UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

BAKALÁŘSKÁ PRÁCE

Statistická analýza dojivosti v programu SAS

Vedoucí diplomové práce:
Mgr. Jaroslav Marek, Ph.D.
Rok odevzdání: 2010

Vypracovala:
Petra Kynčlová
ME, III. ročník

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením pana Mgr. Jaroslava Marka, Ph.D. s použitím uvedené literatury.

V Olomouci dne 20. dubna 2010

Poděkování

Na tomto místě bych chtěla hlavně poděkovat svému vedoucímu bakalářské práce panu Mgr. Jaroslavu Markovi, Ph.D., za jeho nesmírnou trpělivost, užitečné rady a čas, který mi věnoval při konzultacích, a Mgr. Janě Vrbkové za poskytnutí fakultní licence ke statistickému programu SAS Enterprise Guide. Poděkování patří rovněž všem, kteří mě během mého studia podporovali.

Obsah

Úvod

Ve své bakalářské práci se budu věnovat analýze dojivosti krav pomocí vybraných statistických metod. Cílem je pokusit se z dat o chovu dojnic holfštýnského typu získat informace, jejichž využití by mohlo přinést finanční efekt. K dispozici jsem měla dva typy dat. Pro statistickou analýzu dojivosti jsem použila data z více kravínů obsahující údaje o dojivosti krav, kvalitě mléka a jejich sledované biometrické údaje. U analýzy přežívání jsem vycházela z dat o vyřazování dojnic z chovu.

Má práce rovněž představuje statistický program SAS a jeho modul SAS Enterprise Guide, v němž jsem veškeré výpočty zpracovávala. První kapitolu věnuji právě statistickému programu SAS, kde stručně nastíním základní principy práce s touto aplikací.

Následně uvedu vzorek z dat ze sledování dojnic, se kterými jsem pracovala. Současně se pokusím vysvětlit některé základní pojmy.

Ve třetí kapitole se již budu zabývat vlastní statistickou analýzou dojivosti. Nejprve jsem pomocí popisné statistiky zmapovala dojivost ve zkoumaných kravínech. Vyslovené hypotézy ověřím vhodnými statistickými testy.

Ve čtvrté kapitole bude provedena analýza mléčné produkce v jednom zvoleném kravíně. Budu se věnovat kunínskému kravínu a pokusím se najít závislost dojivosti na aktuálním dnu laktačního cyklu.

V poslední kapitole využiji SAS k provedení analýzy přežívání (setrvání) dojnice v chovu. Zkonstruuji distribuční funkci, hustotu pravděpodobnosti, funkci přežití a rizikovou funkci pro dobu setrvání dojnice v chovu.

Ve své práci jsem vycházela z předpokladu, že čtenář je seznámen s teorií pravděpodobnosti a matematické statistiky v rozsahu bakalářského studia odborné matematiky. Neuvádím zde tedy základní definice a pojmy používané v obou těchto disciplínách.

1. SAS Enterprise Guide

1.1. Programový systém SAS

Ve své práci jsem k výpočtům využila svých znalostí práce se statistickým programovým systémem SAS, speciálně pak s jeho modulem SAS Enterprise Guide.

Programový systém SAS, vyvíjený firmou SAS se sídlem v USA, je program určený pro zpracování dat, jejich analýzu a následnou prezentaci. Největší přednost systému SAS spočívá v tom, že se nejedná pouze o úzce specializovaný statistický software, který má své široké využití ve všech možných oblastech. Systém je kromě statistické analýzy dat rovněž zaměřen na operační výzkum a systémovou analýzu, metody vhodné pro řízení podniků, bankovní a finanční služby. Umožňuje také uživateli programování, což znamená možnost přizpůsobit jeho činnost konkrétním požadavkům.

Program SAS je také významným článkem na trhu business inteligence pro jeho uživatelskou jednoduchost, správu a interoperabilitu. Díky těmto aspektům se stal jedním z nejpoužívanějších statistických softwarů v obchodních i akademických kruzích u nás i v zahraničí.

1.2. Modul SAS Enterprise Guide

Programový systém SAS je složen z několika modulů různých funkcí a možností pro kvalitní zpracování dat. Moduly můžeme rozdělit do několika skupin dle účelu jejich použití. Existují moduly zabývající se vkládáním dat a navrhováním formulářů, přístupem k datům z databázových systémů, víceuživatelským přístupem k datovým souborům SASu, statistickou analýzou a prognózováním, řízením jakosti a navrhováním experimentů, finančním plánováním a přípravou výstupů, řízením projektů a operačním výzkumem, laboratorním výzkumem nebo grafickou prezentací dat.

Modul SAS Enterprise Guide je jedním z mnoha modulů systému SAS. Jedná se o uživatelsky snadnou a příjemnou aplikaci, která zajišťuje přístup k většině funkcí SASu. Program nabízí přímý přístup k datům, jednoduchý export a import dat do jiných aplikací, vemi praktické a efektivní řešení zajištěné předem naprogramovanými úlohami pro jednotlivé analýzy nachystané k okamžitému použití.

Práce v programu SAS Enterprise Guide je organizována do jednotlivých projektů. Není potřeba jednotlivé procedury statistických metod žádným způsobem programovat, jelikož programovací kód se generuje automaticky v pozadí vytvářeného projektu s možností si ho během práce kdykoliv zpětně vyvolat.

Díky tomu, že není vyžadován programovací kód, a tudíž i programovací znalosti uživatele, se právě tento modul SASu stává použitelným i pro naprosté laiky. Zároveň umožňuje vytvářet snadná a rychlá řešení expertům, protože významným způsobem šetří čas strávený kódováním a programováním jednotlivých procedur.

1.3. Princip práce

Spustíme-li modul, zobrazí se nám hlavní okno neboli prostředí SAS Enterprise Guide, které se dělí do dalších menších oddělených podoken. Vzhledem k tomu, že modul je plně přizpůsobitelný uživateli, umožňuje nám mezi jednotlivými okny jednoduše přepínat, zavírat je a měnit jejich pozici dle potřeb během práce na statistických analýzách a při vypracovávání jiných úkolů.

Okna v prostředí SAS Enterprise Guide:

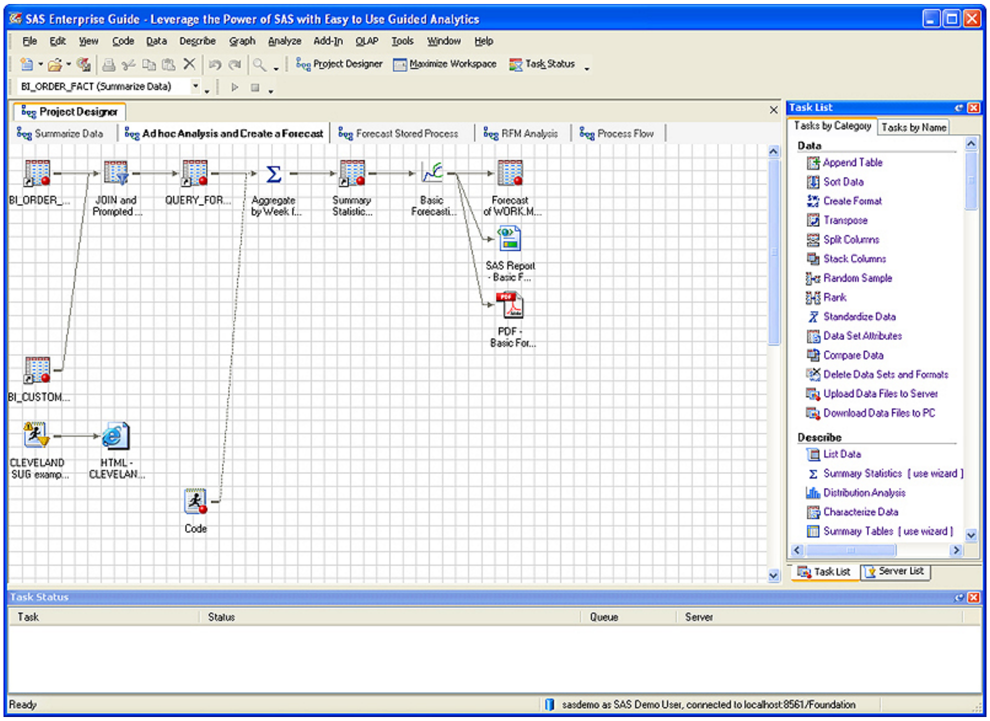
„Project Window“ – zobrazuje aktivní projekt včetně datového souboru, kódu procedur, poznámek a řešení.

„Task List“ – zobrazuje seznam všech možných úkolů a řešení v programu.

„Task Status“ – ukazuje stav, pozici a server právě spuštěné či probíhající procedury.

„Server List“ – představuje seznam všech dostupných serverů SASu.

Prostředí modulu neboli hlavní okno obsahuje kromě výše uvedených oken také pruh nabídek ve stylu aplikací pro Windows (např. File, Edit, View, Tasks, Tools, Help, atd.) a nástrojové lišty.



Obrázek č. 1: Pracovní prostředí v programu SAS Enterprise Guide.

Samotná práce v SAS Enterprise Guide probíhá následovně. Na začátku si vytvoříme nový projekt a pojmenujeme si jej. Pomocí příkazu File/Import Data přidáme datový soubor do projektu. Importovat můžeme textové soubory nebo data jiných formátů, rovněž také lze pracovat se vzdálenými daty. Následně na námi importovaný data set aplikujeme zvolenou proceduru vybranou přes panel nabídek. Po spuštění se zobrazí diaogové okno, kde má uživatel možnost zadefinovat jednotlivé proměnné a specifikovat své požadavky.

Po proběhnutí požadovaných „tasků“ dostaneme výsledné výstupy, se kterými můžeme dále pracovat. Modu SAS Enterprise Guide nám umožňuje jednak vytváření zajímavých reportů, ale také import a export dat různých formátů jako je Excel, ACCESS, HTML atd.

Je tedy zřejmé, že právě modul SAS Enterprise Guide poskytuje velmi jednoduché a rychlé řešení při získávání a zpracování statistických údajů jak kvalitativního tak kvantitativního charakteru. Zároveň uživateli podává přehledné výstupy a tím se stává jedním z uživatelsky velice jednoduchých pomocníků při řešení statistických analýz.

2. Údaje o dojnicích

Veškeré provedené statistické analýzy vycházejí ze souborů dat o chovu dojnic v kravíně Hradecká, Kačina, Křížanovsko, Kunín, Uherčice a Sedlec, viz [2].

Každá dojnice má přiřazené evidenční číslo, na jehož základě můžeme sledovat příslušná data kontroly a data otelení. Dále nám poskytnutá data ukazují pořadí aktuální laktace, neboli kolik telat dojnice porodila. Průměrná délka gravidity u skotu je 280 dní s krajními hodnotami 270 až 300 dní.

Rozdíl mezi datem kontroly a datem otelení nám udává den laktace, v němž se kráva právě nachází. Rovněž se dozvídáme důležité informace o kvalitě mléka, které jsou u dojnice sledovány.

Vzorek dat ze sledování dojivosti je uveden v následující tabulce.

