Univerzita Jana Evangelisty Purkyně v Ústí nad Labem
Přírodovědecká fakulta

Úvod do teorie měření

Prof. Cihlář

Seminář 01

TÉMA:

Průměr, rozptyl a směrodatná odchylka

$$\overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2 \qquad S = \sqrt{S^2}$$

Výpočty pomocí vzorců a pomocí statistických funkcí
Vlastnosti průměru a rozptylu vzhledem k lineárním transformacím hodnot

1. Konstrukce tabulky pro izolované hodnoty:

index i	hodnoty xi	xi - průměr	(xi - průměr)^2
1	3		
2	6		
3	5		
4	3		
5	5		
6	9		
7	4		
8	5		
9	3		
10	2		
součet			
průměr		rozptyl	

   POZOR: průměr definovat jako název
   Povšimnout si nulového součtu odchylek

2. Aplikace statistických funkcí:

   PRUMER, VAR.VYBER a SMODCH.VYBER,

   kontrola výpočtů z tabulky.

3. Experimentování se zadanými čísly (demonstrace změn na číselné ose), sledování příslušných změn vypočítaných charakteristik:

   - všechna čísla zvětší o konstantu (například o 3),
   - všechna čísla se vynásobí konstantou (například číslem 2),
   - čísla se změní tak, aby průměr zůstal zachován a rozptyl se zmenšil (zvětšil),
   - atd.

Seminář 02

TÉMA:

Uspořádaný soubor, minimum, maximum, rozpětí
Medián, kvartily, kvartilové rozpětí
Výpočty těchto charakteristik pomocí statistických funkcí
Vlastnosti těchto charakteristik vzhledem k lineárním transformacím

4. Zvolte si sami 16 různých dvouciferných čísel a pomocí statistických funkcí MIN, MAX, MEDIAN, QUARTIL zjistěte hodnoty pěti požadovaných charakteristik.

index i	hodnoty xi		
1	23		
2	16		
3	15		
4	21	minimum	
5	95	dolní kvartil	
6	29	medián	
7	54	horní kvartil	
8	65	maximum	
9	43		
10	52		
11	31		
12	36		
13	75		
14	83		
15	25		
16	79		

5. Z jejich číselných hodnot odhalte, jaký je jejich význam, a jak se počítají (soubor hodnot si můžete uspořádat).
6. Jaký význam mají pojmy rozpětí a kvartilové rozpětí?
7. Jak se pořadové charakteristiky mění, když hodnoty lineárně transformujeme či jinak měníme, například:
   - všechna čísla zvětšíme o konstantu (například o 3),
   - všechna čísla se vynásobíme konstantou (například číslem 2),
   - jak máme změnit čísla, aby medián zůstal zachován a rozpětí (kvartilové rozpětí) se zmenšilo (zvětšilo),
   - atd.
5. Seznamte se se statistickým nástrojem **Popisná statistika** pro pole hodnot.

Seminář 03

TÉMA:

Generování náhodných veličin (zejména normální rozdělení)
Konstrukce histogramu
Vylučování odlehlých hodnot

8. Naučte se používat nástroj **Generátor pseudonáhodných čísel** (je třeba mít k dispozici doplněk **Analýza Dat**) a statistickou funkci **Četnosti** (dvojhmat).
   - Simulujte 100 hodů mincí a zjistěte počet hozených líců a rubů.
   - Simulujte 200 hodů kostkou a zjistěte, kolikrát padla jednotlivá čísla.
   - Simulujte náhodný výběr rozsahu 500 z výšek lidí pomocí generátoru normálního rozdělení ($\mu = 175$, $\sigma = 10$).
9. Naučte se používat nástroj **Histogram** (v doplňku Analýza dat).

   Není vhodné, aby počet třídních intervalů byl příliš malý anebo příliš velký. Doporučuje se jej volit tak, aby byl přibližně roven číslu ze Sturgesova vzorce: $1 + 3{,}3 \cdot \log n$, kde n je počet měření.
10. Vylučování odlehlých hodnot pomocí **vnitřních hradeb**:

    Pomocí dolního kvartilu *DK*, horního kvartilu *HK* a kvartilového rozpětí *KR*, které je dáno vztahem $KR = HK - DK$, vypočítáme obě vnitřní hradby:

    dolní hradba : $DH = DK - 1{,}5 \cdot KR$,

    horní hradba : $HH = HK + 1{,}5 \cdot KR$.

