Ověřování normality

Grafický způsob

a) Normální pravděpodobnostní graf (NP-plot)

NP-plot umožňuje graficky posoudit, zda data pocházejí z normálního rozložení.
Způsob konstrukce:
na vodorovnou osu vynášíme uspořádané hodnoty $x_{(1)} \le \dots \le x_{(n)}$,
na svislou osu vynášíme kvantily u_{α_j}, kde $\alpha_j = \dfrac{3j-1}{3n+1}$,
přičemž j je pořadí j-té uspořádané hodnoty (jsou-li některé hodnoty stejné, pak za j bereme průměrné pořadí odpovídající takové skupince).
Pocházejí-li data z normálního rozložení, pak všechny dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ budou ležet na přímce.

Příklad na konstrukci N – P plotu:

Desetkrát nezávisle na sobě byla změřena jistá konstanta. Výsledky měření: 2 1,8 2,1 2,4 1,9 2,1 2 1,8 2,3 2,2. Pomocí normálního pravděpodobnostního grafu posuďte, zda se tato data řídí normálním rozložením.

Řešení:

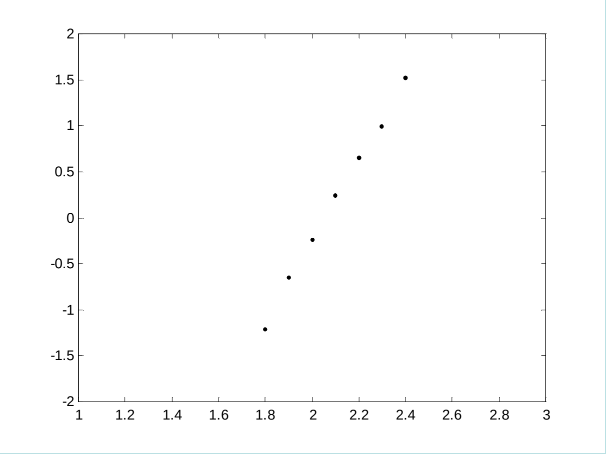
usp. hodnoty	1,8	1,8	1,9	2	2	2,1	2,1	2,2	2,3	2,4
pořadí	1	2	3	4	5	6	7	8	9	10
průměrné pořadí	1,5	1,5	3	4,5	4,5	6,5	6,5	8	9	10

Vektor hodnot průměrného pořadí: j = (1,5 3 4,5 6,5 8 9 10),

vektor hodnot $\alpha_j = \dfrac{3j-1}{3n+1} = (0{,}1129; 0{,}2581; 0{,}4032; 0{,}5968; 0{,}7419; 0{,}8387; 0{,}9355)$,

vektor kvantilů $u_{\alpha_j} = (-1{,}2112; -0{,}6493; -0{,}245; 0{,}245; 0{,}6493; 0{,}9892; 1{,}5179)$.