číslo dojnice	datum kontroly	datum otelení	pořadí laktace	dojivost v kg	tuk
CZ000158239981	7. 8. 2009	28. 7. 2009	1	30.3	3.69
CZ000125133704	7. 8. 2009	16. 7. 2009	5	42.4	3.91
CZ000119382981	8. 7. 2009	18. 1. 2009	3	10.4	3.93
CZ000122401981	8. 7. 2009	10. 12. 2008	2	35.0	3.8
CZ000122373981	8. 7. 2009	28. 10. 2008	2	33.1	3.8

číslo dojnice	bílkovina	laktóza	počet somat. buněk	T/B	den laktace
CZ000158239981	3.56	4.24	143	1.04	10
CZ000125133704	2.72	4.77	216	1.44	22
CZ000119382981	3.52	4.33	1328	1.12	171
CZ000122401981	4.38	3.71	4491	0.87	210
CZ000122373981	3.15	5.18	58	1.21	253

3. Statistická analýza dojivosti ve zkoumaných kravínech

V této kapitole se pokusím analyzovat dojivost ve vybraných kravínech pomocí vhodných statistických metod, viz [3, 4, 5].

Nejprve bych se chtěla věnovat základním číselným charakteristikám náhodné veličiny, se kterými budu nejčastěji v této práci pracovat.

- k-tý obecný moment $\mu'_k = \mathbf{E}[X^k]$.
- 1. obecný moment nazýváváme střední hodnotou
  - pro diskrétní veličinu $\mathbf{E}X = \sum\limits_i x_i P(X = x_i)$,
  - pro spojitou veličinu $\mathbf{E}X = \int\limits_{-\infty}^{+\infty} x f_x dx$.
- k-tý centrální moment $\mu_k = \mathbf{E}[X - \mathbf{E}(X)]^k$.
- 2. centrální moment nazýváme rozptylem a je dán vztahem:
  $\operatorname{var} X = E[X - E(X)]^2$.
- Druhou odmocninu z rozptylu $\sqrt{\operatorname{var} X}$ označujeme jako směrodatnou odchylku.
- α-kvantil (kde $\alpha \in (0, 1)$) náhodné veličiny X je takové reálné číslo x_α, pro které platí $P(X \leq x_\alpha) \leq \alpha$ a současně $P(X \geq x_\alpha) \leq 1 - \alpha$.

  Některé kvantily mají své speciální názvy:
  - Medián je $x_{0,5}$,
  - Dolní kvantil je $x_{0,25}$,
  - Horní kvantil je $x_{0,75}$.
- Modus značíme $\hat{x}$ nebo také M_0.
  - Je-li X absolutně spojitá, je $\hat{x}$ bod, ve kterém má hustota f_X lokální maximum.
  - Je-li X diskrétní, pak je $\hat{x}$ její nejpravděpodobnější hodnota.

Modus nemusí vždy existovat a není určen jednoznačně.

- Kvartilové rozpětí je $R_Q = X_{0,75} - X_{0,25}$.
- Šikmost $A_3 = \frac{\mu_3'}{\sqrt{\mu_2'^3}}$.
- Špičatost $A_4 = \frac{\mu_4'}{\mu_2'^2}$.

Máme-li náhodný výběr $X_1, \ldots, X_n$, pak odhady výše zmíněných charakteristik určíme následovně:

- k-tý výběrový obecný moment $M_k' = \frac{1}{n} \sum_{i=1}^{n} X_i^k$.

  1. obecný výběrový moment se nazývá výběrový průměr a označuje se $\overline{X}$.
- k-tý centrální výběrový moment $M_k = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^k$.

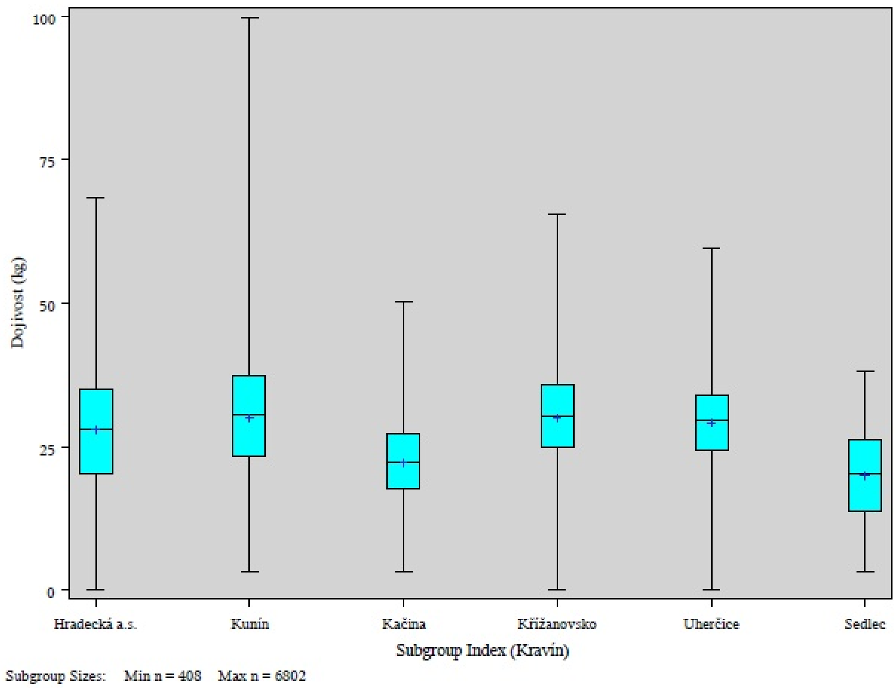
  Nejpoužívanější 2. centrální výběrový moment se nazývá výběrový rozptyl a značíme ho S_X^2 .

  Druhá odmocnina z rozptylu $\sqrt{S_X^2}$ se nazývá výběrová směrodatná odchylka.
- Výběrová šikmost $A_3 = \frac{M_3}{\sqrt{M_2^3}}$.
- Výběrová špičatost $A_4 = \frac{M_4}{M_2^2}$.

3.1. Popisná statistika

Při zpracování práce jsem měla k dispozici jisté statistické výsledky z článku [1]. Tyto poznatky napomohly pro vytypování vhodného postupu při analýze dat. Nejprve se s využitím popisné statistiky a statistické grafiky pokusím navrhnout hypotézy, které je vhodné studovat.

Pomocí krabicového diagramu můžeme porovnat dojivost v jednotlivých kravínech a to podle různých charakteristik - symetrie či asymetrie, mediánu, variability, existence extrémních hodnot.

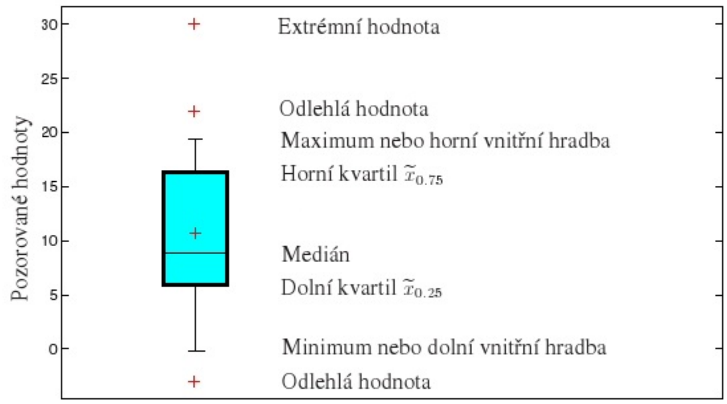


Obrázek č. 2: Variabilita dojivosti v pozorovaných kravínech.

Pro správné pochopení významu boxplotu zavedeme následující pojmy:

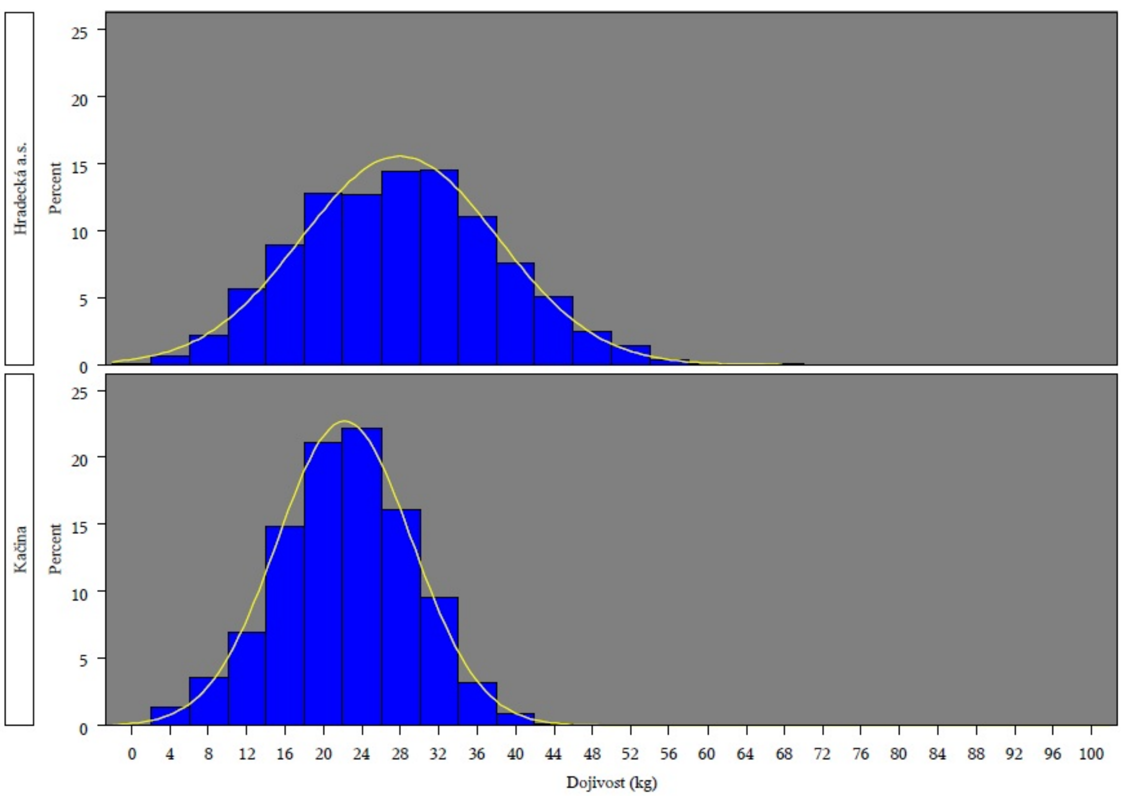
- dolní vnitřní hradba: $X_{0,25} - 1{,}5R_Q$,
- horní vnitřní hradba: $X_{0,75} + 1{,}5R_Q$,
- dolní vnější hradba: $X_{0,25} - 3R_Q$,
- horní vnější hradba: $X_{0,75} + 3R_Q$.

Symbol křížku v obrázku označuje extrémní hodnoty, které leží za vnějšími hradbami.

