Cvičení 4.: Ověřování normality dat, parametrické úlohy o jednom náhodném výběru z normálního rozložení

Kolmogorovův – Smirnovův test normality dat

Testujeme nulovou hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z normálního rozložení s parametry μ a σ^2. Distribuční funkci tohoto rozložení označme $\Phi_T(x)$. Nechť $F_n(x)$ je výběrová distribuční funkce. Testovou statistikou je statistika $D_n = \sup_{-\infty < x < \infty} |F_n(x) - \Phi_1(x)|$.

Nulovou hypotézu zamítáme na hladině významnosti α, když $D_n \geq D_n(\alpha)$, kde $D_n(\alpha)$ je tabelovaná kritická hodnota.
V případě, že neznáme parametry μ a σ^2 normálního rozložení (což je nejčastější případ), změní se rozložení testové statistiky D_n. V takovém případě jde o Lilieforsovu modifikaci Kolmogorovova – Smirnovova testu. Příslušné modifikované kvantily byly určeny pomocí simulačních studií.

Poznámka ke K-S testu ve STATISTICE

Test normality poskytuje hodnotu testové statistiky (ozn. max D) a dvě p-hodnoty. (p-hodnota vyjadřuje pravděpodobnost, s jakou číselné realizace $x_1, ..., x_n$ náhodného výběru $X_1, ..., X_n$ podporují nulovou hypotézu, je-li pravdivá. P-hodnotu porovnáváme s námi zvolenou hladinou významnosti α. Jestliže p-hodnota $\leq \alpha$, pak H_0 zamítáme na hladině významnosti α, je-li p-hodnota $> \alpha$, pak H_0 nezamítáme na hladině významnosti α.) První p-hodnota se vztahuje k případu, kdy střední hodnotu μ a rozptyl σ^2 známe předem, druhá (ozn. Lilieforsovo p) se vztahuje k případu, kdy μ a σ^2 neznáme. Objeví-li se ve výstupu p = n.s. (tj. non significant), pak hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

Shapirův – Wilkův test normality dat

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z rozložení $N(\mu, \sigma^2)$.
Test je založen na zjištění, zda body v Q-Q grafu jsou významně odlišné od regresní přímky proložené těmito body.
(S-W test se používá především pro výběry menších rozsahů, $n < 50$, ale nyní již existuje modifikace pro velká n. V systému STATISTICA je implementováno rozšíření na n kolem 5000.)

Úkol 1. : U 45 studentek VŠE v Praze byla zjišťována výška a obor studia (1 – národní hospodářství, 2 – informatika). Hodnoty jsou uloženy v souboru vyska.sta. Pomocí Lilieforsovy modifikace K-S testu, pomocí S-W testu a pomocí testu dobré shody testujte na hladině významnosti 0,05 hypotézu, že data pocházejí z normálního rozložení. Pomocí N-P grafu posuďte vizuálně předpoklad normality.

Návod:

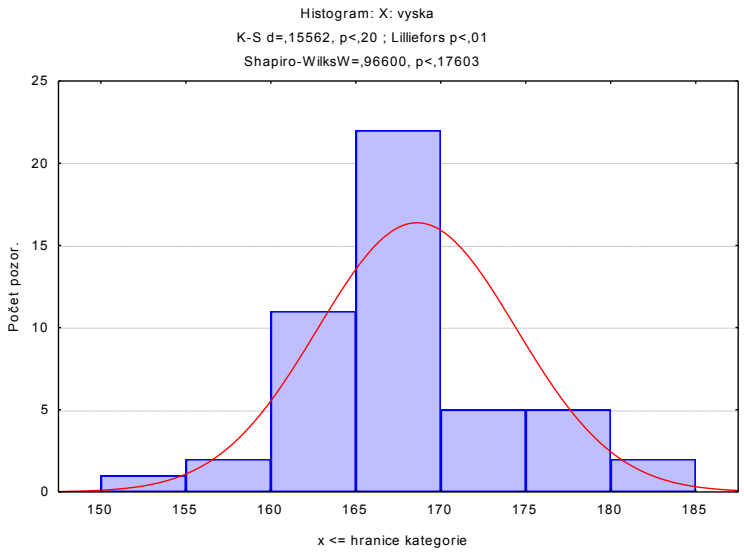
1. způsob provedení Lilieforsova a S-W testu: Statistiky – Základní statistiky/tabulky – Tabulky četností – OK – Proměnné X – OK – Normalita – zaškrtneme Lilieforsův test a S-W test – Testy normality.