    Za odlehlé hodnoty považujeme ty, které jsou **menší než dolní hradba** a **větší než horní hradba**. Tyto hodnoty z výběrového souboru vyloučíme a test pak opakujeme pro redukovaný soubor. Dále zpracováváme jen hodnoty zbývající.
11. Vylučování odlehlých hodnot pomocí **Grubbsova testu** (vhodné pro menší výběrové soubory, kde rozsah výběru n nepřevyšuje číslo 20):

    Nejprve vypočítáme pomocí směrodatné odchylky *S* číslo $S_n = S \cdot \sqrt{\frac{n-1}{n}}$.

    Pak pro minimum souboru *min* vypočítáme hodnotu $T_{\min} = \frac{\overline{X} - \min}{S_n}$,

    a podobně pro maximum souboru *max* vypočítáme hodnotu $T_{\max} = \frac{\max - \overline{X}}{S_n}$.

    Extrémní hodnotu vyloučíme, pokud **vypočtená hodnota $T_{\min}$ či $T_{\max}$ převýší hodnotu $T(n,\alpha)$ uvedenou v následující tabulce**. Tento test pak

opakujeme pro redukovaný soubor do té doby, než extrémní hodnotu již test nevyloučí. Dále zpracováváme jen hodnoty zbývající.

5. Vylučování odlehlých hodnot pomocí **Dean-Dixonova Q-testu** (vhodné pro malé výběrové soubory, kde rozsah výběru n nepřevyšuje číslo 10):
   Pro tento test potřebujeme nejprve vypočítat rozpětí $R = max - min$.
   Hodnoty souboru uspořádáme podle velikosti vzestupně tak, aby bylo
   $$\min = X_1 < X_2 < X_3 < \ \ldots \ < X_{n-1} < X_n = \max .$$
   (Pokud nechceme soubor uspořádávat, můžeme získat druhou nejmenší hodnotu a druhou největší hodnotu pomocí nástroje Popisná statistika.)
   Pak pro minimum souboru *min* vypočítáme hodnotu $Q_{\min} = \frac{X_2 - \min}{R}$,
   a pro maximum souboru *max* vypočítáme hodnotu $Q_{\max} = \frac{\max - X_{n-1}}{R}$.

   Extrémní hodnotu vyloučíme, pokud **vypočtená hodnota $Q_{\min}$ či $Q_{\max}$ převýší hodnotu $Q(n,\alpha)$ uvedenou v následující tabulce**. Tento test pak opakujeme pro redukovaný soubor do té doby, než extrémní hodnotu již test nevyloučí. Dále zpracováváme jen hodnoty zbývající.

6. V následující tabulce se vyskytuje tzv. **hladina významnosti** α . Je to hodnota našeho rizika, že se při použití testu dopustíme chyby (přesněji: je to pravděpodobnost toho, že testem označíme hodnotu za odlehlou, i když tomu tak ve skutečnosti není).

Kritické hodnoty pro testy vylučování odlehlých výsledků

Počet měření n	Grubbsův test $T(n,\alpha)$		Dean-Dixonův Q-test $Q(n,\alpha)$	
	α = 0,05	α = 0,01	α = 0,05	α = 0,01
3	1,412	1,416	0,941	0,988
4	1,689	1,723	0,765	0,889
5	1,869	1,955	0,642	0,760
6	1,996	2,130	0,560	0,698
7	2,093	2,265	0,507	0,637
8	2,172	2,374	0,468	0,590
9	2,237	2,464	0,437	0,555
10	2,294	2,540	0,412	0,527
11	2,343	2,606		
12	2,387	2,663		
13	2,426	2,714		
14	2,461	2,759		
15	2,493	2,800		
16	2,523	2,837		
17	2,551	2,871		
18	2,557	2,903		
19	2,600	2,932		
20	2,623	2,959		