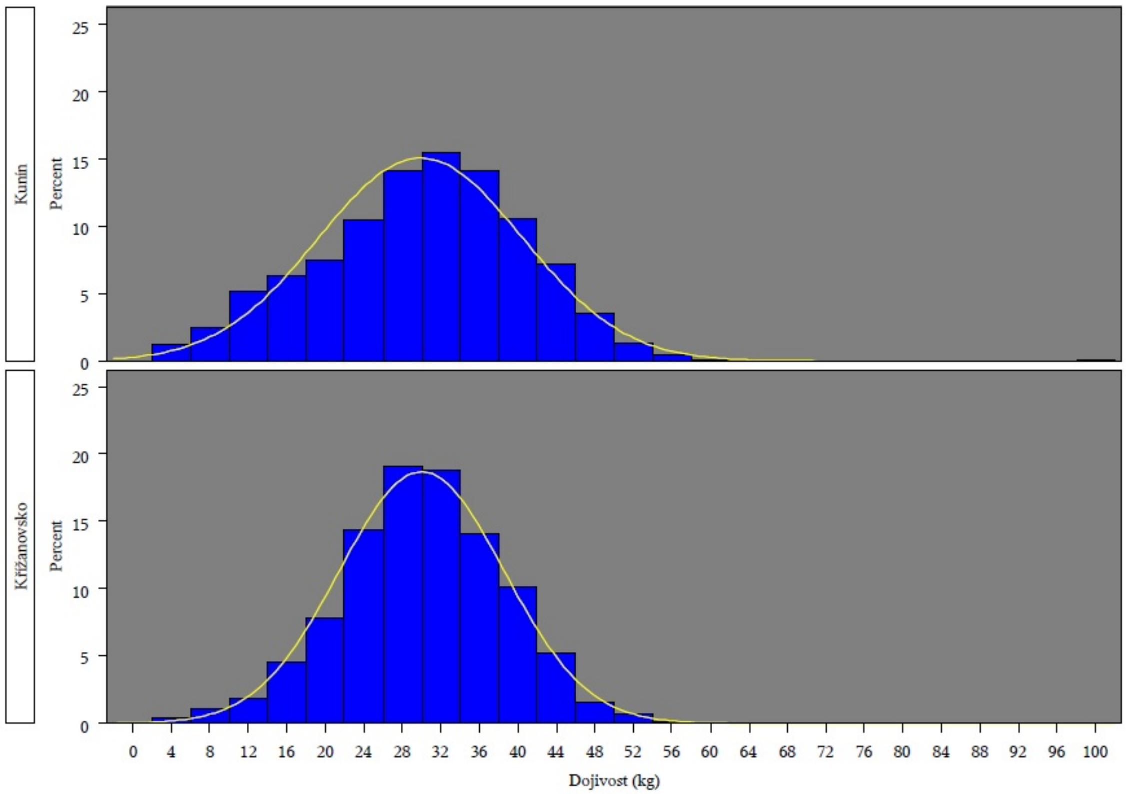


Obrázek 3.: Konstrukce boxplotu.

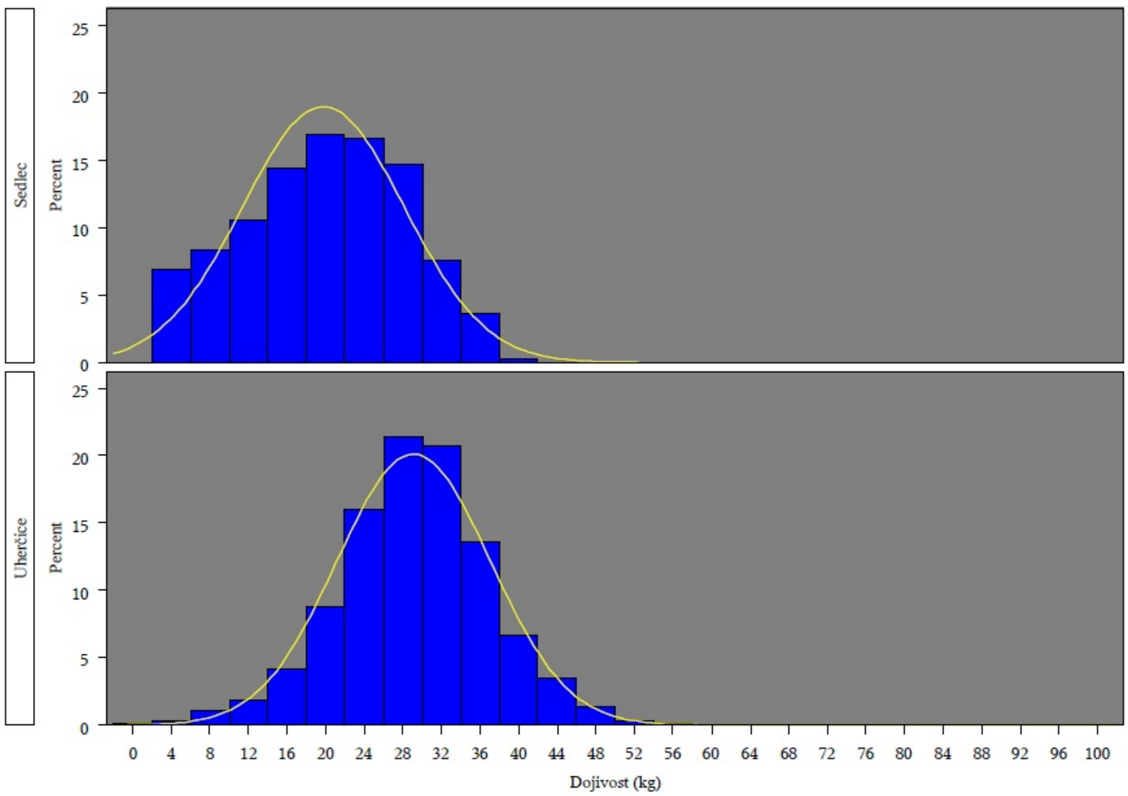
Sestavme si nyní histogramy dojivosti pro testované kravíny aproximované křivkou hustoty distribuční funkce normálního rozdělení.



Obrázek č. 4: Histogram dojivosti v kravínech Hradecká a Kačina.



Obrázek č. 5: Histogram dojivosti v kravínech Kunín a Křížanovsko.



Obrázek č. 6: Histogram dojivosti v kravínech Sedlec a Uherčice.

Na základě uvedených grafů se lze domnívat, že zkoumané výběry pocházejí z normálního rozdělení. V dalších kapitolách se normalitu pokusím ověřit pomocí vhodných statistických testů.

3.2. Testování statistických hypotéz

Testování statistických hypotéz nám umožňuje posoudit, zda experimentálně získaná data odpovídají předpokladu, který jsme vyslovili před provedením testování. Statistickou hypotézou vyslovujeme určitý předpoklad o rozdělení náhodných veličin. Hypotézy se nazývají parametrické, týkají-li se předpoklady hodnot parametrů náhodné veličiny. V opačném případě mluvíme o hypotézách neparametrických.

Hypotézu, kterou chceme testovat, zjistit zda platí či neplatí, nazýváme nulovou hypotézou a značíme ji H_0. Naopak názor pochybnosti o platnosti nulové hypotézy vyjadřuje hypotéza alternativní označovaná jako H_A. Mluvíme o testování nulové hypotézy H_0 ve prospěch alternativy H_A.

Jsou-li vysloveny předpoklady o rozdělení pravděpodoností zkoumané náhodné veličiny X, zvolíme vhodnou statistiku T, kterou označujeme jako testovací kritérium. Testovací statistika je výběrovou funkcí, jež má vztah k nulové hypotéze a za platnosti H_0 musíme znát její rozdělení pravděpodobností. Najdeme vhodnou množinu W, kterou nazveme kritickým oborem. Padne-li výběrová hodnota testovacího kritéria do kritického oboru, pak nulovou hypotézu H_0 zamítáme. V opačném případě H_0 nezamítáme.

Při rozhodování o přijetí či nepřijetí H_0 se můžeme dopustit jedné ze dvou chyb.

	H_0 je správná	H_0 je chybná
H_0 zamítneme	chyba 1. druhu = α	$1 - \beta$
H_0 nezamítneme	$1 - \alpha$	chyba 2. druhu = β

Pravděpodobnost α chyby 1. druhu se nazývá hladina významnosti testu neboli hladina testu. Představuje riziko, že rozhodnutí o zamítnutí nulové hypotézy H_0 je chybné. Číslo $1 - \beta$ označované jako síla testu určuje pravděpodobnost, že H_0 zamítneme a ona skutečně neplatí.

Ve všech testech, které budu ve své práci zmiňovat, bude zvolena hladina významnosti $\alpha = 0.05$.

Vzhledem k tomu, že jsem testy prováděla pomocí statistického programu SAS, rozhodovala jsem o závěrech provedených testů na základě často používané p-hodnoty. P-hodnota značí nejmenší hladinu, při které bychom ještě hypotézu zamítli. Vyjadřuje nám pravděpodobnost, že by testovací kritérium dosáhlo své hodnoty nebo hodnoty ještě více odporující nulové hypotéze. Je-li p-hodnota menší než předem stanovené α, nulovou hypotézu H_0 zamítáme.

3.2.1. Dvouvýběrový t test

Chceme-li porovnat dojivost dvou kravínů, nabízí se nám použít dvouvýběrový t test [3, 5].

Nechť $X_1, \ldots, X_m$ je výběr z $N(\mu_1, \sigma^2)$ a nechť $Y_1, \ldots, Y_n$ je výběr z $N(\mu_2, \sigma^2)$, kde $m \leq 2, n \leq 2, \sigma_2$ 0. Zároveň předpokládáme, že oba výběry jsou na sobě nezávislé. Označme

$$\overline{X} = \frac{1}{m} \sum_{i=1}^{m} X_i,$$

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{m} Y_i,$$

$$S_X^2 = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \overline{X})^2,$$

$$S_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

Za těchto předpokladů platí, že náhodná veličina

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{(m-1)S_X^2 + (n-1)S_Y^2}} \sqrt{\frac{mn(m+n-2)}{m+n}}$$

má rozdělění t_{m+n-2}.

Testujeme přitom hypotézu $H_0 : \mu_1 - \mu_2 = \triangle$, kde $\triangle$ je dané číslo, nejčastěji $\triangle = 0$, oproti alternativě $H_1 : \mu_1 - \mu_2 \neq \triangle$. Je-li $|T| \leq t_{m+n-2}(\alpha)$ zamítáme hypotézu H_0 na hladině α.

Mezi předpoklady dvouvýběrového t testu patří normalita obou výběrů a stejný rozptyl v obou výběrech. Porušení těchto předpokladů však mívá jen malý vliv na výsledek testu. V případě výrazné nenormality bychom dali přednost některému z neparametrických testů.

V našem případě zkoumáme, zda se statisticky významně liší střední hodnoty dojivosti v kravíně Kunín a Hradečná.