	Testy normality (vyska.sta)				
Proměnná	N	max D	Lilliefors p	W	p
X: vyska	48	0,155621	p < ,01	0,965996	0,176031

Výstupní tabulka obsahuje počet pozorování, hodnotu testové statistiky Lilieforsovy modifikace K-S testu (max D = 0,155621), p-hodnotu ($p < 0,01$), testovou statistiku S-W testu (W = 0,965996) a odpovídající p-hodnotu (p = 0,176031). Vidíme, že Lilieforsův test zamítá hypotézu o normalitě na hladině významnosti 0,05, zatímco S-W test nikoli.

2. způsob provedení Lilieforsova a S-W testu: Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné X – OK – Normalita – zaškrtneme K-S test & Lilieforsův test a S-W test – Tabulky četností (nebo Histogram).

	Tabulka četností:X: vyska (vyska.sta) K-S d=,15562, p<,20 ; Lilliefors p<,01 Shapiro-WilksW=,96600, p<,17603					
Kategorie	Četnost	Kumulativní četnost	Rel.četn. (platných)	Kumul. % (platných)	Rel.četn. všech	Kumul. % všech
150,0000<x<=155,0000	1	1	2,08333	2,0833	2,08333	2,0833
155,0000<x<=160,0000	2	3	4,16667	6,2500	4,16667	6,2500
160,0000<x<=165,0000	11	14	22,91667	29,1667	22,91667	29,1667
165,0000<x<=170,0000	22	36	45,83333	75,0000	45,83333	75,0000
170,0000<x<=175,0000	5	41	10,41667	85,4167	10,41667	85,4167
175,0000<x<=180,0000	5	46	10,41667	95,8333	10,41667	95,8333
180,0000<x<=185,0000	2	48	4,16667	100,0000	4,16667	100,0000
ChD	0	48	0,00000		0,00000	100,0000



V tomto případě dostaneme v záhlaví tabulky či histogramu stejné informace jako pomocí předešlého způsobu.

Samostatný úkol: Testy normality a grafické ověření normality proveďte jak pro výšky studentek oboru národní hospodářství, tak pro výška studentek oboru informatiky.

Pro kontrolu:

Výsledky pro obor národní hospodářství:

	Testy normality (vyska.sta) Zhrnout podmínku: z=1				
Proměnná	N	max D	Lilliefors p	W	p
X: vyska	28	0,167473	p < ,05	0,970969	0,606793

Vidíme, že Lilieforsova varianta K-S testu zamítá hypotézu o normalitě na hladině významnosti 0,05 (p-hodnota je menší než 0,05), zatímco S-W test hypotézu o normalitě nezamítá (p-hodnota je větší než 0,05).

Výsledky pro obor informatika:

Proměnná	Testy normality (vyska.sta) Zhrnout podmínku: z=2 N	max D	Lilliefors p	W	p
X: vyska	20	0,172301	p < ,15	0,922747	0,111924

V tomto případě ani jeden z testů hypotézu o normalitě nezamítá na hladině významnosti 0,05.

Upozornění: V archivu závěrečných prací https://is.muni.cz/auth/th/77721/prif_m/ je uložena diplomová práce Dominika Grůzy „Ověřování normality“.

Úkol 2.: Vlastnosti výběrového průměru z normálního rozložení

Předpokládejme, že velký ročník na vysoké škole má výsledky ze statistiky normálně rozloženy kolem střední hodnoty 72 bodů se směrodatnou odchylkou 9 bodů. Najděte pravděpodobnost, že průměr výsledků náhodného výběru 10 studentů bude větší než 80 bodů.