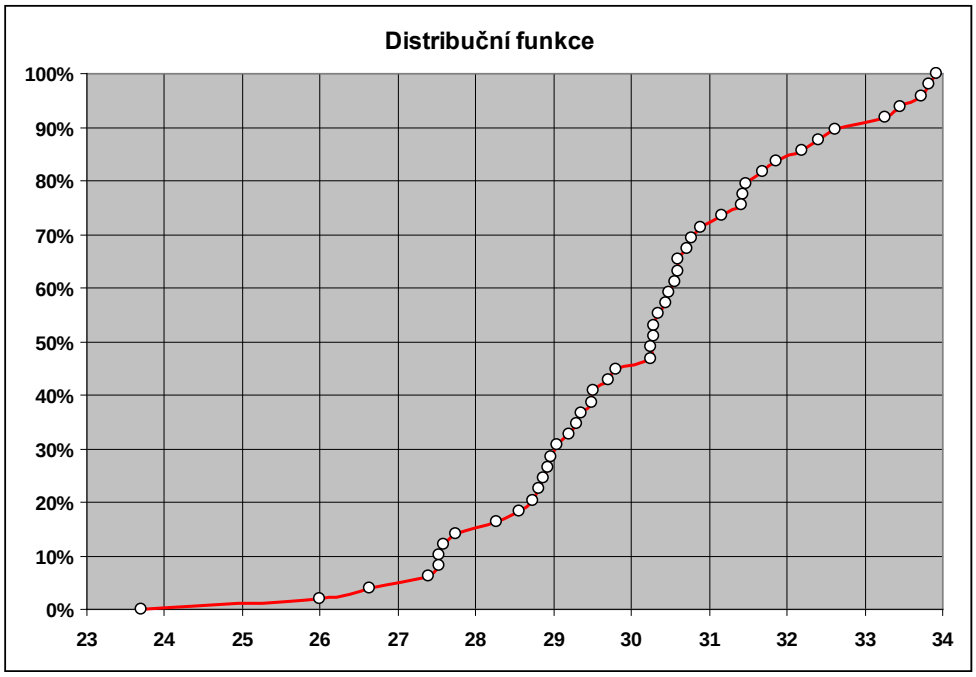
Seminář 04

TÉMA:

Distribuční funkce, kvantily

12. Pomocí generátoru pseudonáhodných čísel si vytvořte soubor 50 čísel s normálním rozdělením (střední hodnotu a směrodatnou odchylku si zvolte libovolně). Na tato data užijte nástroj **Pořadová statistika a percentily**. Odhalte význam údajů ve všech sloupcích získané tabulky.

13. Pomocí údajů v tabulce vytvořte graf tzv. **distribuční funkce**, která pro libovolně zvolenou hodnotu udává, kolik procent čísel z daného souboru je menší, než tato hodnota:



- Odečtěte z grafu přibližnou hodnotu mediánu a zkontrolujte si svůj odhad jeho stanovením pomocí funkce MEDIAN či QUARTIL.
- Totéž udělejte pro oba kvartily.
- Jaký význam mají čísla, která nazýváme **decily**, resp. **centily**?

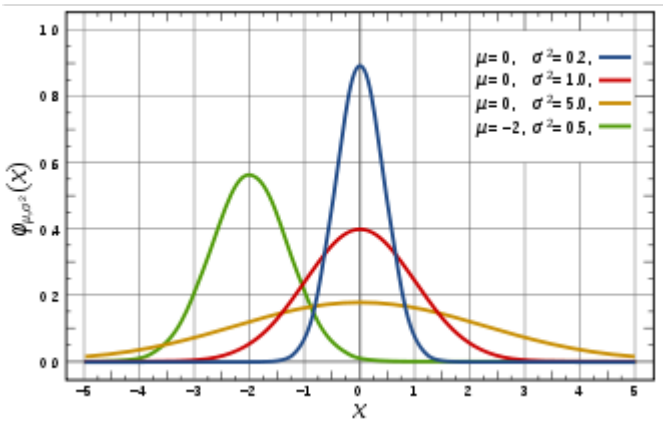
14. Naučte se používat statistické funkce **PERCENTIL** a **PERCENTRANK**. Jaký je jejich vztah ke grafu distribuční funkce?

Seminář 05

TÉMA:

Bodové a intervalové odhady pro parametry normálního rozdělení

Statistické zpracování hodnot opakovaných měření nějaké veličiny vychází tohoto předpokladu: nahodilé chyby způsobují, že naměřené hodnoty x se od správné hodnoty μ liší, přičemž malé odchylky (na obě strany) jsou více pravděpodobné a větší odchylky jsou málo pravděpodobné. Vhodným modelem pro naměřené hodnoty x je tedy normální rozdělení $No(\mu;\sigma^2)$, kde μ je střední hodnota a rozptyl σ^2 je charakteristikou přesnosti měřící metody.



