Vyhodnocování biologických dat pomocí statistických metod

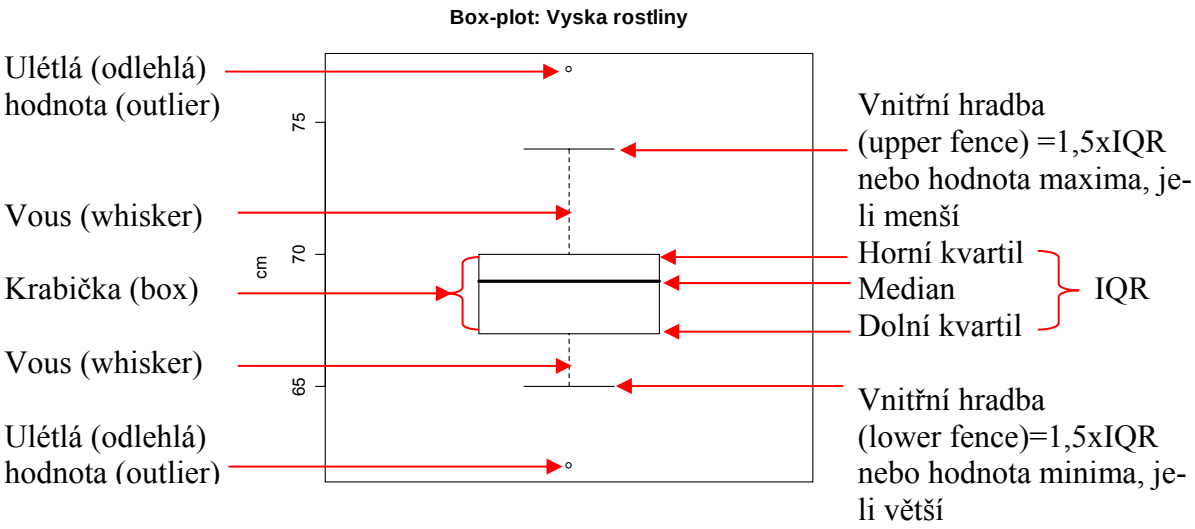
Eva Gelnarová

Úvod

Kdybyste se zeptali studentů nebo absolventů přírodovědecké fakulty, co je to **Statistika**, většina by si vzpomněla na písničku „Statistika nuda je…“a následně poznamenala, že jsou velice rádi, že se rozhodli pro studium botaniky (zoologie, chemie apod.), protože nikdy neměli matematiku a vzorečky v oblibě. A v zajímavém světě biologie či chemie jsou od matematiky a statistiky dostatečně daleko a v bezpečí. Není tomu tak docela pravda…
V současné době je trend, nejen v přírodních vědách ale i medicíně, sledovat velké množství subjektů nebo pacientů, a takto získané informace vyhodnotit statistickými metodami.
Tento text je zaměřený na aplikaci statistických metod, které jsou probírány v rámci středoškolské matematiky, na dva konkrétní datové soubory z oblasti biologie. Tyto základní statistické metody jsou rozšířeny o několik dalších pojmů a doplněny o odkazy na příslušnou literaturu a software. Doufáme, že uvedené příklady budou motivací pro sběr vlastních dat a jejich vyhodnocení.

Statistická metodika

Po provedení experimentu a naměření jsou data zapsaná do datového souboru ve formě obdélníkové **tabulky**. V této tabulce odpovídá každý řádek jednomu měřenému subjektu, např. pacientovi nebo rostlině. A každý sloupec odpovídá jedné sledované veličině, např. krevnímu tlaku pacienta nebo počtu listů rostliny. Při větším rozsahu dat, ale závislosti mezi veličinami nejsou na první pohled zřejmé (tabulky mohou obsahovat řádově desítky až stovky sloupců a řádků). Proto je prvním krokem při každé analýze dat jejich vizualizace-zviditelnění jedné veličiny (**histogram, krabičkový diagram**) nebo závislostí mezi více veličinami (**bodový graf**).
Než se pustíme do vysvětlení dalších statistických pojmů a metod, doporučujeme čtenářům tohoto textu nahlédnout do učebnice Calda,Dupač (1994), kde jsou vysvětleny základní statistické pojmy.
Krabičkový diagram (Box-plot): IQR =mezikvartilové rozpětí =Horní kvartil – Dolní kvartil



Aritmetický průměr, – je jednou z charakteristik polohy (viz. Calda,Dupač (1994)). Další charakteristikou polohy je **medián**, jehož výhodou je malá citlivost na odlehlá pozorování (outliers).

Korelační koeficient – charakterizuje závislost mezi dvěmi naměřenými veličinami, měří těsnost jejich vztahu. Korelační koeficient může nabývat hodnot od -1 do 1. Hodnota blízko 1 nebo -1 znamená, že veličiny jsou navzájem úzce spjaté (+1 - přímo úměrně, -1 - nepřímo úměrně). Hodnota kolem 0 pak indikuje, že mezi veličinami není žádná závislost.. Animace, které názorně ukazují vzájemný vztah veličin a příslušnou hodnotu korelačního koeficientu jsou dostupné na adrese:

http://broiler.stat.vt.edu/~sundar/java/applets/ (anglicky)

(Sekce: Statistical application , podsekce: Correlation)

Vzorce pro výpočet Pearsonova korelačního koeficientu lze nalézt v internetové encyklopedii Wikipedia

http://cs.wikipedia.org/wiki/Korelace

nebo na adrese

http://www.cba.muni.cz/vyuka/sources/biostaticky_seminar/korelace_regrese.pdf

Pro správnost Pearsonova korelačního koeficientu je nutné, aby byly splněny jisté předpoklady.