Kravín	N	Mean	Std Dev	Std Err	Minimum	Maximum
Hradecká a.s.	2456	27.9339	10.2629	0.2071	0	68.4000
Kunín	4966	29.8810	10.5927	0.1503	3.0000	99.9000
Diff (1-2)		-1.9471	10.4847	0.2586		

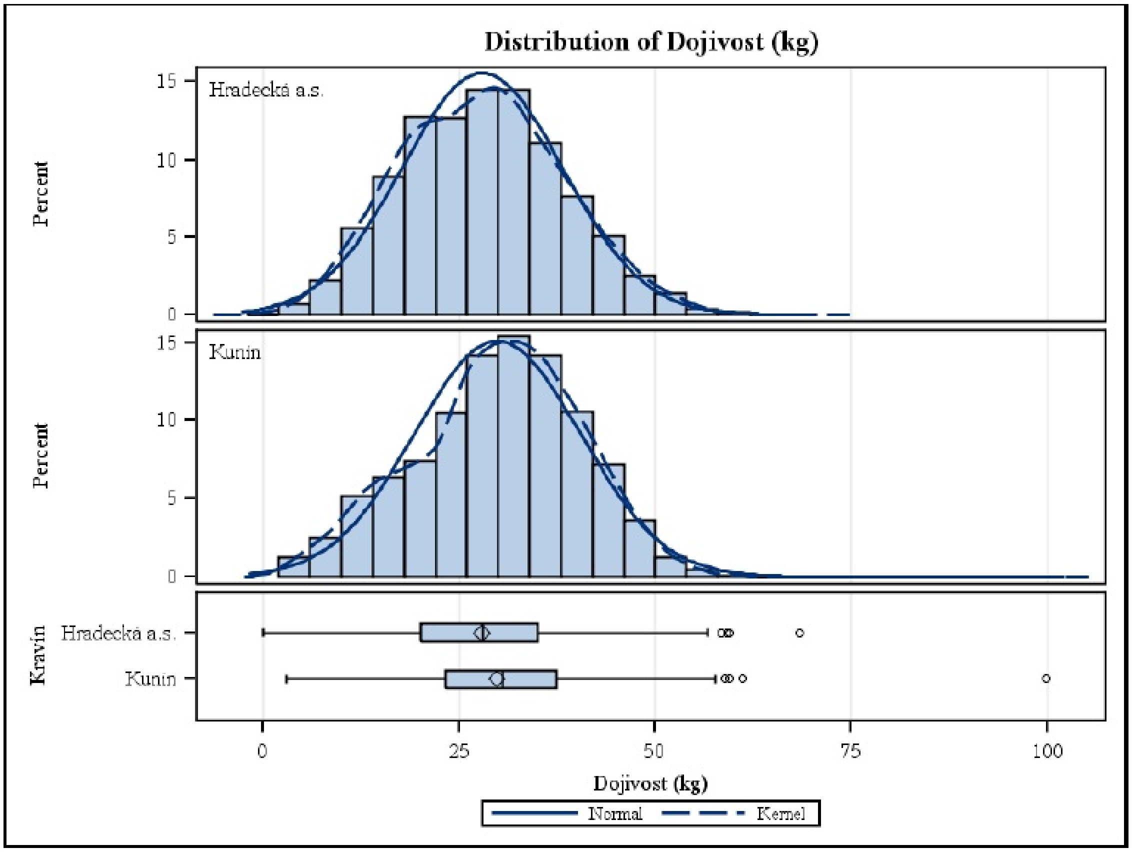
Kravín	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Hradecká a.s.		27.9339	27.5278	28.3400	10.2629	9.9838	10.5583
Kunín		29.8810	29.5863	30.1757	10.5927	10.3884	10.8052
Diff (1-2)	Pooled	-1.9471	-2.4541	-1.4401	10.4847	10.3187	10.6562
Diff (1-2)	Satterthwaite	-1.9471	-2.4487	-1.4454			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	7420	-7.53	<.0001
Satterthwaite	Unequal	5032.6	-7.61	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	4965	2455	1.07	0.0715

Z tabulky vidíme, že p-hodnota je menší než námi zvolená hladina významnosti α nastavená na 0,05. To znamená, že nulovou hypotézu o shodnosti středních hodnot obou kravínu zamítáme.

Použití právě této varianty t testu je možné vzhledem k výsledku testu shody rozptylů ve spodní tabulce, kde p-hodnota je rovna 0,0715.



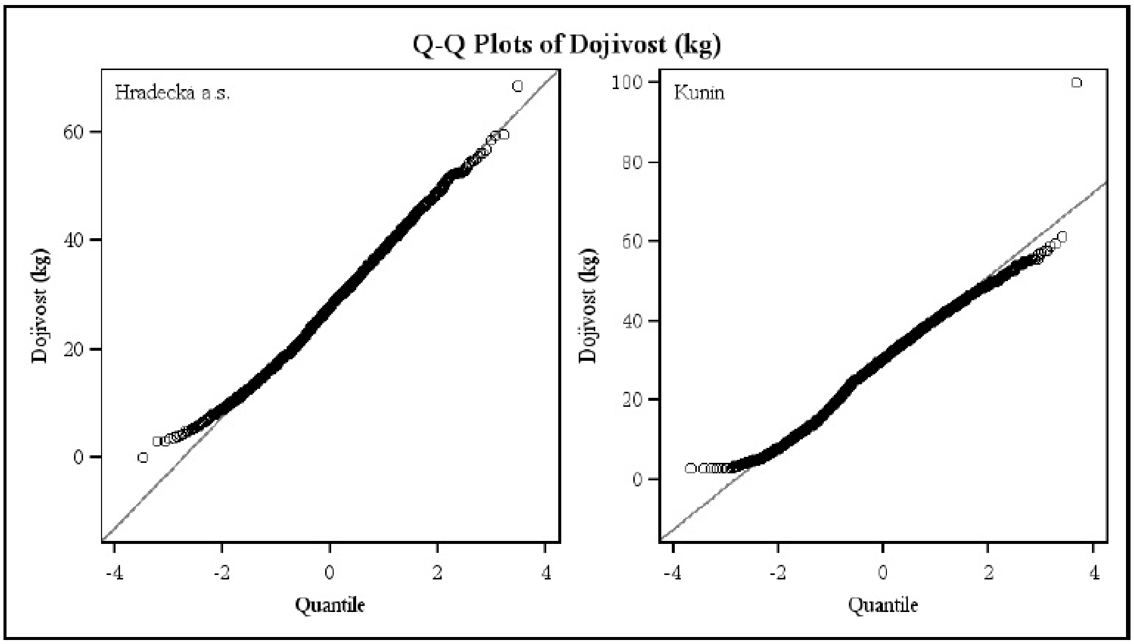
Obrázek č. 7: Srovnání kravínů Hradecká a Kunín v programu SAS.

Normalitu dat můžeme ověřit i pomocí QQ plotu. QQ plot nám umožňuje graficky posoudit, zda data pocházejí z nějakého známého rozložení. Pokud body leží na zobrazené přímce, lze data považovat za výběr z normálního rozdělení. nacházejí-li se dál od přímky, vzdalují se data od předpokládané normality.

Při konstrukci grafu se nejprve na svislou osu vynesou uspořádané hodnoty $x_{(1)}, \ldots, x_{(n)}$ a na osu vodorovnou kvantily K_{α_j} vybraného rozložení.

$$\alpha_j = \frac{j - r_{adj}}{n + n_{adj}}\,,$$

přičemž r_{adj} a $n_{adj} \leq 0{,}5$ jsou korigující faktory, implicitně $r_{adj} = 0{,}375$ a $n_{adj} = 0{,}25$ (Jsou–li některé hodnoty $x_{(1)} \leq \cdots \leq x_{(n)}$) stejné, pak za j bereme průměrné pořadí odpovídající takové skupince. Pokud vybrané rozložení závisí na nějakých parametrech, pak se tyto parametry odhadnou z dat nebo je může zadat uživatel na základě teoretického modelu. Body $(K_{\alpha_j}, x_{(j)})$ se metodou nejmenších čtverců proloží přímka. Čím méně se body $(K_{\alpha_j}, x_{(j)})$ odchylují od této přímky, tím je lepší soulad mezi empirickým a teoretickým rozložením.



Obrázek č. 8: QQ ploty dojivosti pro kravíny Hradecká a Kunín.

3.2.2. Bartlettův test

Dvouvýběrový t test ovšem můžeme použít jen pro porovnání dojivosti u dvou kravínů. Chceme-li zkoumat charakteristiky u většího počtu kravínů, musíme sestrojit ANOVU, kterou však lze použít pouze v případě, že data pochází z normálního rozdělení se stejnou variabilitou.

Pro posouzení homoskedasticity u našich dat použijeme Bartlettův test [3]. Budeme testovat hypotézu

$$H_0 : \sigma_1^2 = \cdots = \sigma_k^2$$

proti alternativě H_1, že H_0 neplatí.

Nechť

$$s_i^2 = \frac{1}{n_i - 1}\left(\sum_{j=1}^{n_i} Y_{ij}^2 - n_i y_i^2\right),$$

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{k}(n_i - 1)s_i^2,$$

$$C = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{n-k}\right),$$

$$B = \frac{1}{C}\left[(n-k)\ln s^2 - \sum_{i=1}^{k}(n_i - 1)\ln s_i^2\right],$$

kde k je počet výběrů.

Za platnosti H_0, má náhodná veličina B přibližně χ^2_{k-1} rozdělení, takže H_0 zamítáme v případě, že $B \geq \chi^2_{k-1}(\alpha)$. Požadujeme však, aby rozsahy výběrů $n_i, \ldots, n_k$ byly dostatečně velké. Nejčastěji se v literatuře uvádí podmínka, že musí platit $n_i > 6$ pro každé $i = 1, \ldots, k$.

Počty dojnic a příslušné výběrové charakteristiky jsou uvedeny v tabulce.

Level of Kravín	N	Dojivost (kg)	
		Mean	Std Dev
Hradecká a.s.	2456	27.9339169	10.2629459
Kačina	6802	22.1773155	7.0346009
Kunín	4966	29.8809907	10.5926816
Křižanovsko	2357	30.0077641	8.5414208
Sedlec	408	19.7448529	8.4064339
Uherčice	1357	29.1128224	7.9326273

Bartlett's Test for Homogeneity of Dojivost (kg) Variance			
Source	DF	Chi-Square	Pr > ChiSq
Kravín	5	1138.7	<.0001

Nulovou hypotézu o homoskedasticitě výběrů tedy zamítáme na hladině $\alpha = 0.05$. Z tohoto důvodu tudíž nemůžeme použít ANOVU. Místo toho použijeme její neparametrickou podobu, a to Kruskal-Wallisův test.

3.2.3. Kruskalův–Wallisův test

Tento test je neparametrickou obdobou analýzy rozptylu jednoduchého třídění a zobecněním dvouvýběrového Wilcoxonova testu. Používá se zejména v případech, máme-li výběry z rozdělení značně se lišících od normálního [3].

Nechť $Y_{i1}, \ldots, Y_{in_i}$ je výběr z nějakého rozdělení se spojitou distribuční funkcí F_i, $i = 1, \ldots, k$. Nechť všechny tyto výběry jsou na sobě nezávislé.

Budeme testovat hypotézu

$$H_0 : F_1(x) = \cdots = F_k(x) \text{ pro všechna } x$$

proti alternativě H_1, že H_0 neplatí.