Normální pravděpodobnostní graf

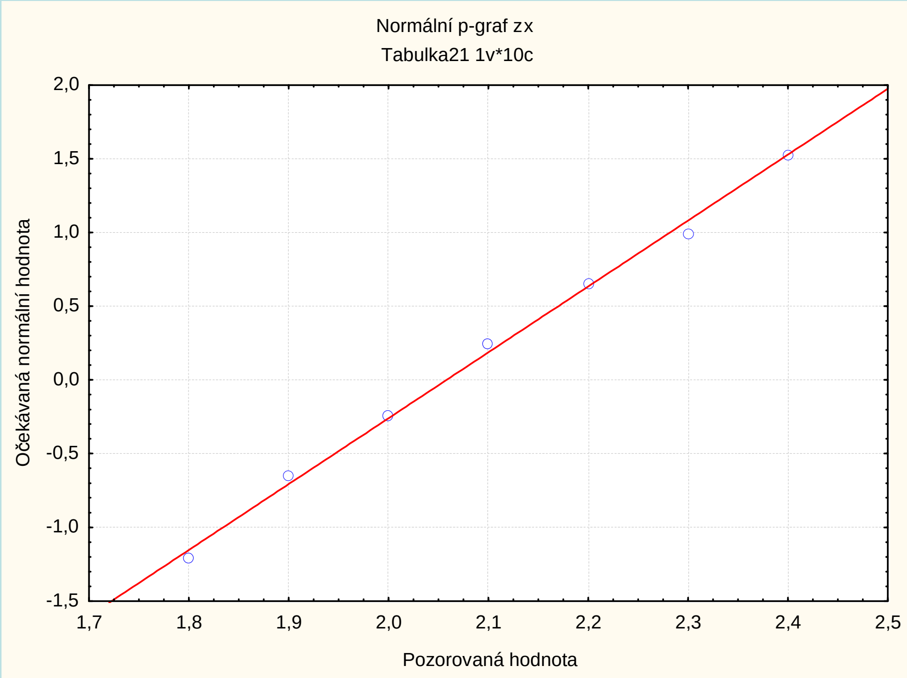


Protože dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ téměř leží na přímce, lze usoudit, že data pocházejí z normálního rozložení.

Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o jedné proměnné a 10 případech. Zjištěné hodnoty zapíšeme do proměnné X.

Grafy – 2D Grafy – Normální pravděpodobnostní grafy – Proměnná X – OK - odškrtneme Neurčovat průměrnou pozici svázaných pozorování - OK.



b) Kvantil-kvantilový graf (Q-Q plot)

Umožňuje graficky posoudit, zda data pocházejí z nějakého známého rozložení (např. systém STATISTICA nabízí 8 typů rozložení: beta, exponenciální, Gumbelovo, gamma, log-normální, normální, Rayleighovo a Weibulovo). Pro nás je nejdůležitější právě normální rozložení.

Způsob konstrukce:

na svislou osu vynášíme uspořádané hodnoty $x_{(1)} \le \ldots \le x_{(n)}$,

na vodorovnou osu kvantily $K_{\alpha_j}(X)$ vybraného rozložení, kde $\alpha_j = \dfrac{j - r_{adj}}{n + n_{adj}}$,

přičemž r_{adj} a n_{adj} jsou korigující faktory $\le 0{,}5$, implicitně $r_{adj} = 0{,}375$ a $n_{adj} = 0{,}25$.

(Jsou-li některé hodnoty $x_{(1)} \le \ldots \le x_{(n)}$ 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.

Body $\left(K_{\alpha_j}(X), x_{(j)}\right)$ se metodou nejmenších čtverců proloží přímka. Čím méně se body odchylují od této přímky, tím je lepší soulad mezi empirickým a teoretickým rozložením.

Příklad na konstrukci Q-Q plotu: Desetkrát nezávisle na sobě byla změřena jistá konstanta. Výsledky měření: 2 1,8 2,1 2,4 1,9 2,1 2 1,8 2,3 2,2. Pomocí Q-Q plotu ověřte, zda se tato data řídí normálním rozložením.

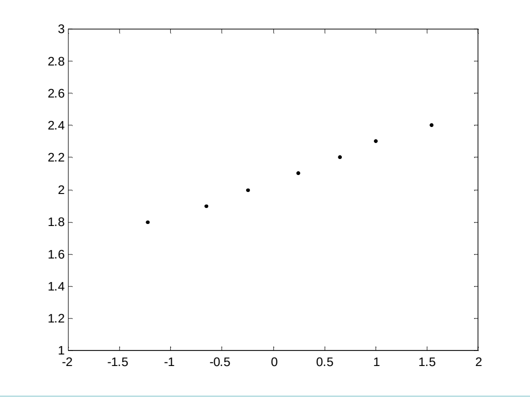
Řešení:

usp.hodnoty	1,8	1,8	1,9	2	2	2,1	2,1	2,2	2,3	2,4
pořadí	1	2	3	4	5	6	7	8	9	10
průměrné pořadí	1,5	1,5	3	4,5	4,5	6,5	6,5	8	9	10

Vektor hodnot průměrného pořadí: j = (1,5 3 4,5 6,5 8 9 10)

vektor hodnot $\alpha_j = \dfrac{j-0{,}375}{n+0{,}25} = (0{,}1098;0{,}2561;0{,}4024;0{,}5976;0{,}7439;0{,}8415;0{,}939)$

vektor kvantilů $u_{\alpha_j} = (-1{,}2278;-0{,}6554;-0{,}247;0{,}247;0{,}6554;1{,}0005;1{,}566)$