Návod:

$X_1, ..., X_{10}$ je náhodný výběr z N(72, 81). Počítáme P(M > 80), přičemž výběrový průměr M má normální rozložení se střední hodnotou E(M) = μ = 72 a rozptylem $D(M) = \frac{\sigma}{n} = \frac{81}{10} =$ 8,1 (viz skripta Základní statistické metody, věta 6.1.1.1., bod 2).

Tedy P(M > 80) = 1 - P(M ≤ 80) = 1 – Φ(80), kde Φ(80) je hodnota distribuční funkce rozložení N(72; 8,1) v bodě 80.

Vytvoříme datový soubor o jedné proměnné a o jednom případu. Do Dlouhého jména této proměnné napíšeme =1 – INormal(80;72;sqrt(8,1)). Zjistíme, že 1 - Φ(80) = 0,00247005.

Funkce INormal(x;μ;σ) počítá hodnotu distribuční funkce rozložení $N(\mu,\sigma^2)$ v bodě x.

	1 Prom1
1	0,00247

Úkol 3.: Intervaly spolehlivosti pro parametry μ, σ^2 normálního rozložení

Z populace stejně starých selat téhož plemene bylo vylosováno šest selat a po dobu půl roku jim byla podávána táž výkrmná dieta. Byly zaznamenávány průměrné denní přírůstky hmotnosti v Dg. Z dřívějších pokusů je známo , že v populaci mívají takové přírůstky normální rozložení, avšak střední hodnota i rozptyl se měnívají. Přírůstky v Dg: 62, 54, 55, 60, 53, 58.

a) Najděte 95% empirický levostranný interval spolehlivosti pro neznámou střední hodnotu μ při neznámé směrodatné odchylce σ.
b) Najděte 95% empirický interval spolehlivosti pro směrodatnou odchylku σ.

Návod:

Vytvoříme nový datový soubor o jedné proměnné X a 6 případech. Do proměnné X napíšeme dané hodnoty.

Ad a) Statistiky – Základní statistiky a tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – zaškrtneme Meze spolehl. prům. (ostatní volby zrušíme) – pro jednostranný interval změníme hodnotu na 90,00 - Výpočet. (Hodnotu změníme na 90, protože dolní mez levostranného 95% intervalu spolehlivosti pro μ je stejná jako dolní mez oboustranného 95% intervalu spolehlivosti pro μ.)

	Popisné statistiky (Tabulka1)	
Proměnná	Int. spolehl. -90,000%	Int. spolehl. 90,000
X	54,05683	59,94317

Vidíme, že $\mu > 54{,}06$ Dg s pravděpodobností aspoň 0,95.

Ad b) Statistiky – Základní statistiky a tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – zaškrtneme Meze sp. směr. odch., ponecháme implicitní hodnotu 95,00 – Výpočet.

	Popisné statistiky (Tabulka1)	
Proměnná	Spolehlivost Sm.Odch. -95,000%	Spolehlivost Sm.Odch. +95,000%
X	2,233234	8,774739

Dostáváme výsledek: $2{,}23\ g < \sigma < 8{,}77\ g$ s pravděpodobností aspoň 0,95.

Úkol 4.: Testování hypotézy o parametru μ normálního rozložení
Systematická chyba měřicího přístroje se eliminuje nastavením přístroje a měřením etalonu, jehož správná hodnota je $\mu = 10{,}00$. Nezávislými měřeními za stejných podmínek byly získány hodnoty: 10,24 10,12 9,91 10,19 9,78 10,14 9,86 10,17 10,05, které považujeme za realizace náhodného výběru rozsahu 9 z rozložení $N(\mu, \sigma^2)$. Je možné při riziku 0,05 vysvětlit odchylky od hodnoty 10,00 působením náhodných vlivů?
Návod:
Na hladině významnosti 0,05 testujeme hypotézu H_0: $\mu = 10$ proti oboustranné alternativě H_1: $\mu \neq 10$. Jde o úlohu na jednovýběrový t-test. Ten je ve STATISTICE implementován. Vytvoříme datový soubor o jedné proměnné a devíti případech, kam zapíšeme naměřené hodnoty. V Základních statistikách/tabulkách vybereme t-test, samostatný vzorek. Do Referenčních hodnot zapíšeme 10. Ve výstupu se podíváme na hodnotu testového kritéria a na p-hodnotu. Pokud p-hodnota bude menší nebo rovna 0,05, zamítneme hypotézu H_0: $\mu = 10$ ve prospěch oboustranné alternativní hypotézy H_1: $\mu \neq 10$ na hladině významnosti 0,05. V opačném případě H_0 nezamítáme. V našem případě je