Regrese – popisuje závislost jedné kvantitativní veličiny (obvykle se značí *y*) na jedné (či více) kvantitativní veličině, která se obvykle značí *x*. Nejjednodušším případem je lineární regresní závislost, kdy závislost *y* na *x* lze vyjádřit přímkou. Názornou ukázku hledání regresní přímky (a dost prostoru pro vlastní pokusy) naleznete na adrese

http://broiler.stat.vt.edu/~sundar/java/applets/ (anglicky)

(Sekce: Statistical application , podsekce: Regression)

Zájemců o bližší seznámení s dalšími statistickými metodami doporučujeme populárně naučnou knížku Moderní statistika (Swoboda 1977), učební texty University Karlovy v Praze (Zvára (2001), Zvárová (2004)) nebo následující internetový odkaz:

http://botany.upol.cz/prezentace/duch

(soubory statistika1.pdf,…,statistika4.pdf)

Způsob zpracování dat

V případě malého množství dat je možné spočítat průměr a další charakteristiky ručně, dosazením naměřených hodnot do vzorců. Nicméně pro větší množství dat by byl tento postup příliš obtížný a pracný. Moderní statistika proto využívá pro analýzu dat počítače a existuje velké množství specializovaných statistických programů. Uveďme alespoň dva programy, které jsou volně přístupné. Prvním z jich je software R, který je možné stáhnout, zcela zdarma, na adrese:

http://www.r-project.org/ (anglicky)

Pomocí tohoto programu byly provedeny analýzy i v tomto výukovém textu. R je programovací jazyk, příkazy se píší na příkazový řádek ve formě kódu, je tedy vhodný pro ty s kladným vztahem k programování. Poněkud přístupnější je program NCSS. Na níže uvedené adrese je zdarma ke stažení jeho zjednodušená verze, NCSS junior:

http://www.ncss.com/download.html#NCSS%206.0%20Junior (anglicky)

Základní statistické funkce jsou implementovány také v EXCELU. Časově omezené demoverze statistických programů STATISTICA a SPSS jsou také volně dostupné na internetu.

Příklad: Studenti

Data: Na přednášce bylo osloveno 22 studentů, aby napsali svou váhu (kg), výšku (m) a obvod pasu (cm). Ke každému údaji navíc známe pohlaví studenta (8 mužů a 14 žen). Data jsou uvedena v apendixu, tabulka č. A.1.
Úkol: Na základě získaných údajů, bychom chtěli odhadnout průměrnou výšku a váhu studenta a vypočítat průměrné BMI (body mass index) a určit procento lidí s nadváhou. Na závěr bychom chtěli určit, zda existuje vzájemná souvislost mezi naměřenými veličinami.

Řešení: Vypočtené hodnoty aritmetických průměrů a mediánů jsou uvedeny v tabulce č.1. Aritmetický průměr výšky vyšel 9,497 metrů, což je naprosto nemyslný výsledek. Na obrázku č.1 vlevo je nakreslen histogram, který poukazuje na existenci člověka, jehož výška je mezi 150 a 200 metry. Jeden ze studentů se spletl a omylem uvedl svou výšku v centimetrech místo v metrech. Při analýze reálných dat je tedy nutné počítat s chybami v datech. Náš datový soubor byl velice malý, proto jsem chybu nalezli okamžitě, ale zpracovávané datové soubory mohou být mnohem větší.. V běžných datových souborech zůstává v průměru 1 až 10% chyb (Hampel (1986)). Data vznikají chybou měření, chybami při opisu dat (z papírového formuláře do počítače) nebo chybným naplánováním experimentu (př. do zkoumané populace určitého druhu malého ptáka se vmísil větší jedinec, který tam nepatří, kukaččí mládě).

Hodnota mediánu byla 1,685. Tento výsledek je mnohem realističtější. V případě podezření na přítomnost odlehlých hodnot nebo v nesymetrických datech je vhodné pro odhad polohy použít medián. Klasickým příkladem nesymetrických zešikměných dat jsou údaje o výšce platu. (http://www.mesec.cz/clanky/bude-vase-mzda-prumerna/)

Tabulka č. 1

	Výška (m) {po opravení}	Váha (kg)	Obvod pasu (cm)	$BMI = \frac{vaha}{vyska^2}$
Průměr – Muži	23,2 {1,79}	76,13	80,38	23,53
Medián – Muži	1,81 {1,79}	72	77	23,11
Průměr – Ženy	1,67	57,82	68,7	20,79
Medián - Ženy	1,67	59	68	21,01
Průměr – obě pohlaví	9,497 {1,71}	64,48	72,95	21,49
Medián – obě pohlaví	1,685	60,5	70	21,36

V další analýze budeme pracovat s opravenou hodnotou výšky.
Podívejme se na obrázek č.2, který zobrazuje histogramy a box-ploty pro měřené veličiny. Dívky jsou v průměru menší než muži, méně váží a mají útlejší pas. Zajímavé je porovnání variability dat v případě váhy a obvodu pasu. Data mužů mají mnohem větší mezikvartilové rozpětí („délka krabičky“) než data žen.
Na základě dat, která jsme vyhodnotili, nelze tvrdit, že obecně „průměrná česká žena má obvod pasu 68 cm“ nebo „muži jsou o 18,5 cm větší než ženy“. Při interpretaci statistik musíme být obezřetní, neboť údaje se vztahují jen na zkoumanou populaci studentů a výsledky nelze přímo zobecnit na širší populaci, např. na všechny muže a ženy v ČR.