15. Pomocí generátoru pseudonáhodných čísel si vytvořte soubor 10 000 čísel s normálním rozdělením (střední hodnotu volte 30 a směrodatnou odchylku volte 3). Tato data uspořádejte do 20 sloupců a 500 řádků. Tato čísla budou modelem měření veličiny se správnou hodnotou 30, která provádělo 500 experimentátorů, z nichž každý hodnotu měřil nezávisle 20krát. Pro měření každého experimentárora (každý řádek) vypočítejte výběrový průměr a výběrový rozptyl.
16. Pomocí nástrojů Popisná statistika a Histogram porovnejte rozdělení hodnot u základních dat (10 000 čísel), výběrových průměrů (500 čísel) a výběrových rozptylů (dalších 500 čísel). Jaké závěry plynou ze získaných informací?

Výběrový průměr je vhodným bodovým odhadem střední hodnoty μ normálního rozdělení (tedy správné hodnoty, kterou měříme).
Výběrový rozptyl je vhodným bodovým odhadem rozptylu σ^2 normálního rozdělení (tedy „přesnosti“ metody, kterou pro měření užíváme).
Oba dva bodové odhady jsou však zatíženy nahodilými chybami, hodnoty bodových odhadů jsou tedy jen přibližně rovny správným hodnotám.

17. Při zpracování měření se pokoušíme stanovit rozmezí (interval) v němž skutečná (neznámá) hodnota s velkou pravděpodobností leží. Například:

95% procentní interval spolehlivosti pokrývá neznámou hodnotu parametru s pravděpodobností (spolehlivostí) 0,95 = 95% .
99% procentní interval spolehlivosti pokrývá neznámou hodnotu parametru s pravděpodobností (spolehlivostí) 0,95 = 95% .

18. Výpočet **intervalu spolehlivosti pro parametr μ normálního rozdělení** provedeme podle tohoto tvrzení:

$$PRAVD\ (\overline{X} - t \cdot \frac{S}{\sqrt{n}} < \mu < \overline{X} + t \cdot \frac{S}{\sqrt{n}}) = 1 - \alpha$$

kde kvantil (percentil) t Studentova rozdělení získáme pomocí funkce TINV s těmito hodnotami parametrů: Prst = α , Volnost = $n - 1$, anebo použijeme dále uvedenou tabulku kvantilů.

- Zjistěte si pro zvolenou hodnotu $\alpha = 0{,}05$ a hodnotu $n = 20$ číselnou hodnotu kvantilu t.
- Vypočítejte pro každý řádek dolní mez intervalu spolehlivosti $\overline{X} - t.\frac{S}{\sqrt{n}}$.
- Vypočítejte pro každý řádek horní mez intervalu spolehlivosti $\overline{X} + t.\frac{S}{\sqrt{n}}$.
- Zjistěte logickou operací v každém řádku, zda byla experimentátorem intervalem spolehlivosti zachycena správná hodnota $\mu = 30$ (dolní mez je menší než 30 a současně je horní mez větší než 30).
- Zjistěte (pomocí funkce Průměr) u kolika procent experimentátorů se intervalem spolehlivosti podařilo pokrýt správnou hodnotu 30.

19. Výpočet **intervalu spolehlivosti pro parametr σ^2 normálního rozdělení** provedeme podle tohoto tvrzení:

$$PRAVD\ (\frac{(n-1) \cdot S^2}{\chi_1^2} < \sigma^2 < \frac{(n-1) \cdot S^2}{\chi_2^2}) = 1 - \alpha$$

kde kvantily χ_1^2 a χ_2^2 získáme funkcí CHIINV s těmito hodnotami parametrů: pro χ_1^2 : Prst = α /2 , Volnost = n – 1 ,
pro χ_2^2 : Prst = 1 - α /2 , Volnost = n – 1 ,
anebo použijeme dále uvedenou tabulku kvantilů.

- Zjistěte si pro zvolenou hodnotu $\alpha = 0{,}05$ a hodnotu $n = 20$ číselnou hodnotu kvantilů χ_1^2 a χ_2^2 .
- Vypočítejte pro každý řádek dolní mez intervalu spolehlivosti $\frac{(n-1).S^2}{\chi_1^2}$.

- Vypočítejte pro každý řádek horní mez intervalu spolehlivosti $\frac{(n-1).S^2}{\chi_2^2}$.
- Zjistěte logickou operací v každém řádku, zda byla experimentátorem intervalem spolehlivosti zachycena správná hodnota $\sigma^2 = 9$ (dolní mez je menší než 9 a současně je horní mez větší než 9).
- Zjistěte (pomocí funkce Průměr) u kolika procent experimentátorů se intervalem spolehlivosti podařilo pokrýt správnou hodnotu 9.