	Test průměrů v úči referenční konstantě (hodnotě)							
Proměnná	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
Prom1	10,05111	0,162669	9	0,054223	10,00000	0,942611	8	0,373470

Protože p-hodnota 0,373470 > 0,05 nulovou hypotézu nezamítáme na hladině významnosti 0,05. S rizikem omylu nejvýše 5% lze tedy odchylky od hodnoty 10 vysvětlit působením náhodných vlivů.
Všimněme si ještě hodnoty testového kriteria: $t_0 = 0{,}942611$. Kritický obor

$$W = \langle -\infty, -t_{1-\alpha/2}(n-1)\rangle \cup \langle t_{1-\alpha/2}(n-1), \infty\rangle = \langle -\infty, -t_{0,975}(8)\rangle \cup \langle t_{0,975}(8), \infty\rangle =$$
$$= \langle -\infty, -2{,}306\rangle \cup \langle 2{,}306, \infty\rangle$$

Protože $t_0 \notin W$, nezamítáme na hladině významnosti 0,05 hypotézu H_0.

Úkol 5.: Interval spolehlivosti pro rozdíl parametrů μ_1 - μ_2 dvourozměrného rozložení
Bylo vylosováno 6 vrhů selat a z nich vždy dva sourozenci. Jeden z nich vždy dostal náhodně dietu č. 1 a druhý dietu č. 2. Přírůstky v Dg jsou následující: (62,52), (54,56), (55,49), (60,50), (53,51), (58,50). Za předpokladu, že rozdíly uvedených dvojic tvoří náhodný výběr z normálního rozložení se střední hodnotou μ_1 - μ_2, sestrojte 95% interval spolehlivosti pro rozdíl středních hodnot.
Návod:
Vytvoříme datový soubor o třech proměnných a šesti případech. Do proměnných v1 a v2 zapíšeme naměřené přírůstky, do proměnné v3 uložíme rozdíly v1 - v2.
Ve STATISTICE je implementován výpočet oboustranného intervalu spolehlivosti pro μ, když σ neznáme. Pomocí Popisných statistik zjistíme meze 95% intervalu spolehlivosti pro střední hodnotu proměnné v3 tak, že zaškrtneme Meze spoleh. prům.

	Popisné statistiky	
Proměnná	Int. spolehl. -95,000%	Int. spolehl. +95,000%
Prom3	0,626461	10,70687

Dostaneme výsledek: 0,63 Dg < μ < 10,71 Dg s pravděpodobností aspoň 0,95.

Úkol 6.: Testování hypotézy o rozdíl parametrů μ_1 - μ_2 dvourozměrného rozložení
Pro data z úkolu 5. testujte na hladině významnosti 0,05 hypotézu, že obě výkrmné diety mají stejný vliv.
Návod:
Označme $\mu = \mu_1 - \mu_2$. Na hladině významnosti 0,05 testujeme hypotézu H_0: $\mu = 0$ proti oboustranné alternativě H_1: $\mu \neq 0$. Jde o úlohu na párový t-test. Ten je ve STATISTICE implementován.Vytvoříme datový soubor o dvou proměnných a šesti případech. Do proměnných v1 a v2 zapíšeme naměřené přírůstky. V menu Základní statistiky/tabulky vybereme t-test, závislé vzorky. Zadáme názvy obou proměnných a ve výstupu se podíváme na hodnotu testového kritéria a na p-hodnotu.

	t-test pro závislé vzorky Označ. rozdíly jsou významné na hlad. p < ,05000							
Proměnná	Průměr	Sm.odch.	N	Rozdíl	Sm.odch. rozdílu	t	sv	p
Prom1	57,00000	3,577709						
Prom2	51,33333	2,503331	6	5,666667	4,802777	2,890087	5	0,034183