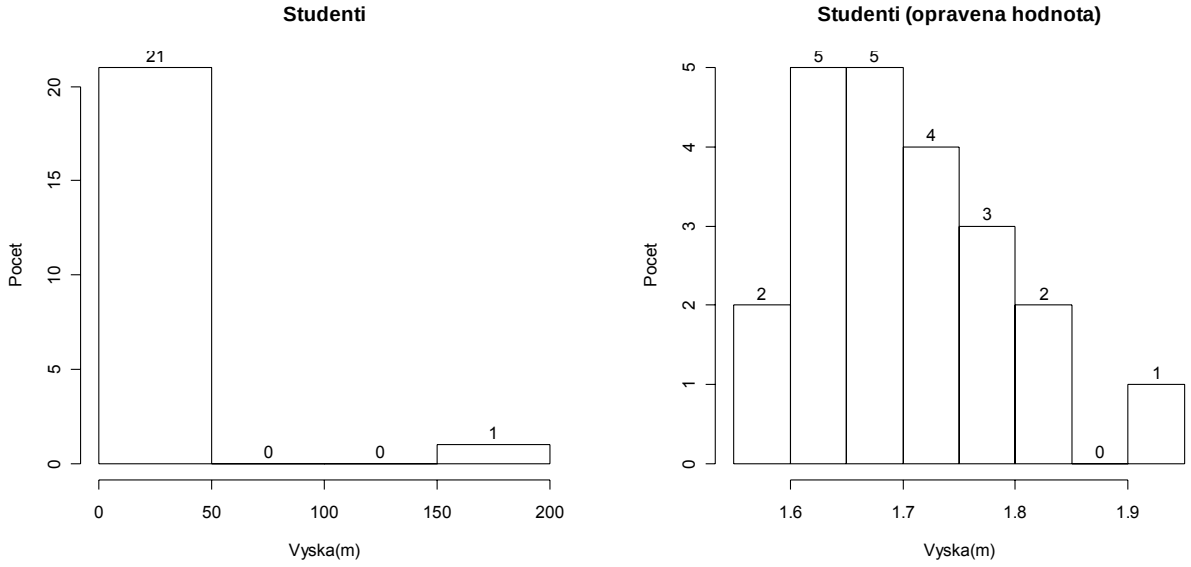
V posledním sloupci tabulky jsou uvedeny odhady polohy pro BMI. Průměrná hodnota BMI pro muže je 23,53, ale uvědomme si, že

$$23{,}53 \neq \frac{76{,}13}{1{,}79^2} = 23{,}76.$$

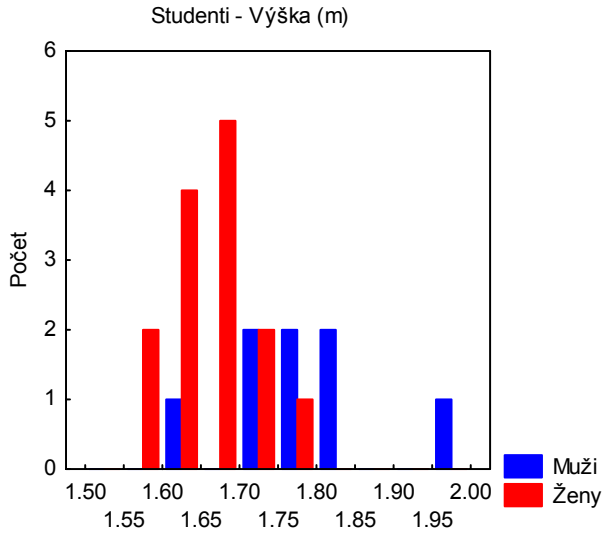
Průměrnou hodnotu BMI nelze vypočítat z průměrné výšky a váhy, ale je nutné vypočítat hodnotu BMI pro každého studenta individuálně a až z takto získaných hodnot spočítat průměr nebo medián.
Na základě BMI lze určit, zda student má normální váhu nebo trpí nadváhou. V tabulce č.2 jsou uvedeny hraniční hodnoty BMI a zároveň počet studentů v jednotlivých kategoriích. Zvýšenou váhu mají 3 studenti, což je 13.6% souboru. Téměř 70% studentů má normální váhu (12+3).

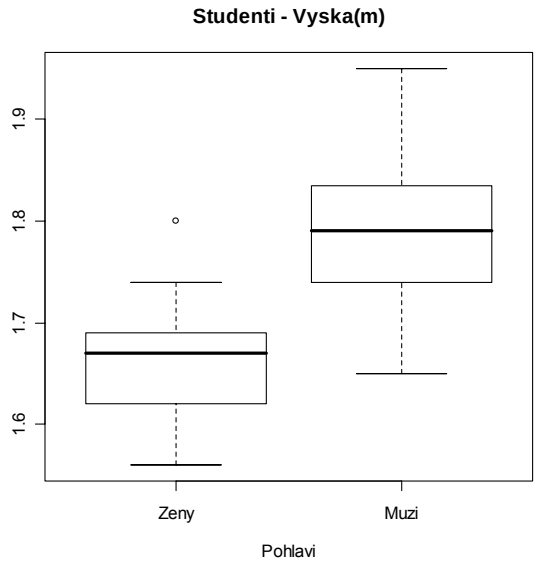
Tabulka č. 2

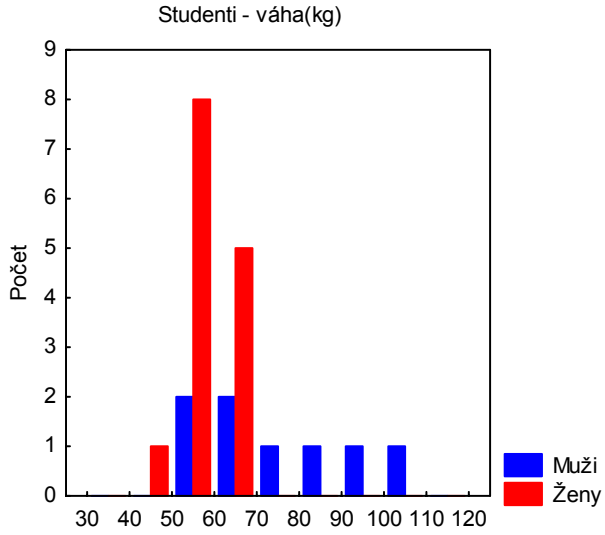
		Podváha	Normální váha	Nadváha	Obezita	Těžká obezita
Muži	BMI	BMI<20	20 - 24,9	25-29,9	30-39,9	BMI>40
	Počet	**2**	**3**	**2**	**1**	**0**
Ženy	BMI	BMI<19	19-23.9	24-28.9	29-38.9	BMI>39
	Počet	**2**	**12**	**0**	**0**	**0**