Tabulky kvantilů:

n	$\alpha = 0{,}05$			$\alpha = 0{,}01$		
	t	χ_1^2	χ_2^2	***t***	χ_1^2	χ_2^2
3	4,30266	7,37778	0,05064	9,92499	10,59653	0,01002
4	3,18245	9,34840	0,21579	5,84085	12,83807	0,07172
5	2,77645	11,14326	0,48442	4,60408	14,86017	0,20698
6	2,57058	12,83249	0,83121	4,03212	16,74965	0,41175
7	2,44691	14,44935	1,23734	3,70743	18,54751	0,67573
8	2,36462	16,01277	1,68986	3,49948	20,27774	0,98925
9	2,30601	17,53454	2,17972	3,35538	21,95486	1,34440
10	2,26216	19,02278	2,70039	3,24984	23,58927	1,73491
11	2,22814	20,48320	3,24696	3,16926	25,18805	2,15585
12	2,20099	21,92002	3,81574	3,10582	26,75686	2,60320
13	2,17881	23,33666	4,40378	3,05454	28,29966	3,07379
14	2,16037	24,73558	5,00874	3,01228	29,81932	3,56504
15	2,14479	26,11893	5,62872	2,97685	31,31943	4,07466
16	2,13145	27,48836	6,26212	2,94673	32,80149	4,60087
17	2,11990	28,84532	6,90766	2,92079	34,26705	5,14216
18	2,10982	30,19098	7,56418	2,89823	35,71838	5,69727
19	2,10092	31,52641	8,23074	2,87844	37,15639	6,26477
20	2,09302	32,85234	8,90651	2,86094	38,58212	6,84392
21	2,08596	34,16958	9,59077	2,84534	39,99686	7,43381
22	2,07961	35,47886	10,28291	2,83137	41,40094	8,03360
23	2,07388	36,78068	10,98233	2,81876	42,79566	8,64268
24	2,06865	38,07561	11,68853	2,80734	44,18139	9,26038
25	2,06390	39,36406	12,40115	2,79695	45,55836	9,88620
26	2,05954	40,64650	13,11971	2,78744	46,92797	10,51965
27	2,05553	41,92314	13,84388	2,77872	48,28978	11,16022
28	2,05183	43,19452	14,57337	2,77068	49,64504	11,80765
29	2,04841	44,46079	15,30785	2,76326	50,99356	12,46128
30	2,04523	45,72228	16,04705	2,75639	52,33550	13,12107

Seminář 06

TÉMA:

Principy testování statistických hypotéz

Testy o parametrech normálního rozdělení $No(\mu;\sigma^2)$ - jeden výběr

1. Ilustrativní příklad:
   Hraji se soupeřem hru, při níž záleží na tom, jak nám padají šestky na hozených kostkách. Zatímco u mé kostky padá šestka podle očekávání zhruba v jedné šestině případů, zdá se mi, že na jeho kostce padá šestka daleko častěji. Hlodá ve mně podezření, že jeho kostka je „falešná“, on to ale popírá.
   Dohodli jsme se, že test jeho kostky uděláme takto: hodí 24krát kostkou a spočítáme, kolikrát mu padne šestka. Když bude počet hozených šestek „moc velký“, prohlásíme kostku za falešnou a vyřadíme ji ze hry. Jaký význam ale máme dát slovům „moc velký“? Pomůže nám následující tabulka?

Počet hozených šestek	Pravděpodobnost tohoto jevu za podmínky, že kostka je „správná“	Hladina významnosti 0,05
0	0,01257911521248	**0,95 = 95%**
1	0,06037975301988	
2	0,13887343194573	
3	0,20368103352040	
4	0,21386508519642	
5	0,17109206815714	
6	0,10835830983285	
7	0,05572713077118	
8	0,02368403057775	**0,05 = 5%**
9	0,00842098864987	
10	0,00252629659496	
11	0,00064305731508	
12	0,00013932908493	
13	0,00002572229260	
14	0,00000404207455	
15	0,00000053894327	
16	0,00000006063112	
17	0,00000000570646	
18	0,00000000044384	
19	0,00000000002803	
20	0,00000000000140	
21	0,00000000000005	
22	0,00000000000000	
23	0,00000000000000	
24	0,00000000000000	