Protože p-hodnota 0,034183 < 0,05, zamítáme hypotézu H_0: $\mu = 0$ ve prospěch alternativní hypotézy H_1: $\mu \neq 0$ na hladině významnosti 0,05. Znamená to, že jsme s rizikem omylu nejvýše 5% prokázali rozdíl v účinnosti obou výkrmných diet.
Všimněme si ještě hodnoty testového kriteria: $t_0 = 2{,}890087$. Kritický obor

$$W = \langle -\infty, -t_{1-\alpha/2}(n-1)\rangle \cup \langle t_{1-\alpha/2}(n-1), \infty\rangle = \langle -\infty, -t_{0,975}(5)\rangle \cup \langle t_{0,975}(5), \infty\rangle =$$

$$= \langle -\infty, -2{,}5706\rangle \cup \langle 2{,}5706, \infty\rangle$$

Protože $t_0 \in W$, zamítáme na hladině významnosti 0,05 hypotézu H_0.

Příklady k samostatnému řešení

Příklad 1.: Měřením délky deseti válečků byly získány hodnoty (v mm): 5,38 5,36 5,35 5,40 5,41 5,34 5,29 5,43 5,42 5,32. Těchto deset hodnot považujeme za realizace náhodného výběru rozsahu 10 z normálního rozložení $N(\mu, \sigma^2)$.

a) Sestrojte 99% interval spolehlivosti pro neznámou střední hodnotu μ
b) Sestrojte 99% interval spolehlivosti pro neznámou směrodatnou odchylku σ.
c) Na hladině významnosti 0,01 testujte hypotézu, že střední hodnota délky válečků je 5,3 mm proti oboustranné alternativě.

Výsledky:

ad a)

5,3248 mm < μ < 5,4152 mm s pravděpodobností aspoň 0,99

ad b)

0,0272 mm < σ < 0,1002 mm s pravděpodobností aspoň 0,99.

ad c) Testujeme H_0: $\mu = 5{,}3$ proti H_1: $\mu \neq 5{,}3$ na hladině významnosti 0,01. Nulovou hypotézu zamítáme na hladině významnosti 0,01 a přijímáme alternativní hypotézu.

Příklad 2.: Bylo náhodně vybráno 15 desetiletých chlapců a byla zjištěna jejich výška (v cm). Výsledky měření 130, 140, 136, 141, 139, 133, 149, 151, 139, 136, 138, 142, 127, 139, 147 považujeme za realizace náhodného výběru rozsahu 15 z rozložení $N(\mu,\sigma^2)$. Podle názoru odborníků by střední hodnoty výšky desetiletých chlapců měla být 136,1 cm. Testujte tuto hypotézu na hladině významnosti 0,05.

Pomocí N-P plotu a S-W testu ověřte normalitu dat.

Výsledky:

S-W test poskytl p-hodnotu 0,7998, tedy hypotézu o normalitě nezamítáme na hladině významnosti 0,05. Dále testujeme H_0: $\mu = 136{,}1$ proti H_1: $\mu \neq 136{,}1$ na hladině významnosti 0,05. Protože $p = 0{,}0947 > 0{,}05$, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

Příklad 3.: Pět mužů se rozhodlo, že budou hubnout. Zjistili svou hmotnost před zahájením diety a po ukončení diety.

Číslo osoby	1	2	3	4	5
Hmotnost před dietou	84	77,5	91,5	84,5	97,5
Hmotnost po dietě	78,5	73,5	88,5	80	97

Na hladině významnosti 0,05 testujte hypotézu, že dieta neměla vliv na hmotnost.

Výsledky:

Testujeme H_0: $\mu_1 - \mu_2 = 0$ proti H_1: $\mu_1 - \mu_2 \neq 0$. Testová statistika nabývá hodnoty 4,1105, odpovídající p-hodnota je 0,0174, tedy nulovou hypotézu zamítáme na hladině významnosti 0,05. S rizikem omylu nejvýše 5% jsme tedy prokázali, že dieta má vliv na střední hodnotu hmotnosti.