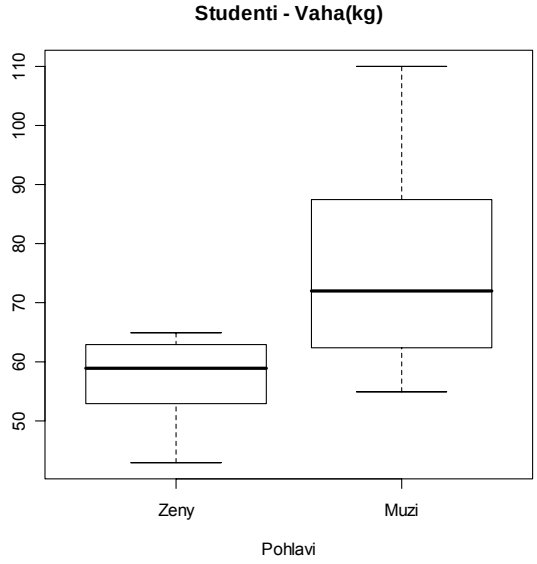


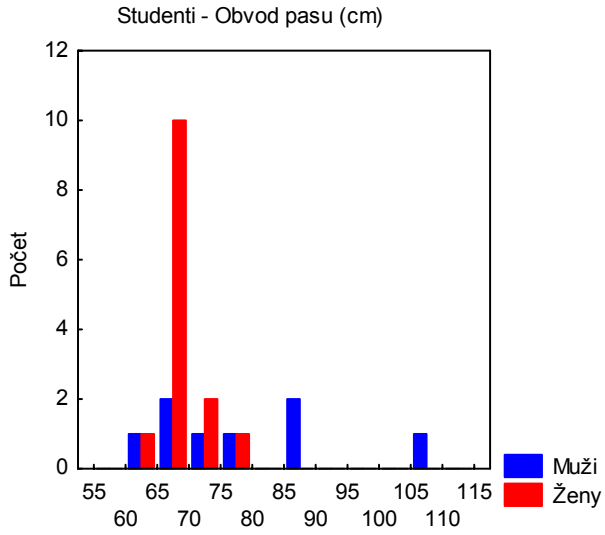
Obrázek č.1: Histogram, výška studentů uvedená v metrech.

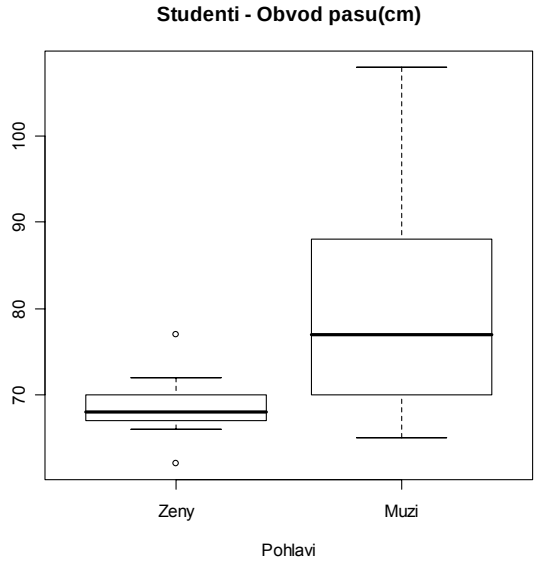

Studenti - Výška (m)
Počet
0
1
2
3
4
5
6
1.50
1.55
1.60
1.65
1.70
1.75
1.80
1.85
1.90
1.95
2.00
Muži
Ženy

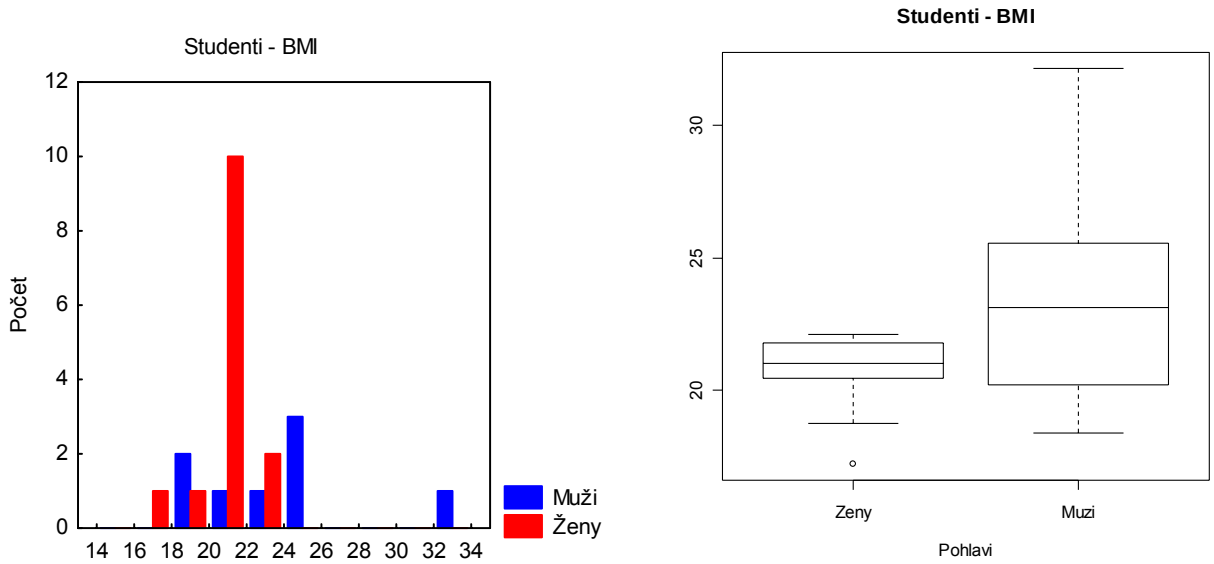

Studenti - Vyska(m)
1.6
1.7
1.8
1.9
Zeny
Muzi
Pohlavi


Studenti - váha(kg)
Počet
0
1
2
3
4
5
6
7
8
9
30
40
50
60
70
80
90
100
110
120
Muži
Ženy


Studenti - Vaha(kg)
50
60
70
80
90
100
110
Zeny
Muzi
Pohlavi


Studenti - Obvod pasu (cm)
Počet
0
2
4
6
8
10
12
55
60
65
70
75
80
85
90
95
100
105
110
115
Muži
Ženy


Studenti - Obvod pasu(cm)
70
80
90
100
Zeny
Muzi
Pohlavi



Obrázek č.2: Histogramy a box-ploty pro měřené veličiny

Předpokládejme, že vztah mezi výšku, váhou a obvodem pasu je stejný pro obě pohlaví. Data tedy budeme analyzovat pro obě pohlaví dohromady. Tabulce č.3 jsou uvedeny Pearsonovy korelační koeficienty. Všimněme si, že tabulka je symetrická. Jako nejtěsnější se jeví závislost váhy a obvodu pasu, korelační koeficient je roven 0,96. Na diagonále tabulky jsou hodnoty 1, protože každá veličina je ideálně korelovaná sama se sebou.