2. Obecný postup při testování statistických hypotéz o parametrech normálního rozdělení:
   - Nejprve zformulujeme tzv. ***nulovou hypotézu H_0*** o vybraném parametru rozdělení. Nulová hypotéza má tvar rovnosti, například: $\mu = 175$ nebo $\sigma^2 = 10{,}2$, a podobně.
   - Proti této hypotéze postavíme tzv. ***alternativní hypotézu H_a***, která má obvykle tvar nerovnosti, například: $\mu > 175$ nebo $\sigma^2 < 10{,}2$, a podobně.
   - Vybereme vhodnou náhodnou veličinu G, tzv. ***testové kritérium***.
   - Zvolíme malé kladné číslo α (bývá zvykem volit zejména hodnoty $\alpha = 0{,}10$, resp. $\alpha = 0{,}05$, resp. $\alpha = 0{,}01$), které budeme nazývat ***hladinou významnosti***.
   - Určíme tzv. ***kritický obor W***. Ten má tuto vlastnost: jestliže platí nulová hypotéza H_0, pak hodnota testového kritéria G padne do W s malou pravděpodobností α, a naopak skoro jistě (s pravděpodobností 1 - α) hodnota G nepadne do W.
   - Z dat vypočteme hodnotu testového kritéria a porovnáme s kritickým oborem:

jestliže $G \in W$, pak zamítneme nulovou hypotézu H_0,
jestliže $G \notin W$, pak nezamítneme nulovou hypotézu H_0.

3. Testová kritéria a kritické obory pro jednotlivé hypotézy a pro jejich alternativy:

T-test pro nulovou hypotézu H_0: $\mu = konst$

Proti nulové hypotéze stavíme alternativní hypotézu H_a: $\mu < konst$, když $\overline{X} < konst$.

Proti nulové hypotéze stavíme alternativní hypotézu H_a: $\mu > konst$, když $\overline{X} > konst$.

Testovým kritériem je náhodná veličina $G = \dfrac{\left| \overline{X} - konst \right|}{S} \cdot \sqrt{n}$

Zvolíme hladinu významnosti α, nejčastěji $\alpha = 0{,}05$.

Kritickým oborem bude interval $W = (t, +\infty)$, kde kvantil t Studentova rozdělení získáme funkcí TINV s volbou parametrů Prst = 2.α, Volnost = n – 1.

χ^2-test pro nulovou hypotézu H_0: $\sigma^2 = konst$

Proti nulové hypotéze stavíme alternativní hypotézu H_a: $\sigma^2 < konst$, když $S^2 < konst$.

Proti nulové hypotéze stavíme alternativní hypotézu H_a: $\sigma^2 > konst$, když $S^2 > konst$.

Testovým kritériem je náhodná veličina $G = \dfrac{(n-1) \cdot S^2}{konst}$

Zvolíme hladinu významnosti α, nejčastěji $\alpha = 0{,}05$.

Kritickým oborem při alternativě $\sigma^2 < konst$ bude interval $W = (0, \chi^2)$, kde kvantil χ^2 získáme funkcí CHIINV s volbou parametrů Prst = 1 - α, Volnost = n – 1.

Kritickým oborem při alternativě $\sigma^2 > konst$ bude interval $W = (\chi^2, +\infty)$, kde kvantil χ^2 získáme funkcí CHIINV s volbou parametrů Prst = α, Volnost = n – 1.

4. Vygenerujte si data a testujte různé hypotézy na různých hladinách významnosti.

Seminář 07

TÉMA:

Testy o parametrech normálního rozdělení – dva výběry

1. Předpokládáme, že: jeden výběr pochází z rozdělení $No(\mu_1;\sigma_1^2)$
   a druhý výběr pochází z rozdělení $No(\mu_2;\sigma_2^2)$.
   Používáme nástroj Popisná statistika pro zjištění poměrů ve výběrech a následně dále uvedené testy.
2. Mohou nastat dva případy:
   - **Výběry jsou závislé** (jde o dvě opakovaná měření na týchž statistických jednotkách, oba datové soubory tedy mají stejný počet měření).
     V tomto případě pro test nulové hypotézy: $\mu_1 = \mu_2$ použijeme tzv. **Dvouvýběrový párový t-test.**
   - **Výběry jsou nezávislé** (hodnoty z výběrů se navzájem neovlivňují, rozsah obou souborů nemusí být obecně stejný).
     V tomto případě pro test nulové hypotézy: $\mu_1 = \mu_2$ máme k dispozici dva tzv. t-testy, a to:
     Dvouvýběrový t-test s rovností rozptylů
     a **Dvouvýběrový t-test s nerovností rozptylů.**
     O tom, který z těchto testů použijeme se rozhodujeme na základě tzv.
     Dvouvýběrového F-testu pro rozptyl,
     při kterém testujeme nulovou hypotézu $\sigma_1^2 = \sigma_2^2$.