Všechny veličiny Y_{ij} dohromady vytvoří sdružený náhodný výběr o rozsahu $N = n_1 + \cdots + n_k$. Veličiny uspořádáme do rostoucí posloupnosti a určíme pořadí R_{ij} každé veličiny Y_ij ve sdruženém výběru. Toto pořadí je možné zapsat do schématu uvedeného v tabulce.

Tabulka 1 Kruskalův-–Walisův test [3]

Výběr	**Pořadí veličin ve sdruženém náhodném výběru**				**Součet pořadí**
1	R_{11}	R_{12}	$\cdots$	R_{1n_1}	T_1
2	R_{21}	R_{22}	$\cdots$	R_{2n_2}	T_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
I	R_{k1}	R_{k2}	$\cdots$	R_{In_k}	T_k

Celkový součet všech pořadí je $T_1 + \cdots + T_k = N(N+1)/2$.

Jako testová statistika se používá

$$Q = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(N+1)\,.$$

Nulovou hypotézu zamítáme, pokud

$$Q \geq \chi^2_{k-1}(\alpha)\,.$$

V našem případě získáme následující hodnoty:

Wilcoxon Scores (Rank Sums) for Variable Dojivost (kg) Classified by Variable Kravín					
Kravín	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Hradecká a.s.	2456	24288757.0	22530116.0	244267.037	9889.5590
Kunín	4966	54918387.5	45555601.0	318727.949	11058.8779
Kačina	6802	45499810.0	62398147.0	346485.613	6689.1811
Křížanovsko	2357	26579933.0	21621939.5	240037.545	11277.0187
Uherčice	1357	14709070.5	12448439.5	187742.487	10839.4035
Sedlec	408	2301073.0	3742788.0	105780.551	5639.8848
Average scores were used for ties.					

Kruskal-Wallis Test	
Chi-Square	2858.6387
DF	5
Asymptotic Pr > Chi-Square	<.0001
Exact Pr >= Chi-Square	.

Získaná p-hodnota je mnohem menší než hladina významnosti 0,05, tedy nulovou hypotézu zamítáme ve prospěch alternativy, že distribuční funkce jednotlivých výběrů si nejsou rovny.

V tomto případě se tedy nabízí zjistit, které dvojice se od sebe významně liší. Pomocí Schéffeho metody se pokusíme porovnat střední hodnoty dojivosti jednotlivých výběrů.

Alpha	0.05
Error Degrees of Freedom	18340
Error Mean Square	78.41896
Critical Value of F	2.21459

Comparisons significant at the 0.05 level are indicated by ***.				
Kravín Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Křížanovsko - Kunín	0.1268	-0.6103	0.8638	
Křížanovsko - Uherčice	0.8949	-0.1092	1.8991	
Křížanovsko - Hradecká a.s.	2.0738	1.2242	2.9235	***
Křížanovsko - Kačina	7.8304	7.1261	8.5348	***
Křížanovsko - Sedlec	10.2629	8.6828	11.8430	***
Kunín - Křížanovsko	-0.1268	-0.8638	0.6103	
Kunín - Uherčice	0.7682	-0.1345	1.6708	
Kunín - Hradecká a.s.	1.9471	1.2202	2.6740	***
Kunín - Kačina	7.7037	7.1537	8.2537	***
Kunín - Sedlec	10.1361	8.6185	11.6537	***
Uherčice - Křížanovsko	-0.8949	-1.8991	0.1092	
Uherčice - Kunín	-0.7682	-1.6708	0.1345	
Uherčice - Hradecká a.s.	1.1789	0.1822	2.1756	***
Uherčice - Kačina	6.9355	6.0594	7.8116	***
Uherčice - Sedlec	9.3680	7.7042	11.0317	***
Hradecká a.s. - Křížanovsko	-2.0738	-2.9235	-1.2242	***
Hradecká a.s. - Kunín	-1.9471	-2.6740	-1.2202	***
Hradecká a.s. - Uherčice	-1.1789	-2.1756	-0.1822	***
Hradecká a.s. - Kačina	5.7566	5.0629	6.4503	***
Hradecká a.s. - Sedlec	8.1891	6.6137	9.7644	***

Comparisons significant at the 0.05 level are indicated by ***.				
Kravin Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Kačina - Křížanovsko	-7.8304	-8.5348	-7.1261	***
Kačina - Kunín	-7.7037	-8.2537	-7.1537	***
Kačina - Uherčice	-6.9355	-7.8116	-6.0594	***
Kačina - Hradecká a.s.	-5.7566	-6.4503	-5.0629	***
Kačina - Sedlec	2.4325	0.9305	3.9344	***
Sedlec - Křížanovsko	-10.2629	-11.8430	-8.6828	***
Sedlec - Kunín	-10.1361	-11.6537	-8.6185	***
Sedlec - Uherčice	-9.3680	-11.0317	-7.7042	***
Sedlec - Hradecká a.s.	-8.1891	-9.7644	-6.6137	***
Sedlec - Kačina	-2.4325	-3.9344	-0.9305	***

Ze získaných tabulek tudíž vyplývá, že střední hodnoty dojivosti se signifikantně neliší mezi kravíny Křížanovsko, Kunín a Uherčice. Ostatní dvojice distribučních funkcí se již od sebe významně liší.

3.2.4. Test šikmosti a špičatosti

Normalitu dat bychom také mohli otestovat pomocí šikmosti a špičatosti [3].

Jestliže platí hypotéza, že $\xi_1, \ldots, \xi_n$ je výběr z normálního rozdělení, pak A_3 a A_4 mají asymptoticky normální rozdělení s parametry

$$E(A_3) = 0,$$

$$\operatorname{var}(A_3) = \frac{6(n-2)}{(n+1)(n+3)},$$

$$E(A_4) = 3 - \frac{6}{(n+1)},$$

$$\operatorname{var}(A_4) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}.$$

Zároveň platí, že A_3 a A_4 jsou asymptoticky nekorelované.

Využít asymptotické normality však můžeme pouze pro velká n (v praxi pro $n \geq 200$). Vypočteme veličinu

$$U_3 = \frac{A_3}{\sqrt{var(A_3)}}.$$

Je-li $|U_3| \geq u(\alpha/2)$, zamítáme hypotézu, že jde o výběr z normálního rozdělení. Test proti alternativám s odlišnou špičatostí je založen na A_4. Vypočteme

$$U_4 = \frac{A_4 - E(A_4)}{\sqrt{var(A_4)}}.$$

Hypotézu zamítáme v případě, že $|U_4| \geq u(\alpha/2)$.

Výpočtem získáme pro tato data následující výběrové charakteristiky

moment/kravín	Hradecká	Kačina	Kunín	Křižanovsko
n	2456	6802	4966	2357
$\overline{X}$	27,9339	22,1773	29,8810	30,0078
S_X	10,2629	7,0346	10,5927	8,5414
šikmost A_3	0,1719	−0,0836	−0,1597	−0,1214
var(A_3)	0,0024	0,0009	0,0012	0,0025
U_3	3,4811	−2,8155	−4,5979	−2,4090
špičatost A_4	−0,3202	−0,1139	−0,0454	0,3118
var(A_4)	0,0097	0,0035	0,0048	0,0101
$E(A_4)$	2,9976	2,9991	2,9988	2,9975
U_4	−33,6651	−52,4657	−43,8551	−26,6995

moment/kravín	Sedlec	Uherčice
n	408	1357
$\overline{X}$	19,7449	29,1128
S_X	8,4064	7,9326
šikmost A_3	−0,1388	−0,1030
var(A_3)	0,0145	0,0044
U_3	−1,1528	−1,5527
špičatost A_4	−0,7473731	0,6726
var(A_4)	0,0567	0,0175
$E(A_4)$	2,9853	2,9956
U_4	−15,6753	−17,5643

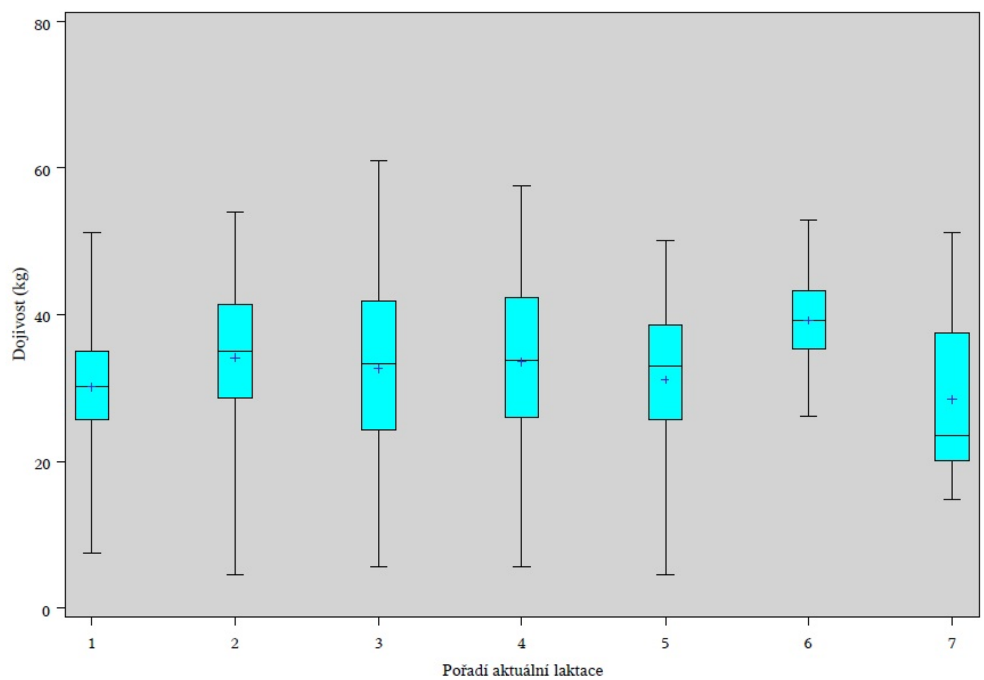
Dle tabulek je kritická hodnota $u(\alpha/2) = 1{,}96$. Na základě testu špičatosti zamítáme nulovou hypotézu, že data pocházejí z normálního rozdělení, pro všechny testované kravíny.

Testem šikmosti nulovou hypotézu zamítáme pro kravíny Hradecká, Kačina, Kunín a Křížanovsko, u nichž jsou testovací statistiky značně vyšší než kritická hodnota. Naopak pro kravíny Sedlec a Uherčice hypotézu o normalitě dat nezamítáme.

4. Statistická analýza dojivosti v kravíně Kunín

Dojivost můžeme sledovat s přihlédnutím k různým faktorům. Vzhledem k největšímu počtu dat jsem si vybrala Kunín a budu analyzovat dojivost s ohledem na faktor laktačního dne pomocí obdobných statistických metod jako v předchozí části.

Nejprve si opět vykreslíme krabicový diagram závislosti dojivosti na laktačním cyklu.



Obrázek č. 9: Variabilita dojivosti v různých laktačních cyklech.

Z obrázku se nabízí možnost otestovat shodnost středních hodnot v jednotlivých laktačních cyklech pomocí analýzy rozptylu jednoduchého třídění. Jelikož podmínkou je homoskedascidita dat, použijeme stejně jako v předchozí kapitole Bartlettův test.

