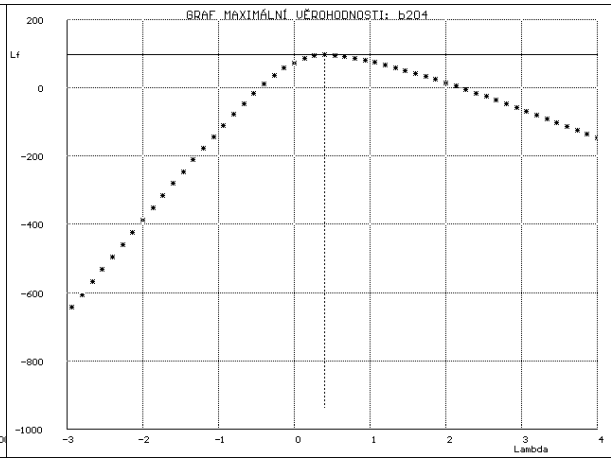
3. Transformace dat: Jelikož se na základě průzkumové analýzy dat zjistilo, že rozdělení výběru dat se systematicky odlišuje od rozdělení normálního, vyvstává zde problém, jak data vůbec vyhodnotit. V takovém případě je vhodná transformace dat, která vede ke stabilizaci rozptylu, zesymetričtění rozdělení a v případě Box-Coxovy transformace i k normalitě.

(a) *Mocninná transformace:* Zesymetričtění rozdělení výběru je možné provést užitím jednoduché (prosté) mocninné transformace, která sice nezachovává měřítko, není vzhledem k hodnotě λ všude spojitá a hodí se pouze pro kladná data.

Pro odhad exponentu λ se hledají optimální hodnoty charakteristik asymetrie (šikmosti) a špičatosti. K určení optimálního λ lze ale také užít orientační grafické metody, selekčního grafu dle Hinese a Hinesové, obr. 2.21a. Podle umístění experimentálních bodů v okolí teoretických křivek selekčního grafu byla odhadnuta velikost $\hat{\lambda} \approx 0.5$.



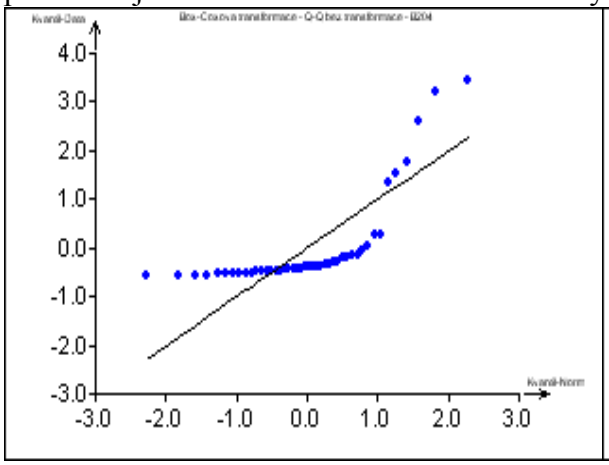
Obr. 2.21a Hinesův-Hinesové selekční graf pro výběr z exponenciálního rozdělení, ***ADSTAT***



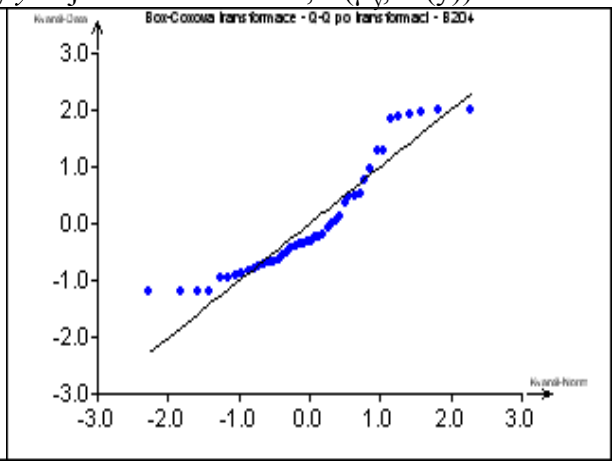
Obr. 2.21b Graf logaritmu věrohodnostní funkce na λ pro výběr z exponenciálního rozdělení, ***ADSTAT***

(b) **Box-Coxova transformace:** Pro přiblížení rozdělení výběru k rozdělení normálnímu vzhledem k šikmosti a špičatosti se užívá Boxovy-Coxovy transformace. Pro odhad parametru

λ v Boxově-Coxově transformaci lze užít metodu maximální věrohodnosti (obr. 2.21b) s tím, že pro $\lambda = \hat{\lambda}$ je rozdělení transformované veličiny *y* nejblíže normálnímu, $N(\mu_y, \sigma^2(y))$.



Obr. 2.22a Q-Q graf původních dat, ADSTAT



Obr. 2.22b Q-Q graf dat po Box-Coxově transformaci, ADSTAT

Průběh věrohodnostní funkce ln $L = f(\lambda)$ lze znázornit ve zvoleném intervalu např. $-3 \le \lambda \le 3$ a identifikovat maximum křivky v grafu věrohodnostní funkce tak, že *x*-ová souřadnice indikuje odhad $\hat{\lambda}$. Dva průsečíky křivky ln $L(\lambda)$ s rovnoběžkou s *x*-ovou osou indikují 100(1-α)%ní interval spolehlivosti parametru λ, tj. $\langle \lambda_D, \lambda_H \rangle$. Jelikož tento interval spolehlivosti neobsahuje číslo +1, je mocninná a Boxova-Coxova transformace ze statistického hlediska výhodná a má smyl ji užívat.

Zpětná transformace: Po vhodné transformaci určíme $\bar{y}$, $s^2(y)$ a potom pomocí zpětné transformace s využitím Taylorova rozvoje v okolí $\bar{y} = 0.35465$ odhadneme retransformované parametry $\bar{x}_R = 0.14318$ a $s^2_R = 0.020931$ původních dat. Uvedený postup vede vesměs k lepším odhadům polohy a rozptylu a je vhodný zvláště v případech takového asymetrického (exponenciálního) rozdělení výběru.

Závěr: Výběr pochází z exponenciálního rozdělení, a proto nejlepší odhad střední hodnoty získáme transformací dat, a to $\bar{x}_R = 0.143$ a směrodatnou odchylku $s = 0.145$. Konečně 95%ní interval spolehlivosti bude $L_D = 0.102$ a $L_H = 0.190$.