U všech těchto testů volíme za 1. soubor vždy ten, který má větší odhad testovaného parametru (tedy buď výběrový průměr nebo výběrový rozptyl) a za 2. soubor ten, který má odhad testovaného parametru menší.

3. Nulové hypotézy testujeme na hladině významnosti α (obvykle volíme 0,05). Počítač nám ale hladinu významnosti sám vypočítá, je to **hodnota P**, která se objeví v tabulce.

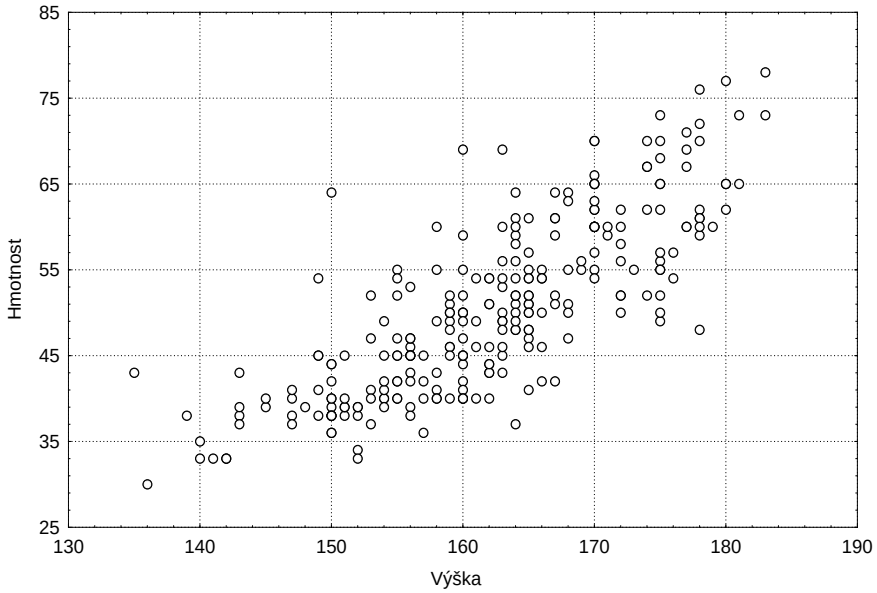
Nulovou hypotézu tedy zamítáme, když je P–hodnota menší než 0,05 (resp. jiná zvolená hladina významnosti).
Tomu také odpovídá situace v tabulce, kdy vypočtená hodnota testového kritéria převyšuje tzv. kritickou hodnotu.

4. Generujte si soubory dat a používejte výše uvedené testy.

Seminář 08

TÉMA:

Závislost normálně rozdělených náhodných veličin, korelace, grafické znázornění



Na obrázku je typická statistická závislost.

Statistickou závislost obvykle modelujeme vhodnou funkční závislostí, v nejjednodušším případě prokládáme body přímku.

Těsnost lineární statistické závislosti měříme koeficientem korelace, který se počítá podle následujícího vzorce, resp. pomocí funkce CORREL.

$$r = \frac{n \cdot \sum x_i y_i - \sum x_i \cdot \sum y_i}{\sqrt{(n \cdot \sum x_i^2 - (\sum x_i)^2) \cdot (n \cdot \sum y_i^2 - (\sum y_i)^2)}}$$

Koeficient korelace nabývá hodnoty od -1 do 1 a přitom:

hodnotě $r = 1$	odpovídá rostoucí funkční lineární závislost,
hodnotě r mezi 0 a 1	odpovídá rostoucí statistická lineární závislost,
hodnota $r = 0$	signalizuje neexistenci lineární závislosti,
hodnotě r mezi -1 a 0	odpovídá klesající statistická lineární závislost,
hodnotě $r = -1$	odpovídá klesající funkční lineární závislost.

Například těsnost statistické závislosti na hořejším obrázku je charakterizována hodnotou korelačního koeficientu $r = 0{,}7962$.