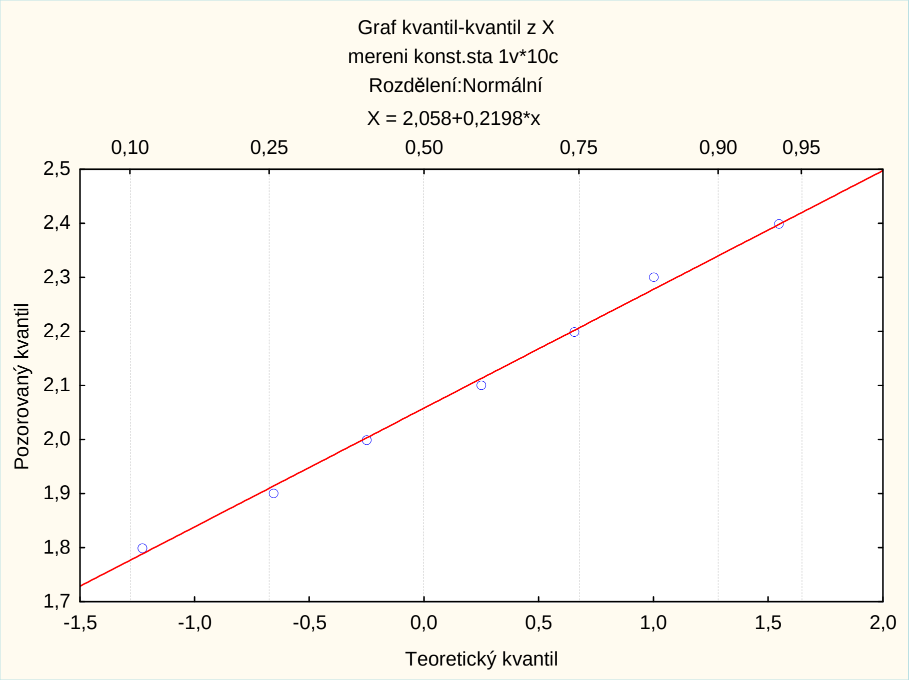


Vzhled grafu nasvědčuje tomu, že data pocházejí z normálního rozložení.

Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o jedné proměnné a 10 případech. Zjištěné hodnoty zapíšeme do proměnné X.

Grafy – 2D Grafy – Grafy typu Q-Q– Proměnná X – OK - odškrtneme Neurčovat průměrnou pozici svázaných pozorování - OK.



c) Histogram

Umožňuje porovnat tvar hustoty četnosti s tvarem hustoty pravděpodobnosti vybraného teoretického rozložení. (Ve STATISTICE je pojem histogramu širší, skrývá se za ním i sloupkový diagram.)

Způsob konstrukce ve STATISTICE:

na vodorovnou osu se vynášejí třídicí intervaly (implicitně 10, jejich počet lze změnit, stejně tak i meze třídicích intervalů) či varianty znaku a na svislou osu absolutní nebo relativní četnosti třídicích intervalů či variant. Do histogramu se zakreslí tvar hustoty (či pravděpodobnostní funkce) vybraného teoretického rozložení. Kromě 8 typů rozložení uvedených u Q-Q plotu umožňuje STATISTICA použít ještě další 4 rozložení: Laplaceovo, logistické, geometrické, Poissonovo.

Příklad na konstrukci histogramu:

U 70 domácností byly zjišťovány týdenní výdaje na nealkoholické nápoje (v Kč).

Výdaje	$\langle 35,65\rangle$	$\langle 65,95\rangle$	$\langle 95,125\rangle$	$\langle 125,155\rangle$	$\langle 155,185\rangle$	$\langle 185,215\rangle$
Počet dom.	7	16	27	14	4	2

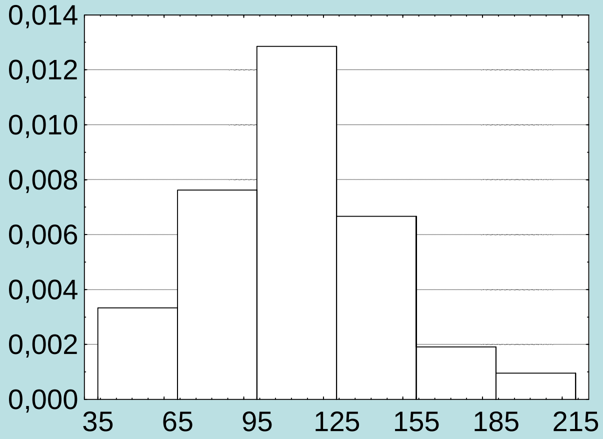
Nakreslete histogram.