4.1. Bartlettův test

V našem případě dostáváme:

Level of poradi_aktualni_laktace	N	dojivost	
		Mean	Std Dev
1	639	30.1092332	6.9218539
2	496	34.1401210	9.5027260
3	284	32.6172535	11.7809731
4	135	33.5881481	12.1144052
5	57	31.1368421	10.5823916
6	12	39.2333333	6.7653305
7	12	28.4416667	12.6162563

Bartlett's Test for Homogeneity of dojivost Variance			
Source	DF	Chi-Square	Pr > ChiSq
poradi_aktualni_lakt	6	156.9	<.0001

P-hodnota je signifikantně menší než hladina významnosti $\alpha = 0{,}05$, tudíž nulovou hypotézu o shodnosti rozptylů zamítáme.

4.2. Kruskal–Wallisův test

Po provedení Bartlettova testu ani v tomto případě nemůžeme sestavit ANOVU. Data znovu otestujeme pomocí neparametrického Kruskal–Wallisova testu.

Wilcoxon Scores (Rank Sums) for Variable Dojivost (kg) Classified by Variable Kravín					
Kravín	**N**	**Sum of Scores**	**Expected Under H0**	**Std Dev Under H0**	**Mean Score**
Hradecká a.s.	2456	24288757.0	22530116.0	244267.037	9889.5590
Kunín	4966	54918387.5	45555601.0	318727.949	11058.8779
Kačina	6802	45499810.0	62398147.0	346485.613	6689.1811
Křížanovsko	2357	26579933.0	21621939.5	240037.545	11277.0187
Uherčice	1357	14709070.5	12448439.5	187742.487	10839.4035
Sedlec	408	2301073.0	3742788.0	105780.551	5639.8848
Average scores were used for ties.					

Kruskal-Wallis Test	
Chi-Square	2858.6387
DF	5
Asymptotic Pr > Chi-Square	<.0001
Exact Pr >= Chi-Square	.

Nulovou hypotézu zamítáme na hladině významnosti $\alpha = 0{,}05$ oproti alternativě, že distribuční funkce dojivosti pro jednotlivé výběry jsou shodné.

I v tomto případě můžeme prostřednictvím Schéffeho metody pro mnohonásobné porovnávání zjistit, které konkrétní výběry se od sebe významně liší.

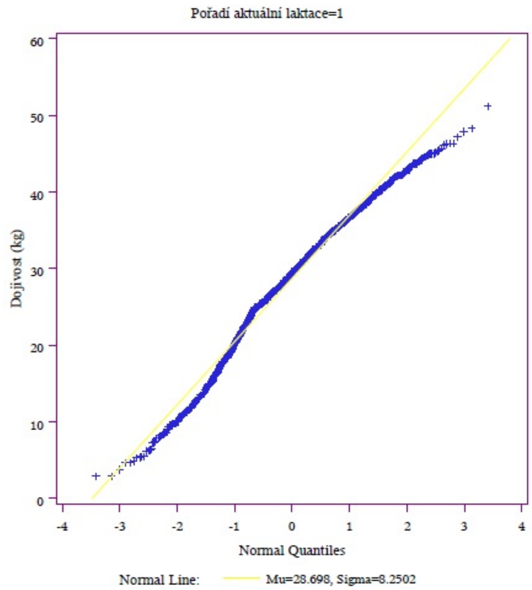
Comparisons significant at the 0.05 level are indicated by ***.				
poradi_aktualni_laktace Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
6 - 2	5.0932	-4.6265	14.8129	
6 - 4	5.6452	-4.3768	15.6672	
6 - 3	6.6161	-3.1890	16.4211	
6 - 5	8.0965	-2.4705	18.6635	
6 - 1	9.1241	-0.5699	18.8181	
6 - 7	10.7917	-2.7908	24.3741	
2 - 6	-5.0932	-14.8129	4.6265	
2 - 4	0.5520	-2.6777	3.7817	
2 - 3	1.5229	-0.9528	3.9986	
2 - 5	3.0033	-1.6498	7.6563	
2 - 1	4.0309	2.0399	6.0218	***
2 - 7	5.6985	-4.0213	15.4182	
4 - 6	-5.6452	-15.6672	4.3768	
4 - 2	-0.5520	-3.7817	2.6777	
4 - 3	0.9709	-2.5071	4.4489	
4 - 5	2.4513	-2.8040	7.7066	
4 - 1	3.4789	0.3275	6.6303	***
4 - 7	5.1465	-4.8755	15.1685	
3 - 6	-6.6161	-16.4211	3.1890	

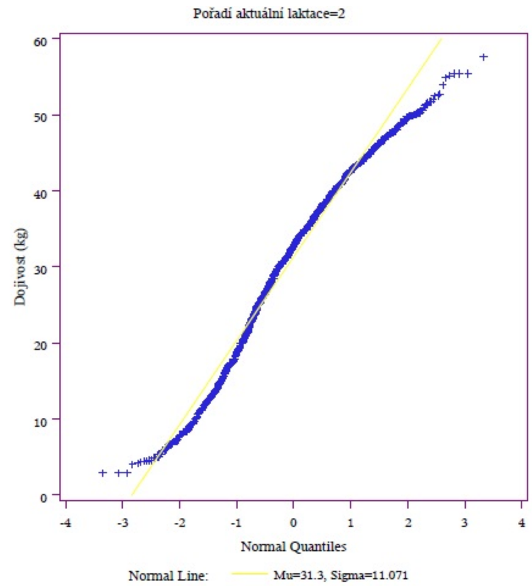
Comparisons significant at the 0.05 level are indicated by ***.				
poradi_aktualni_laktace Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
3 - 2	-1.5229	-3.9986	0.9528	
3 - 4	-0.9709	-4.4489	2.5071	
3 - 5	1.4804	-3.3483	6.3092	
3 - 1	2.5080	0.1353	4.8807	***
3 - 7	4.1756	-5.6295	13.9806	
5 - 6	-8.0965	-18.6635	2.4705	
5 - 2	-3.0033	-7.6563	1.6498	
5 - 4	-2.4513	-7.7066	2.8040	
5 - 3	-1.4804	-6.3092	3.3483	
5 - 1	1.0276	-3.5715	5.6267	
5 - 7	2.6952	-7.8718	13.2621	
1 - 6	-9.1241	-18.8181	0.5699	
1 - 2	-4.0309	-6.0218	-2.0399	***
1 - 4	-3.4789	-6.6303	-0.3275	***
1 - 3	-2.5080	-4.8807	-0.1353	***
1 - 5	-1.0276	-5.6267	3.5715	
1 - 7	1.6676	-8.0264	11.3616	
7 - 6	-10.7917	-24.3741	2.7908	
7 - 2	-5.6985	-15.4182	4.0213	
7 - 4	-5.1465	-15.1685	4.8755	
7 - 3	-4.1756	-13.9806	5.6295	
7 - 5	-2.6952	-13.2621	7.8718	
7 - 1	-1.6676	-11.3616	8.0264	

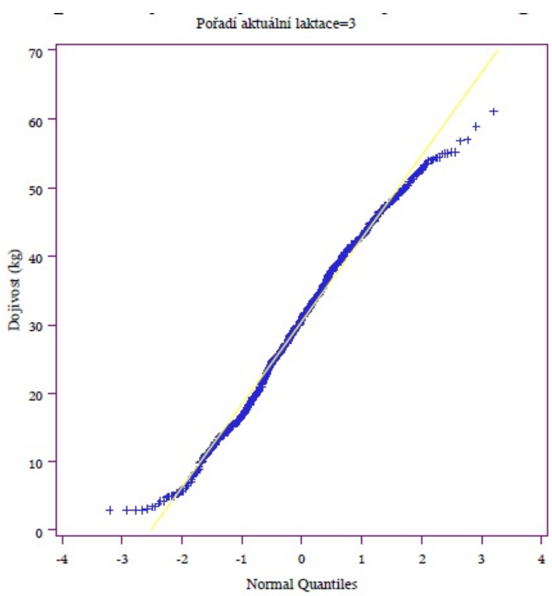
Signifikantně se od sebe liší dvojice výběrů laktačních cyklů 1 a 2, 1 a 3, 1 a 4.

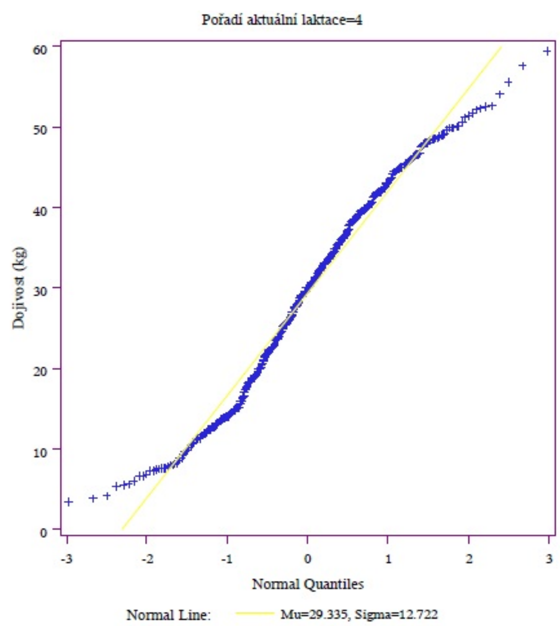
4.3. Posouzení normality laktačních cyklů