Tabulka č. 3: Pearsonův korelační koeficient

	Výška (m)	Váha (kg)	Obvod pasu (cm)
Výška (m)	1	0,80	0,69
Váha (kg)	0,80	1	0,96
Obvod pasu (cm)	0,69	0,96	1

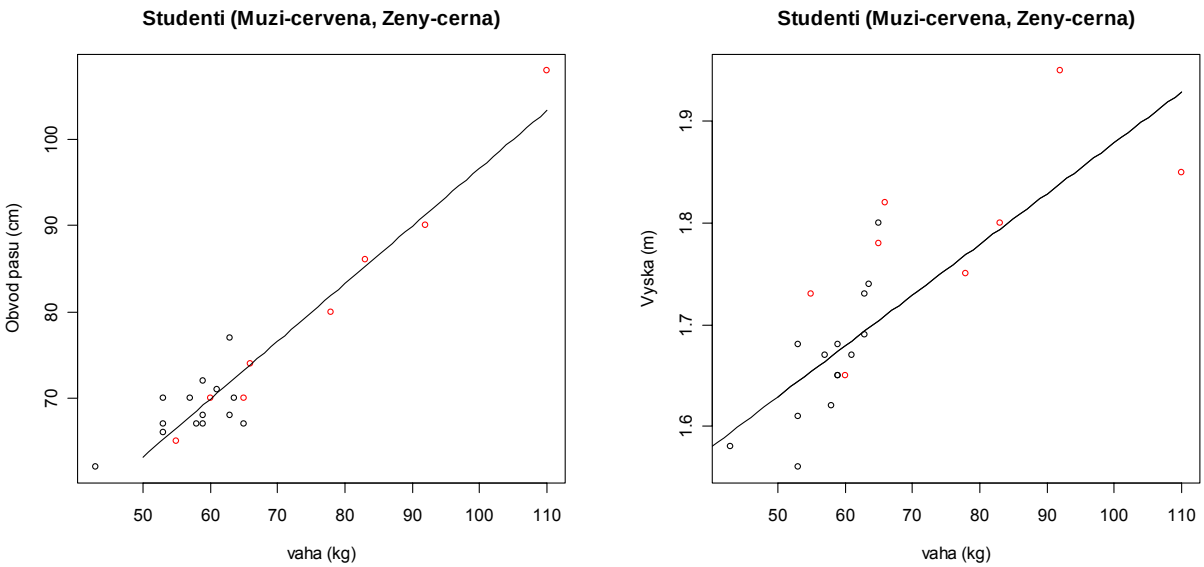
Na obrázku 3 jsou bodové grafy znázorňující vztah mezi výškou a váhou (vpravo) a váhou a obvodem pasu. Každý bod odpovídá jednomu studentovi. Vztahy mezi veličinami jsou lineární, body jsme proložili regresní přímky, které jsou tvaru

$$Vyska = 1.38 + 0.005\,Vaha$$

$$Obvod\ \ pasu = 29.67 + 0.671\,Vaha$$

Směrnice první přímky je 0,005. Je-li jeden student o 1kg těžší než druhý, pak by měl být také v průměru o 0,005m=0,5 cm větší.
Směrnice druhé přímky je 0,671. Je-li jeden student o 1kg těžší než druhý, pak by měl mít v pase v průměru o 0,671cm více.

Další náměty pro analýzu: Výše uvedená analýza samozřejmě není zcela kompletní. Bylo by zajímavé spočítat např. směrodatnou odchylku (viz. Calda,Dupač (1994)) nebo analyzovat souvislost výšky, váhy a obvodu pasu pro obě pohlaví odděleně.



Obrázek č.3: Bodový graf: závislost mezi vahou studenta a obvodem jeho pasu (vlevo) a závislost mezi vahou a výškou (vpravo).

Příklad: Květy kosatců

Bylo změřeno celkem 150 květů kosatců, které patří mezi 3 druhy: setosa (50 měřených květů), versicol (50 měřených květů) a virginica (50 měřených květů). U každého květu byla změřena délka a šířka kališního lístku a korunního plátku. Data jsou uvedena v apendixu, tabulka č. A.2. Data (soubor iris) jsou k dispozici také jako součást volně šířitelného statistického software R (http://www.r-project.org/).

Setosa

Versicolor

Virginica

Úkol: Na základě dat srovnejte tvary květů 3 druhů kosatců.

Řešení: Podívejme se nejdříve na krabičkové diagramy v obrázku č.4 a na bodové diagramy na obrázku č. 6. Nejmenší **korunní plátek** má druh setosa. V porovnání s versicolorem a virginicou je nejkratší a také nejužší. Naopak nejširší a nejdelší korunní plátky má virginica. Poměr šířky ku délce korunních plátků je zobrazen na obrázku č.5 (vpravo).

V případě **kališních lístků** je situace jiná. Opět nejkratší kališní lístek má setoza, tento kališní lístek je však nejširší. Z obrázku č. 5 (vlevo) je zřejmé, že versicolor a virginica mají stejný tvar kališních lístků (každý měřený květ má zhruba stejný poměr šířky a délky kališního lístku) a liší se od sebe pouze velikostí. Medián poměru mezi šířkou a délkou je u setozy 0.68. U versicolorem a virginicz je medián tohoto poměru mnohem menší (0,462;0.460). Čím blíže k hodnotě 1, tím kulatější jsou kališní lístky.