1. Experimentujte s daty a ověřujte vlastnosti koeficientu korelace.

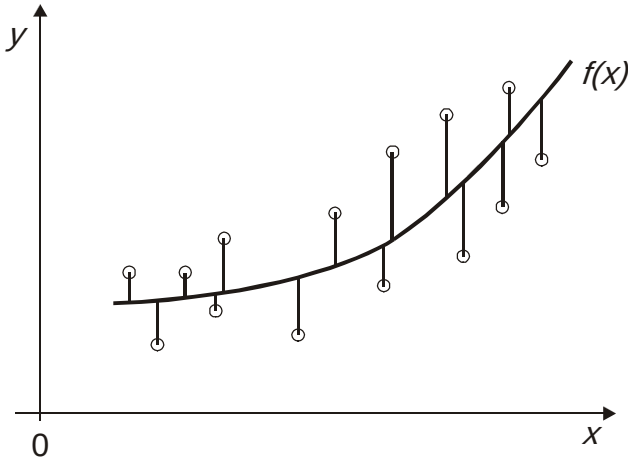
Seminář 09

TÉMA:

Regrese, metoda nejmenších čtverců, pás spolehlivosti pro regresní funkci

Statistickou závislost obvykle modelujeme vhodnou funkční závislostí. Tuto funkci hledáme tak, aby součet druhých mocnin odchylek měření od hodnoty regresní funkce byl minimální (používáme tzv. **metodu nejmenších čtverců**).

V jednoduchých situacích volíme lineární závislost, jejímž grafem je přímka.



1. Budeme tedy předpokládat, že závislost veličiny *y* na veličině *x* je lineární a regresní funkce má tvar $y = f(x) = b_1 + b_2 \cdot x$.
2. Potřebné výpočty uspořádáme do podobné tabulky, kterou jsme používali při výpočtu koeficientu korelace:

i	x_i	y_i	x_i^2	$x_i \cdot y_i$	y_i^2
1					
2					
3					
4					
5					
atd.					
Součet					

Neznámá čísla b_1 a b_2 v rovnici regresní funkce vypočítáme z údajů posledního součtového řádku podle těchto vzorců:

$$b_1 = \frac{\sum x_i^2 \cdot \sum y_i - \sum x_i \cdot \sum x_i y_i}{n \cdot \sum x_i^2 - \left(\sum x_i\right)^2}, \qquad b_2 = \frac{n \cdot \sum x_i y_i - \sum x_i \cdot \sum y_i}{n \cdot \sum x_i^2 - \left(\sum x_i\right)^2}.$$

3. Zadejte si libovolně dvojice čísel reprezentující výsledky měření, vypočtěte údaje v tabulce, nalezněte rovnici regresní přímky, naprogramujte její výpočet do dalšího sloupce tabulky a vytvořte přehledný graf.
4. Produkt Excel umožňuje pro statistickou závislost rychle nalézt regresní přímku – v nabídce grafu stačí zadat požadavky na vytvoření spojnice trendu a zobrazení její rovnice. Seznamte se s touto možností a zkuste použít i další různé regresní funkce.
5. Vhodnost regresní funkce posuzujeme velikostí čísla, které se nazývá **reziduální součet čtverců**:

$$s_r = \sum (y_i - f(x_i))^2$$

Doplňte výpočetní tabulku o další sloupec a vypočítejte reziduální součet čtverců. Přesvědčte se, že jej pro případ regresní přímky lze počítat i podle následujícího vzorce:

$$s_r = \sum y_i^2 - b_1 \cdot \sum y_i - b_2 \cdot \sum x_i y_i \ .$$

Reziduální součet čtverců slouží i k odhadu rozptylu chyb, kterých jsme se při měření dopustili. Odhad rozptylu je dán tímto vzorcem:

$$\sigma^2 \approx \frac{s_r}{n-2}$$ (platí pro přímkovou regresi).

6. Pomocí reziduálního součtu čtverců můžeme také vypočítat 95% interval spolehlivosti pro hodnotu regresní funkce $f(x)$ pomocí vzorce:

$$f(x) \ \pm \ t \cdot \sqrt{\frac{s_r \cdot (\sum x_i^2 - 2 \cdot x \cdot \sum x_i + n \cdot x^2)}{(n-2) \cdot (n \cdot \sum x_i^2 - \left(\sum x_i \right)^2)}} \ ,$$

kde kvantil t vyhledáme pomocí statistické funkce TINV s těmito hodnotami parametrů: Prst = 0,05 , Volnost = $n - 2$.

Doplňte výpočetní tabulku o další dva sloupce a naprogramujte do nich dolní a horní mez intervalu spolehlivosti pro funkční hodnotu regresní funkce. Doplňte i graf – vytvoří se vám tzv. **pás spolehlivosti pro regresní funkci.**