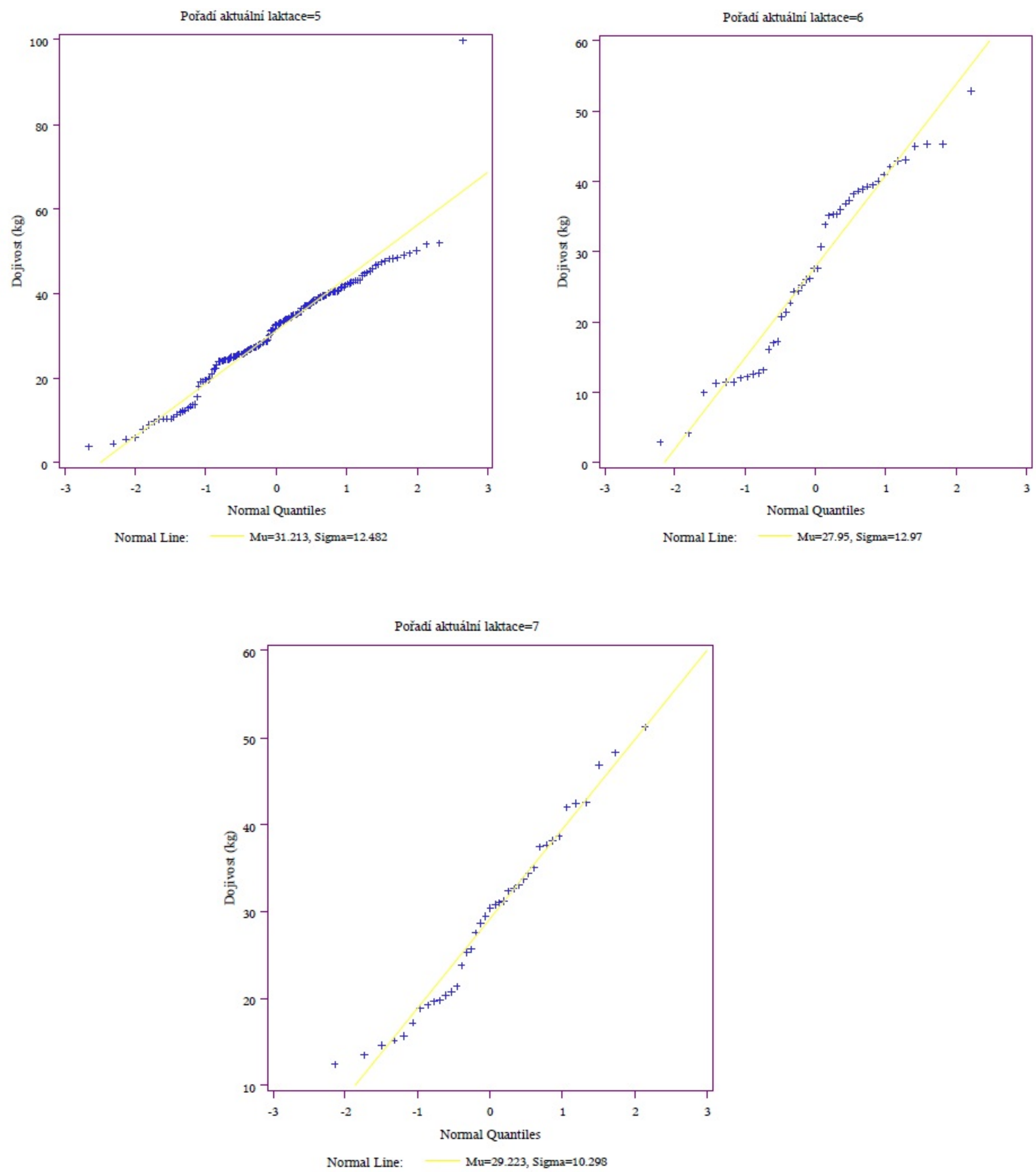
Zda data pocházejí z normálního rozdělení můžeme posoudit při sestrojení QQ plotů pro dané laktační cykly.


Pořadí aktuální laktace=1
Dojivost (kg)
Normal Quantiles
Normal Line:
Mu=28.698, Sigma=8.2502


Pořadí aktuální laktace=2
Dojivost (kg)
Normal Quantiles
Normal Line:
Mu=31.3, Sigma=11.071


Pořadí aktuální laktace=3
Dojivost (kg)
Normal Quantiles


Pořadí aktuální laktace=4
Dojivost (kg)
Normal Quantiles
Normal Line:
Mu=29.335, Sigma=12.722



Obrázek č. 10: Q-Q ploty dojivosti pro jednotlivé laktační cykly.

5. Regresní analýza

5.1. Kubická regrese

Pojmem kubická regrese rozumíme model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + e_i,$$

kde $i = 1, \dots, n$ a $e_i \sim N(0, \sigma^2)$.

Zde máme

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^2 \\ & \dots & & \\ 1 & x_n & x_n^2 & x_n^3 \end{pmatrix},$$

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \sum x_i^5 \\ \sum x_i^3 & \sum x_i^4 & \sum x_i^5 & \sum x_i^6 \end{pmatrix},$$

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \\ \sum x_i^2 Y_i \\ \sum x_i^3 Y_i \end{pmatrix}.$$

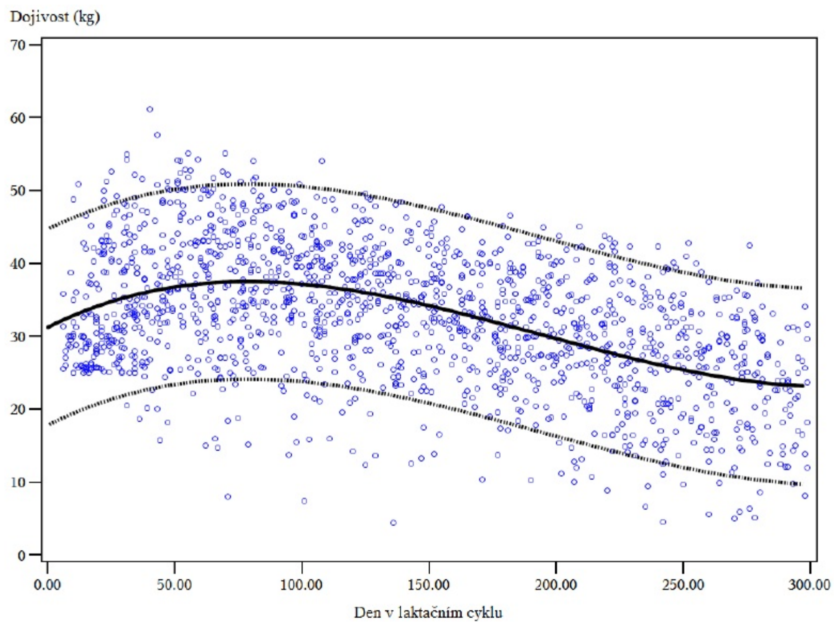
Odhad $\mathbf{b} = (b_0, b_1, b_2, b_3)'$ vektoru $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ vypočítáme řešením soustavy rovnic

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}.$$

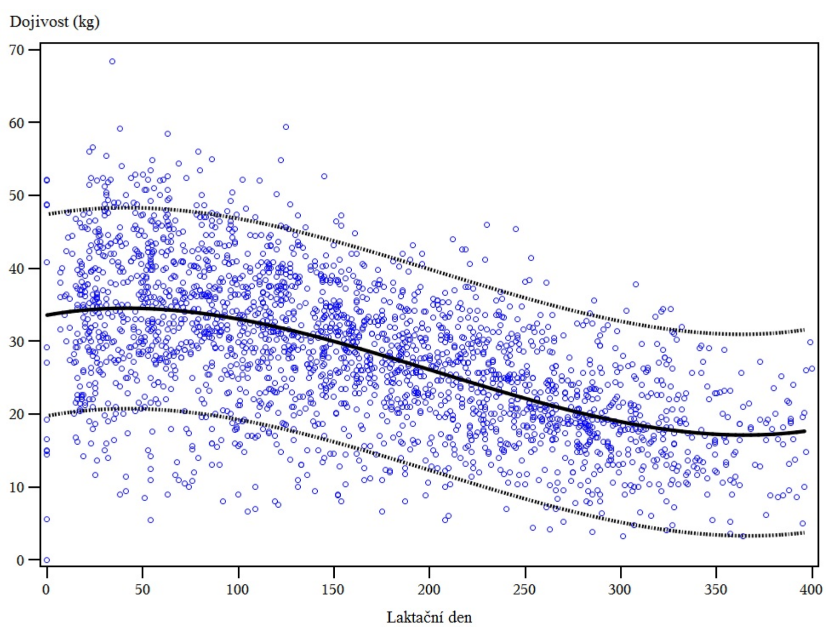
Poté určíme

$$s^2 = \frac{1}{n-3}\left(\sum Y_i^2 - b_0 \sum Y_i - b_1 \sum x_i Y_i - b_2 \sum x_i^2 Y_i - b_3 \sum x_i^3 Y_i\right).$$

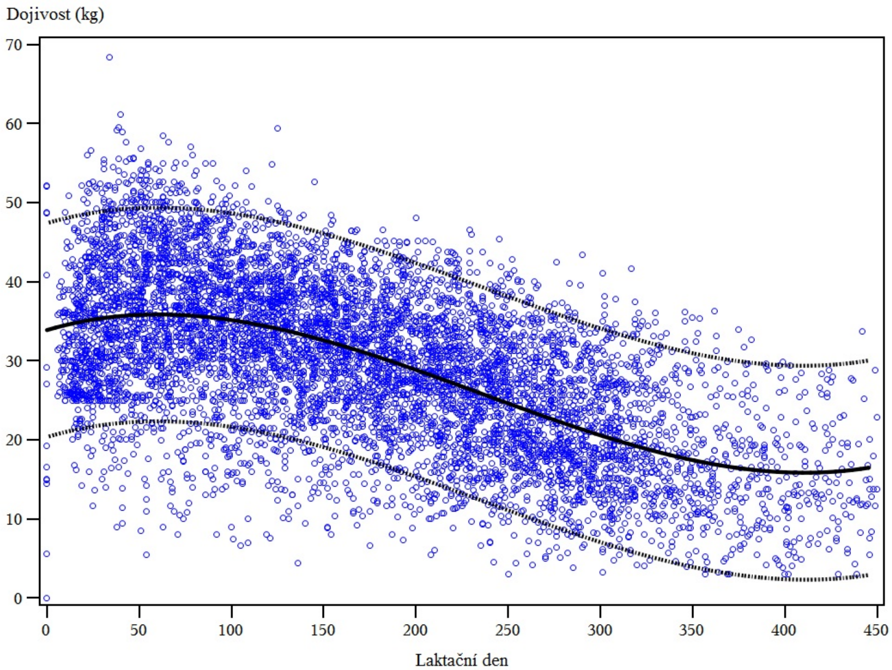
V programu SAS Enterprise Guide jsme si nechali vykreslit graf závislosti dojivosti na dnu laktace pro kravíny v Hradečné a Kuníně.



Obrázek č. 11: Laktační křivka pro kravín Kunín.



Obrázek č. 12: Laktační křivka pro kravín Hradecká.



Obrázek č. 13: Laktační křivka pro kravíny Hradecká a Kunín.

6. Analýza přežívání dojnic

6.1. Údaje o vyřazování dojnic

S využitím poznatků z pojistné matematiky a praktických znalostí se statistickým modulem SAS Enterprise Guide se pokusím v této kapitole věnovat analýze přežívání dojnic. K jejímu provedení jsem měla k dispozici data týkající se vyřazování dojnic z kravína. Data nám poskytují informační údaje o dojnici (její identifikační číslo, původ, stáj), datum jejího narození, datum a důvod vyřazení dojnice z kravína.

Ukázku dat prezentuje následující tabulka.

kráva	kód	původ	stáj	datum narození	datum vyřazení	důvod vyřazení
36257	544	H100	100062	01.12.1996	02.11.2006	jatky
44307	265	H63 C37	100062	26.11.1998	07.02.2006	zánět vemene
77654	931	H100	100062	27.03.2004	09.07.2007	zánět čelisti
101865	205	C50 H50	100062	10.10.1999	20.03.2007	nevstávala po otelení
125050	205	H81 C19	200021	20.09.2002	07.11.2008	chronická bronchitida
181917	931	H100	100062	27.02.2005	25.07.2008	ochrnutí zadních končetin

6.2. Analýza přežívání

Analýza přežívání se používá k popisu dat, která odpovídají času od vstupní události do výskytu nějaké sledované události, a posléze tato data analyzuje. Za vstupní událost můžeme například považovat datum narození jedince, začátek onemocnění či léčby nebo zavedení nového přístroje do výroby. Koncovou událostí může být úmrtí jedince, uzdravení pacienta či porucha přístroje. Dobu mezi vstupní a koncovou událostí označujeme jako dobu přežití.