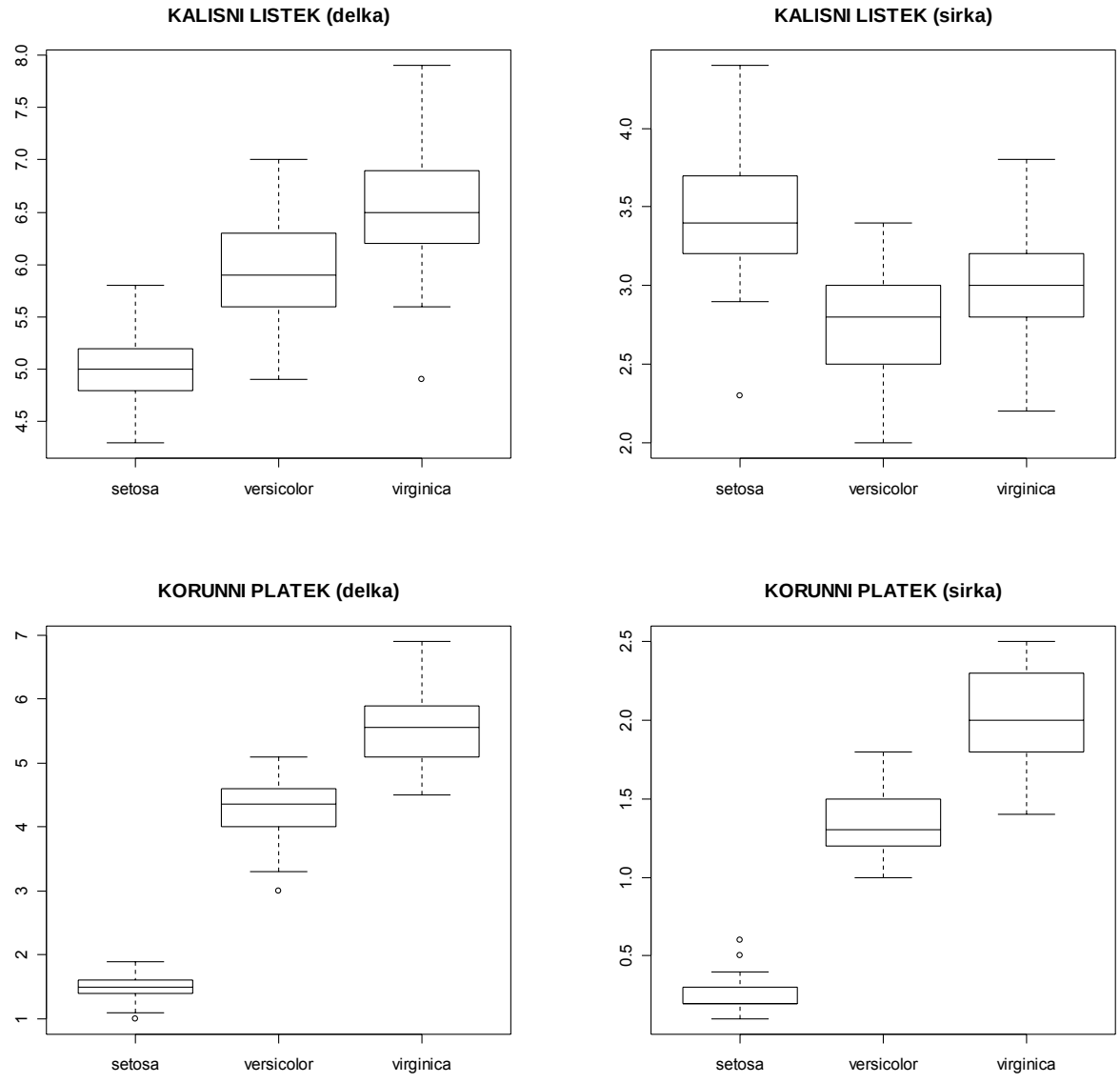
Na základě datového souboru máme zhruba představu, jak se chovají typičtí zástupci jednotlivých druhů kosatců. V případě nového nezařazeného měření je tedy možné odhadnout podle tvaru a velikosti, ke kterému druhu kosatců měřený květ patřil.
Na základě znalosti rozměrů korunních plátků lze sestavit určitou posloupnost pravidel, které pomohou k zařazení nového měření. Je-li délka korunního plátku menší než 2,1cm, pak květ paří kosatci druhu Setosa. Jeli korunní plátek květu delší než 2,1cm a zároveň není širší než 1,64, pak patří ke druhu Versicolor. Je-li korunní plátek květu delší než 2,1cm a zároveň širší než 1,64cm, pak se jedná o druh Virginica. Schematicky je tato posloupnost rozhodnutí zakreslena na obrázku č.7.

Tip na další analýzy: Tento příklad byl poněkud umělý, protože při novém měření je obvykle známo, ke kterému druhu kosatce květ patří. Na adrese

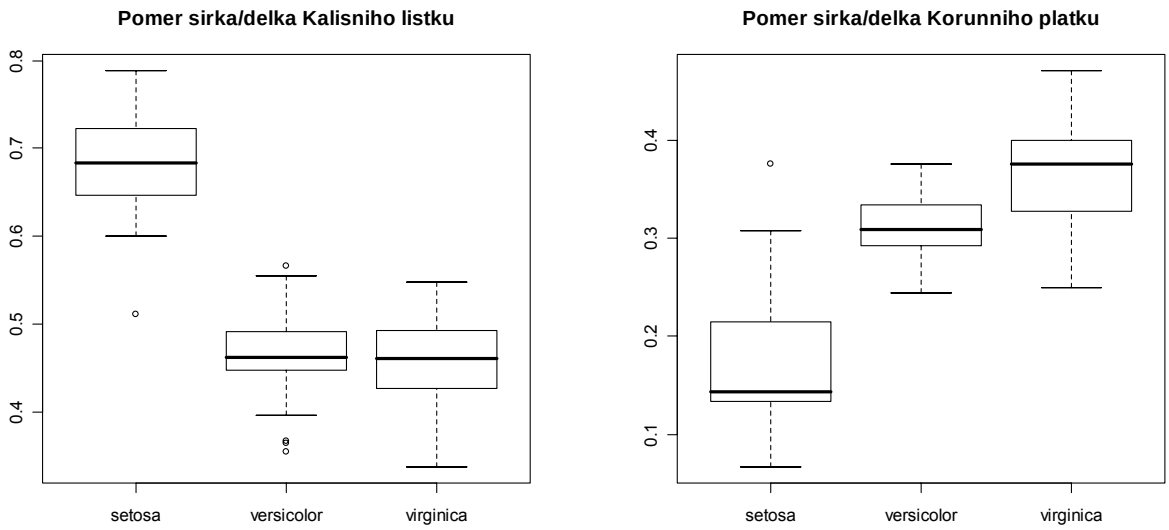
http://lib.stat.cmu.edu/DASL/Stories/EgyptianSkullDevelopment.html