Řešení:

Nejprve sestavíme tabulku rozložení četností:

$\langle u_j, u_{j+1}\rangle$	$x_{[j]}$	d_j	n_j	p_j	N_j	F_j	f_j
$\langle 35,65\rangle$	50	30	7	7/70=0,1	7	7/70=0,1	7/2100=0,0033
$\langle 65,95\rangle$	80	30	16	16/70=0,23	23	23/70=0,33	16/2100=0,0076
$\langle 95,125\rangle$	110	30	27	27/70=0,38	50	50/70=0,71	23/2100=0,0109
$\langle 125,155\rangle$	140	30	14	14/70=0,2	64	64/70=0,91	14/2100=0,0067
$\langle 155,185\rangle$	170	30	4	4/70=0,06	68	68/70=0,97	4/2100=0,0019
$\langle 185,215\rangle$	200	30	2	2/70=0,03	70	70/70=1	2/2100=0,00010

S pomocí této tabulky sestrojíme histogram:

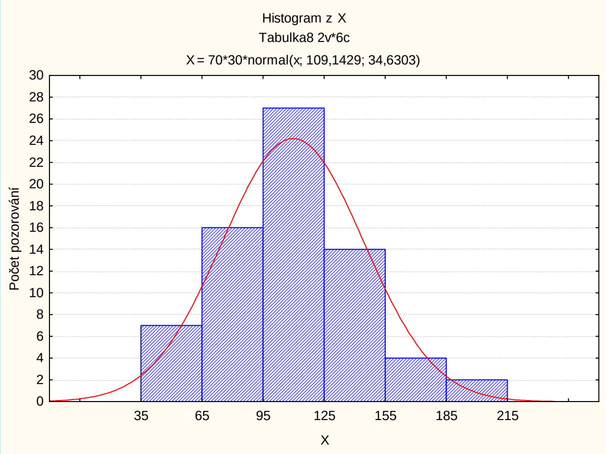


Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o dvou proměnných a 6 případech. První proměnnou nazveme X, druhou cetnost. Do proměnné X napíšeme středy třídicích intervalů, do proměnné cetnost odpovídající absolutní četnosti:

	1 X	2 cetnost
1	50	7
2	80	16
3	110	27
4	140	14
5	170	4
6	200	2

Grafy – Histogramy – zadáme proměnnou vah cetnost – Proměnná X - zaškrtneme Hranice – Určit hranice – zaškrtneme Zadejte hraniční rozmezí: Minimum 35, Krok 30, Maximum 215 – OK – OK. Dostaneme graf:

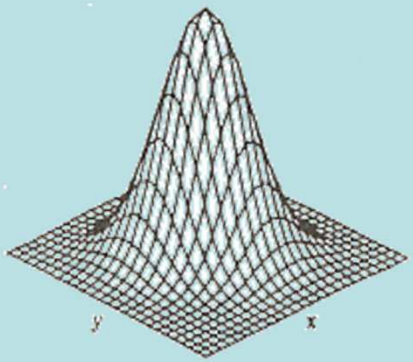


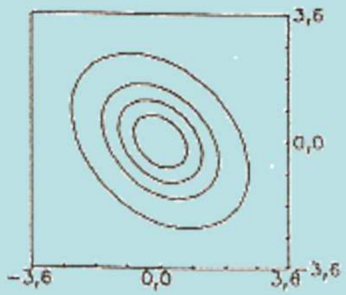
Na rozdíl od histogramu konstruovaného ručně jsou na svislé ose absolutní četnosti, nikoliv četnostní hustoty. V porovnání s grafem hustoty normálního rozložení je vidět, že naše rozložení četností je lehce kladně zešikmené. Naše data tedy nepocházejí z normálního rozložení.