V našem případě budeme za vstupní událost považovat datum narození dojnice a za koncovou údálost datum jejího vyřazení z kravína.

6.2.1. Distribuční funkce

Na aktuální dobu přežití dojnice může pohlížet jako na hodnotu proměnné T, jenž může nabývat pouze nezáporných hodnot. T je tedy nezáporná náhodná veličina udávající čas, který uplynul od narození dojnice. Pravděpodobnostní rozdělení náhodné veličiny T analýza přežívání popisuje pomocí distribuční funkce

$$F(t) = P(T \leq t).$$

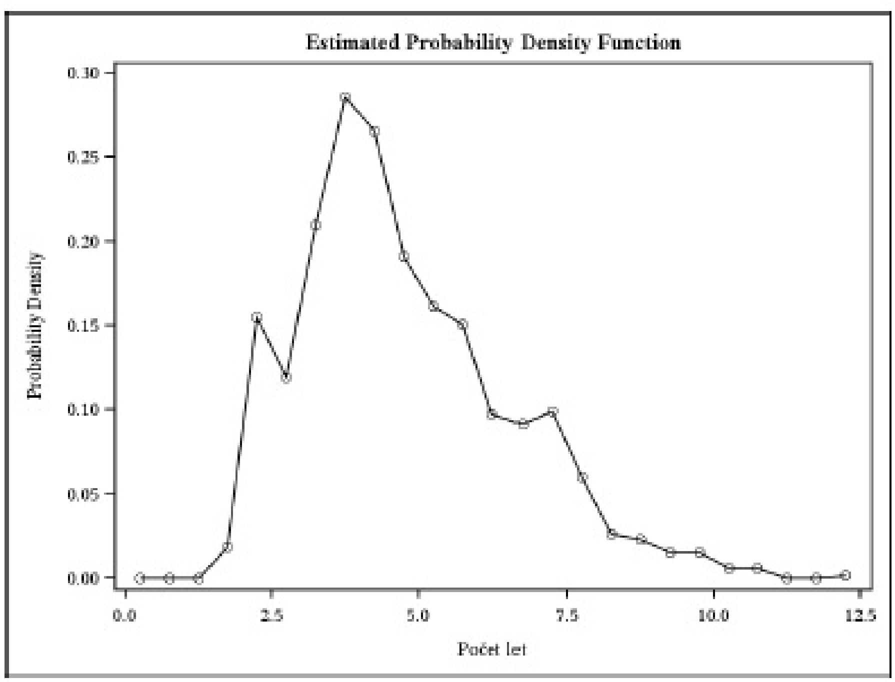
Udává nám pravděpodobnost, že doba přežití dojnice je menší nebo rovna hodnotě t, rovněž také pravděpodobnost, že dojnice již v čase t nežije.

6.2.2. Hustota pravděpodobnosti

U náhodné veličiny T se spojitým rozdělením můžeme distribuční funkci vyjádřit pomocí hustoty pravděpodobnosti $f(t)$ takto

$$F(t) = \int_0^t f(u)du.$$

Hustota $f(t)$ musí být nezáporná borelovsky měřitelná funkce z $R \to R$, viz [5], str. 45.



Obrázek č. 14: Hustota pravděpodobnosti náhodné veličiny T vygenerovaná systémem SAS Enterprise Guide.

6.2.3. Funkce přežití

Dále se zavádí funkce přežití $S(t)$ vyjádřená vztahem

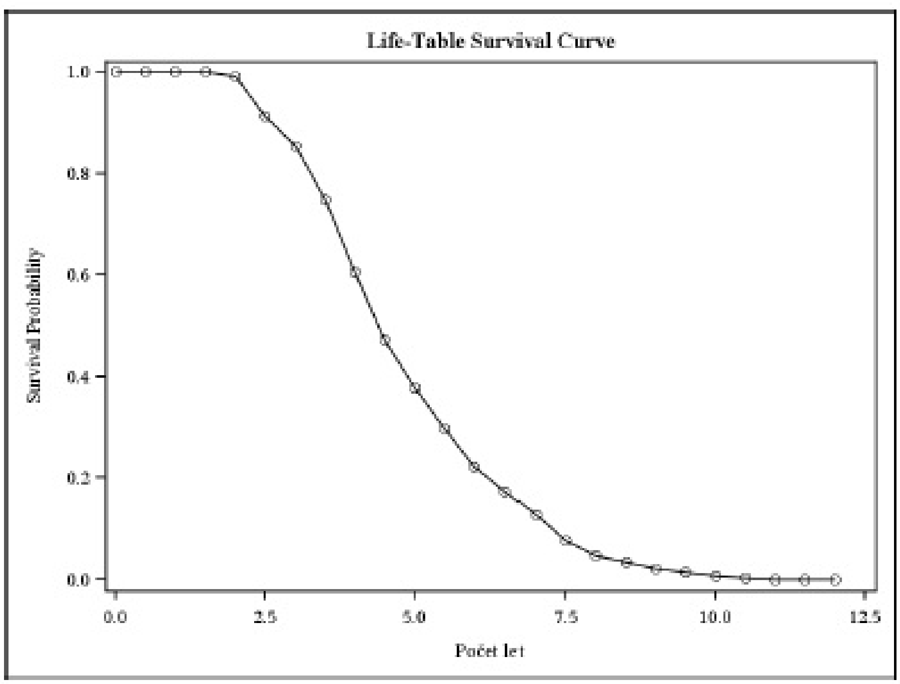
$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)du.$$

Reprezentuje pravděpodobnost, že doba přežití sledované dojnice bude větší než t, neboli značí situaci, že bude v čase t naživu.

Funkce přežití je nerostoucí funkcí, tudíž z vlastností hustoty a distribuční funkce pro ni platí:

$$S(0) = 1,$$

$$\lim_{t\to\infty} S(t) = 0.$$



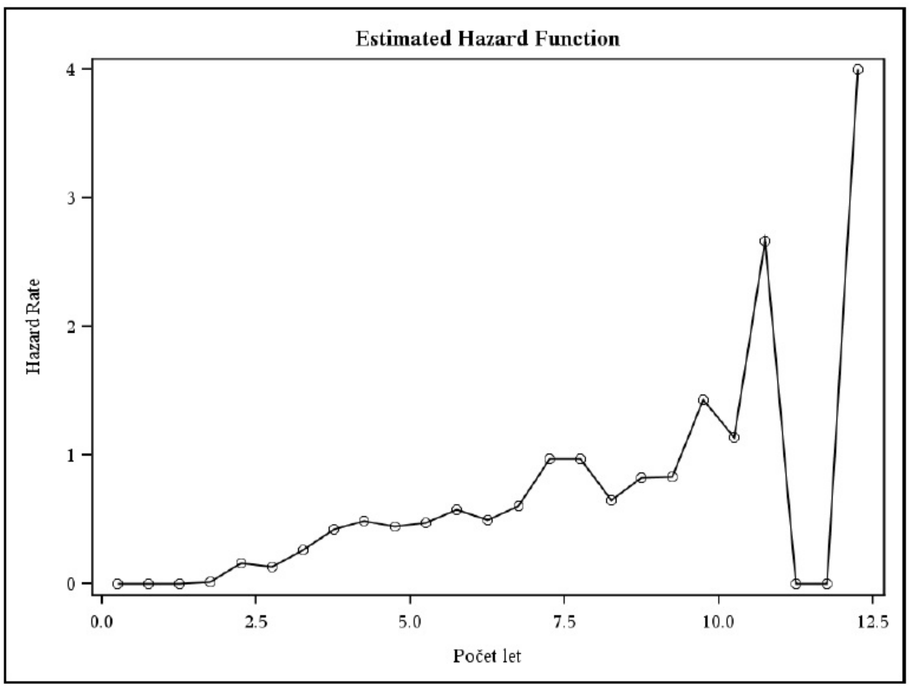
Obrázek č. 15: Funkce přežití dojnic vygenerovaná systémem SAS Enterprise Guide.

6.2.4. Riziková funkce

V analýze přežívání se také používá riziková funkce $h(t)$, často nazývána jako míra úmrtnosti. Je definována vztahem

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} = 0$$

a vyjadřuje pravděpodobnost, že dojnice zemře v čase t za podmínky, že do tohoto času přežila.



Obrázek č. 16: Riziková funkce vygenerovaná systémem SAS Enterprise Guide.

7. Závěr

V práci jsem se snažila pomocí vybraných statistických metod analyzovat dojivost krav ve dvou základních situacích – v závislosti na kravíně a na aktuálním laktačním cyklu.

Zjistila jsem, že dojivost v různých kravínech se značně liší. Na to má jistě vliv spousta okolních faktorů, neboť dojnice byly všechny stejného typu.

Nejvyšší dojivost byla v Kuníně, Křižanovsku a Uherčicích. Nejmenší dojivost měl kravín v Sedleci.

Ve všech kravínech se objevil obdobný průběh dojivosti, která je funkcí laktačního dne. Dojivost u dojnice od narození telete stoupá přibližně do 1/3 laktačního cyklu a poté začne klesat.

Dále jsem zjistila, že se liší dojivost krav v prvním laktačním cyklu a v následujících laktačních cyklech, kde je vyšší. Chovatelské důvody pro nezařazení dojnice k dalšímu chovu jsou nejspíše důvodem pro tento jev.

Vyřazování dojnic jsem zmapovala v poslední kapitole. Zkonstruovaná funkce přežití říká, že pouze třetina dojnic je v chovu zařazena dále po pátém roku svého věku.

Při zpracovávání práce jsem uplatnila své znalosti statistického softwaru SAS, jenž mi velice usnadnil veškeré výpočty, které jsem musela provést. Nejvíce si ovšem cením toho, že jsem se zjistila, jak používat teorii, se kterou jsem byla seznámena během studia na konkrétních aplikacích. V neposlední řadě jsem si osvojila základy práce v programu TeX.

Doufám, že úsilí věnované vypracování této práce bude přínosem pro mou budoucí praxi.

Literatura

[1] Zelinková, G., Marek, J.: Dependence of milk yield, fat: protein ratio and somatic cell counts. Proceedings of XI. Middle-European Congress, Brno, 2010.

[2] Zelinková, G.: Data ze sledování dojivosti v kravínech Hradecká, Křižanovsko, Kunín, Uherčice, Sedlec. 2010.

[3] Anděl, J.: Statistické metody. MATFYZPRESS, Praha, 2003.

[4] Pavlík, J.: Aplikovaná statistika. Vysoká škola chemicko-technologická v Praze, Praha, 2005.

[5] Kunderová, P.: Úvod do teorie pravděpodobnosti a matematické statistiky. UP, Olomouc, 2004.

[6] Cipra, T.: Pojistná matematika - teorie a praxe. EKOPRESS, s.r.o., Praha, 2006.

[7] Využití modulu SAS Enterprise Guide při statistických analýzách v agrárním sektoru
http://www.agris.cz/etc/textforwarder.php?iType=2&iId=137523&PHPSESSID=a3
[online 15. 3. 2010]

[8] SAS
http://www.sas.com[online 15. 3. 2010]