http://lib.stat.cmu.edu/DASL/Datafiles/EgyptianSkulls.html

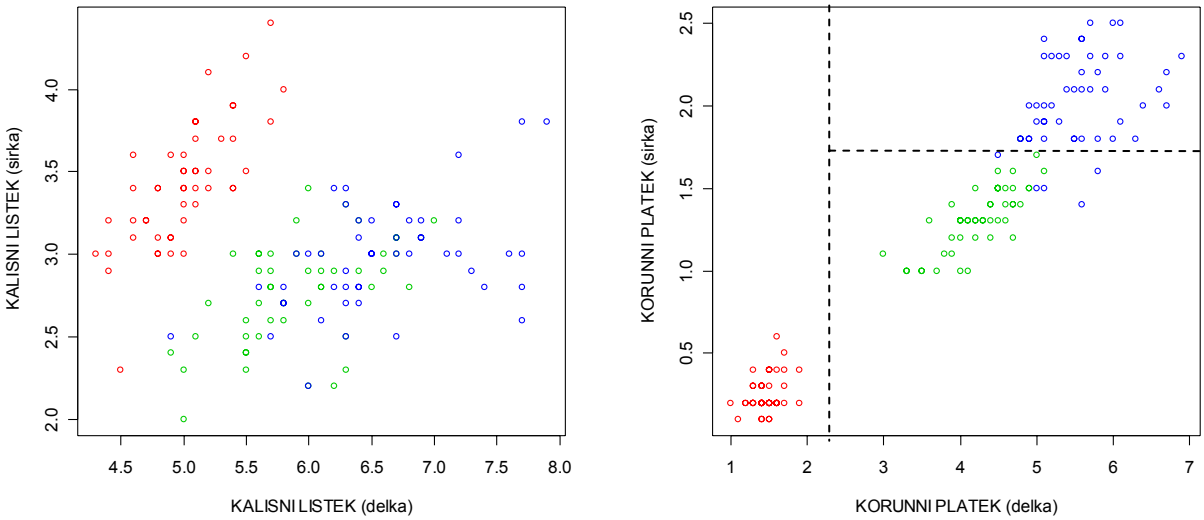
lze stáhnout datový soubor obsahující údaje o rozměrech lebek, které byly nalezeny při vykopávkách v Egyptě. Podle okolních nálezů bylo možné identifikovat do kterého období lebka patří, lebky jednotlivých období se mezi sebou lišily nejen velikostí, ale i tvarem. Při nálezu nové lebky je možno časově ji zařadit právě podle jejího charakteristického tvaru a rozměrů.



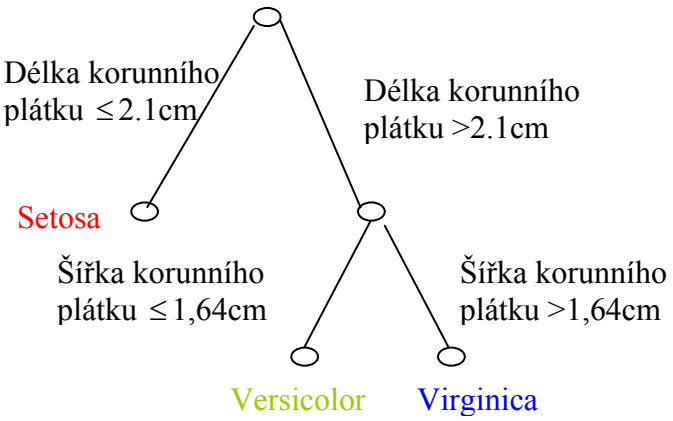
Příklad č.4: Box-ploty: Šířky a délky kališních lístků a korunních plátků.



Obrázek č.5: Box-ploty**:** Poměr šířky ku délce korunních plátků a kališních lístků



Obrazek č.6: Bodový graf: červená – setosa, zelená – versicolor, modrá – virginica.



Obrázek č.7: Schématické rozdělení kosatců podle velikosti korunního plátku. Odpovídá dělení datového souboru naznačeném přerušovanými čarami na obrázku č.6 vpravo.

Literatura a odkazy:

Calda E., Dupač V.(1994): Matematika pro gymnázia – Kombinatorika a pravděpodobnost. nakladatelství Prometheus
Hampel (1986): Robust statistics. Wiley-Interscience
Swoboda H. (1977): Moderní statistika. Svoboda, Praha.
Zvára K.(2001): Biostatistika. Nakladatelství Karolinum
Zvárová J.(2004): Základy statistiky pro biomedicínské obory: Biomedicínská statistika I.. Universita karlova v Praze

Appendix: Data

Tabulka č.A.1: Studenti, reálný datový soubor, studenti Masarykovy university v Brně, 2004

Pohlaví	Výška (m)	Váha (kg)	Obvod Pasu (cm)
M	1,85	110	108
Z	1,65	59	68
Z	1,69	63	77
M	1,95	92	90
Z	1,8	65	67
M	1,75	78	80
M	1,8	83	86
M	1,78	65	70
M	1,65	60	70
Z	1,62	58	67
M	1,82	66	74
Z	1,67	57	70
Z	1,58	43	62
Z	1,68	53	70
Z	1,67	61	71
Z	1,68	59	72
Z	1,74	63,5	70
Z	1,61	53	67
Z	1,65	59	67
Z	1,56	53	66
Z	1,73	63	68
M	173	55	65