Ověřování dvourozměrné normality pomocí dvourozměrného tečkového diagramu

Máme dvourozměrný datový soubor $(x_1, y_1), \ldots, (x_n, y_n)$, který je realizací dvourozměrného náhodného výběru $(X_1, Y_1), \ldots, (X_n, Y_n)$ z dvourozměrného rozložení. Na vodorovnou osu vyneseme hodnoty x_j, na svislou hodnoty y_k a do příslušných průsečíků nakreslíme tolik teček, jaká je absolutní četnost dvojice (x_j, y_k). Jedná-li se o náhodný výběr z dvourozměrného normálního rozložení, měly by tečky zhruba rovnoměrně vyplnit vnitřek elipsovitého obrazce. Vrstevnice hustoty dvourozměrného normálního rozložení jsou totiž elipsy – viz následující obrázek.

Graf hustoty a vrstevnice dvourozměrného normálního rozložení s parametry $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\rho = -0{,}75$:





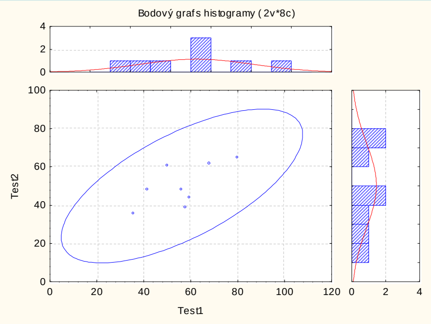
Do dvourozměrného tečkového diagramu můžeme ještě zakreslit $100(1-\alpha)\%$ elipsu konstantní hustoty pravděpodobnosti. Bude-li více než 100α % teček ležet vně této elipsy, svědčí to o porušení dvourozměrné normality. Bude-li mít hlavní osa elipsy kladnou resp. zápornou směrnici, znamená to, že mezi veličinami X a Y existuje určitý stupeň přímé resp. nepřímé lineární závislosti.

Příklad: Máme k dispozici výsledky testů ze dvou předmětů zjištěné u osmi náhodně vybraných studentů určitého oboru.

Číslo studenta	1	2	3	4	5	6	7	8
Počet bodů v 1. testu	80	50	36	58	42	60	56	68
Počet bodů ve 2. testu	65	60	35	39	48	44	48	61

Pomocí dvourozměrného tečkového diagramu se zakreslenou 95% elipsou konstantní hustoty pravděpodobnosti a histogramy pro počty bodů v 1. a 2. testu posuďte, zda tato data lze považovat za realizace náhodného výběru z dvourozměrného normálního rozložení.

Řešení: Vytvoříme nový datový soubor se dvěma proměnnými Test1 a Test2 a osmi případy. Nyní nakreslíme dvourozměrný tečkový diagram: Grafy – 2D Grafy - Bodové grafy s histogramy. V typu proložení pro bodový graf vypneme lineární proložení. Proměnné – X – Test1, Y – Test2 – OK. Dostaneme dvourozměrný tečkový diagram pro vektorovou proměnnou (Test1, Test2) a histogramy pro Test1 a Test2. Nyní do diagramu zakreslíme 95% elipsu konstantní hustoty pravděpodobnosti: 2x klikneme na pozadí grafu a otevře se okno s názvem Vš. možnosti. Vybereme Graf: Elipsa, zvolíme Přidat novou elipsu. Po vykreslení elipsy změníme měřítko: na vodorovné ose bude minimum 0, maximum 120, na svislé ose bude minimum 0, maximum 100. (Stačí 2x kliknout na číselný popis osy a na záložce Měřítka vybrat manuální mód.)



Obrázek svědčí o tom, že předpoklad dvourozměrné normality je oprávněný a že mezi počty bodů z 1. a 2. testu bude existovat určitý stupeň přímé lineární závislosti, tzn., že u studentů, kteří měli vysoký resp. nízký počet bodů v 1. testu, lze očekávat vysoký resp. nízký počet bodů ve 2. testu.

Testy normality dat

K ověřování normality dat slouží celá řada testů, které jsou podrobně popsány ve statistické literatuře. Zde se omezíme na tři testy, které jsou implementovány v systému STATISTICA, a to Kolmogorovův – Smirnovův test a jeho Lilieforsovu variantu, Shapirův – Wilksův test a Andersenův – Darlingův test.

K závěrům těchto testů však přistupujeme s určitou opatrností. Máme-li k dispozici rozsáhlejší datový soubor (orientačně n > 30) a test zamítne na obvyklé hladině významnosti 0,01 nebo 0,05 hypotézu o normalitě, i když vzhled diagnostických grafů svědčí jenom o lehkém porušení normality, nedopustíme se závažné chyby, pokud použijeme statistickou metodu založenou na normalitě dat.

Kolmogorovův – Smirnovův test a jeho Lilieforsova varianta

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, \ldots, X_n$ 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 $D_n = \sup\limits_{-\infty<x<\infty} |F_n(x) - \Phi_T(x)|$.

Nulovou hypotézu zamítáme na hladině významnosti α, když $D_n \geq D_n(\alpha)$, kde $D_n(\alpha)$ je tabelovaná kritická hodnota.

Pro $n \geq 30$ lze $D_n(\alpha)$ aproximovat výrazem $\sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}$.

Shapirův – Wilkův test normality dat

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z normálního rozložení $N(\mu, \sigma^2)$.

Testová statistika má tvar:

$$W = \frac{\sum_{i=1}^{m} a_i^{(n)} \left[X_{(n-i+1)} - X_{(i)}\right]^2}{\sum_{i=1}^{m} (X_i - M)^2},$$

kde m = n/2 pro n sudé a m = (n-1)/2 pro n liché. Koeficienty $a_i^{(n)}$ jsou tabelovány.

Na testovou statistiku W lze pohlížet jako na korelační koeficient mezi uspořádanými pozorováními a jim odpovídajícími kvantily standardizovaného normálního rozložení. V případě, že data vykazují perfektní shodu s normálním rozložením, bude mít W hodnotu 1. Hypotézu o normalitě tedy zamítneme na hladině významnosti α, když se na této hladině neprokáže korelace mezi daty a jim odpovídajícími kvantily rozložení N(0,1).

Lze také říci, že S – W 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 v systému STATISTICA je implementováno jeho rozšíření i na výběry velkých rozsahů, kolem 2000.)

Andersonův – Darlingův test

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, \ldots, X_n$ pochází z normálního rozložení $N(\mu, \sigma^2)$.

Testová statistika má tvar:

$$AD = -\frac{1}{n}\left[\sum_{i=1}^{n}(2i-1)\left\{\ln\Phi\left(\frac{x_{(i)}-m}{s}\right)+\ln\left(1-\Phi\left(\frac{x_{n+1-(i)}-m}{s}\right)\right)\right\}\right]-n\,,$$

kde $x_{(i)}$ jsou vzestupně uspořádané realizace náhodného výběru, Φ je distribuční funkce rozložení $N(0,1)$.

Hypotéza H_0 se zamítá na hladině významnosti α, je-li vypočítaná hodnota testové statistiky AD větší než kritická hodnota $D_{1-\alpha}$. Pro velký rozsah výběru se přibližná 95% kritická hodnota počítá podle vzorce

$$D_{0,95} = 1{,}0348\left(1-\frac{1{,}013}{n}-\frac{0{,}93}{n^2}\right)$$

Příklad:
Jsou dány hodnoty 10, 12, 8, 9, 16. Pomocí Lilieforsova testu, S – W testu a A – D testu testujte na hladině významnosti 0,05 hypotézu, že tato data pocházejí z normálního rozložení.

Řešení:
Vytvoříme nový datový soubor o jedné proměnné nazvané X a pěti případech. Do proměnné X zapíšeme uvedené hodnoty.
Provedení Lilieforsova a S-W testu:
V menu vybereme Statistiky – Základní statistiky/tabulky – Tabulky četností – OK, Proměnné X – OK. Na záložce zvolíme Normalita a zaškrtneme Lilieforsův test a Shapiro – Wilksův W test – Testy normality.