Tabulka č.A.2: Květy kosatců

KLd – Kališní lístek, délka (Sepal Length)
KLš – Kališní lístek, šířka (Sepal Width)
KPd – Korunní plátek, délka (Petal Length)
KPš – Korunní plátek, šířka (Petal Width)

Setosa				Versicolor				Virginica			
KLd	KLš	KPd	KPš	KLd	KLš	KPd	KPš	KLd	KLš	KPd	KPš
5,1	3,5	1,4	0,2	7	3,2	4,7	1,4	6,3	3,3	6	2,5
4,9	3	1,4	0,2	6,4	3,2	4,5	1,5	5,8	2,7	5,1	1,9
4,7	3,2	1,3	0,2	6,9	3,1	4,9	1,5	7,1	3	5,9	2,1
4,6	3,1	1,5	0,2	5,5	2,3	4	1,3	6,3	2,9	5,6	1,8
5	3,6	1,4	0,2	6,5	2,8	4,6	1,5	6,5	3	5,8	2,2
5,4	3,9	1,7	0,4	5,7	2,8	4,5	1,3	7,6	3	6,6	2,1
4,6	3,4	1,4	0,3	6,3	3,3	4,7	1,6	4,9	2,5	4,5	1,7
5	3,4	1,5	0,2	4,9	2,4	3,3	1	7,3	2,9	6,3	1,8
4,4	2,9	1,4	0,2	6,6	2,9	4,6	1,3	6,7	2,5	5,8	1,8
4,9	3,1	1,5	0,1	5,2	2,7	3,9	1,4	7,2	3,6	6,1	2,5
5,4	3,7	1,5	0,2	5	2	3,5	1	6,5	3,2	5,1	2
4,8	3,4	1,6	0,2	5,9	3	4,2	1,5	6,4	2,7	5,3	1,9
4,8	3	1,4	0,1	6	2,2	4	1	6,8	3	5,5	2,1
4,3	3	1,1	0,1	6,1	2,9	4,7	1,4	5,7	2,5	5	2
5,8	4	1,2	0,2	5,6	2,9	3,6	1,3	5,8	2,8	5,1	2,4
5,7	4,4	1,5	0,4	6,7	3,1	4,4	1,4	6,4	3,2	5,3	2,3
5,4	3,9	1,3	0,4	5,6	3	4,5	1,5	6,5	3	5,5	1,8
5,1	3,5	1,4	0,3	5,8	2,7	4,1	1	7,7	3,8	6,7	2,2
5,7	3,8	1,7	0,3	6,2	2,2	4,5	1,5	7,7	2,6	6,9	2,3
5,1	3,8	1,5	0,3	5,6	2,5	3,9	1,1	6	2,2	5	1,5
5,4	3,4	1,7	0,2	5,9	3,2	4,8	1,8	6,9	3,2	5,7	2,3
5,1	3,7	1,5	0,4	6,1	2,8	4	1,3	5,6	2,8	4,9	2
4,6	3,6	1	0,2	6,3	2,5	4,9	1,5	7,7	2,8	6,7	2
5,1	3,3	1,7	0,5	6,1	2,8	4,7	1,2	6,3	2,7	4,9	1,8
4,8	3,4	1,9	0,2	6,4	2,9	4,3	1,3	6,7	3,3	5,7	2,1
5	3	1,6	0,2	6,6	3	4,4	1,4	7,2	3,2	6	1,8
5	3,4	1,6	0,4	6,8	2,8	4,8	1,4	6,2	2,8	4,8	1,8
5,2	3,5	1,5	0,2	6,7	3	5	1,7	6,1	3	4,9	1,8
5,2	3,4	1,4	0,2	6	2,9	4,5	1,5	6,4	2,8	5,6	2,1
4,7	3,2	1,6	0,2	5,7	2,6	3,5	1	7,2	3	5,8	1,6
4,8	3,1	1,6	0,2	5,5	2,4	3,8	1,1	7,4	2,8	6,1	1,9
5,4	3,4	1,5	0,4	5,5	2,4	3,7	1	7,9	3,8	6,4	2
5,2	4,1	1,5	0,1	5,8	2,7	3,9	1,2	6,4	2,8	5,6	2,2
5,5	4,2	1,4	0,2	6	2,7	5,1	1,6	6,3	2,8	5,1	1,5
4,9	3,1	1,5	0,2	5,4	3	4,5	1,5	6,1	2,6	5,6	1,4
5	3,2	1,2	0,2	6	3,4	4,5	1,6	7,7	3	6,1	2,3
5,5	3,5	1,3	0,2	6,7	3,1	4,7	1,5	6,3	3,4	5,6	2,4
4,9	3,6	1,4	0,1	6,3	2,3	4,4	1,3	6,4	3,1	5,5	1,8
4,4	3	1,3	0,2	5,6	3	4,1	1,3	6	3	4,8	1,8
5,1	3,4	1,5	0,2	5,5	2,5	4	1,3	6,9	3,1	5,4	2,1
5	3,5	1,3	0,3	5,5	2,6	4,4	1,2	6,7	3,1	5,6	2,4
4,5	2,3	1,3	0,3	6,1	3	4,6	1,4	6,9	3,1	5,1	2,3
4,4	3,2	1,3	0,2	5,8	2,6	4	1,2	5,8	2,7	5,1	1,9
5	3,5	1,6	0,6	5	2,3	3,3	1	6,8	3,2	5,9	2,3
5,1	3,8	1,9	0,4	5,6	2,7	4,2	1,3	6,7	3,3	5,7	2,5
4,8	3	1,4	0,3	5,7	3	4,2	1,2	6,7	3	5,2	2,3
5,1	3,8	1,6	0,2	5,7	2,9	4,2	1,3	6,3	2,5	5	1,9
4,6	3,2	1,4	0,2	6,2	2,9	4,3	1,3	6,5	3	5,2	2
5,3	3,7	1,5	0,2	5,1	2,5	3	1,1	6,2	3,4	5,4	2,3
5	3,3	1,4	0,2	5,7	2,8	4,1	1,3	5,9	3	5,1	1,8