	Testy normality (Tabulka1)				
Proměnná	N	max D	Lilliefors p	W	p
X	5	0,224085	p > .20	0,912401	0,482151

Vidíme, že testová statistika K-S testu je d = 0,22409, odpovídající Lilieforsova p-hodnota je větší než 0,2, tedy hypotézu o normalitě nezamítáme na hladině významnosti 0,05.
Testová statistika S-W testu je W = 0,9124, odpovídající p-hodnota je 0,48215, tedy hypotézu o normalitě nezamítáme na hladině významnosti 0,05.
Provedení A - D testu:
Statistiky – Rozdělení & simulace – proložení dat rozděleními – OK – Proměnné Spojité: X – na záložce Spojité proměnné ponecháme zaškrtnuté pouze Normální, na záložce Možnosti vybereme Anderson – Darling – OK – Souhrnné statistiky rozdělení.

	Souhrn rozdělení for Proměnná: x (Tabulka4)							
	K-S d	K-S p-hodn.	AD stat.	AD p-hodn.	Chí-kvadrát	Chí-kvadr. p-hodn.	Chí-kvadr. SV	Posun (práh/poloha)
Normální (poloha,měřítko)	0,224085	0,915101	0,295219	0,940172				

Testová statistika A – D testu je 0,2952, odpovídající p-hodnota je 0,9402, tedy hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

Parametrické úlohy o jednom náhodném výběru z normálního rozložení

Mnoho náhodných veličin, s nimiž se setkáváme ve výzkumu i praxi, se řídí normálním rozložením. Za jistých předpokladů obsažených v centrální limitní větě se dá rozložení jiných náhodných veličin aproximovat normálním rozložením. Proto je zapotřebí věnovat velkou pozornost právě náhodným výběrům
z normálního rozložení.

Rozložení statistik odvozených z výběrového průměru a rozptylu

Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení $N(\mu, \sigma^2)$. Pak platí

a) $M \sim N(\mu, \frac{\sigma^2}{n})$, tedy $U = \frac{M-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$.

(Pivotová statistika U slouží k řešení úloh o μ, když σ^2 známe.)

b) $K = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n\text{-}1)$.

(Pivotová statistika K slouží k řešení úloh o σ^2, když μ neznáme.)

c) $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$.

(Tato pivotová statistika slouží k řešení úloh o σ^2, když μ známe.)

d) $T = \frac{M-\mu}{\frac{S}{\sqrt{n}}} \sim t(n\text{-}1)$.

(Pivotová statistika T slouží k řešení úloh o μ, když σ^2 neznáme.)

Vysvětlení

ad a) Výběrový průměr M je lineární kombinace náhodných veličin s normálním rozložením, má tedy normální rozložení s parametry E(M) = μ, D(M) = σ^2/n. Statistika U se získá standardizací M.

ad b) Vhodnou úpravou výběrového rozptylu S^2, kde použijeme obrat $X_i - M = (X_i - \mu) - (M - \mu)$, lze statistiku K vyjádřit jako součet kvadrátů n - 1 stochasticky nezávislých náhodných veličin se standardizovaným normálním rozložením. Tento součet se řídí rozložením $\chi^2(n-1)$.

ad c) Tato statistika je součet kvadrátů n stochasticky nezávislých náhodných veličin se standardizovaným normálním rozložením, řídí se tedy rozložením $\chi^2(n)$.

ad d) U ~ N(0, 1), K ~ $\chi^2(n-1)$ jsou stochasticky nezávislé, protože M a S^2 jsou stochasticky nezávislé, tudíž statistika

$$T = \frac{U}{\sqrt{\frac{K}{n-1}}} = \frac{M-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1).$$

Příklad: Hmotnost balíčku krystalového cukru baleného na automatické lince se řídí normálním rozložením se střední hodnotou 1002 g a směrodatnou odchylkou 8 g. Kontrolor náhodně vybírá 9 balíčků z jedné série a zjišťuje, zda jejich průměrná hmotnost je alespoň 999 g. Pokud ne, podnik musí zaplatit pokutu 20 000 Kč. Jaká je pravděpodobnost, že podnik bude muset zaplatit pokutu?

Řešení:

X ~ N(1002, 64), M ~ $N\left(1002, \frac{64}{9}\right)$

$$P(M \le 999) = P\left(\frac{M-1002}{\sqrt{\frac{64}{9}}} \le \frac{999-1002}{\sqrt{\frac{64}{9}}}\right) = P\left(U \le -\frac{9}{8}\right) = \Phi\left(\frac{-9}{8}\right) = 1-\Phi\left(\frac{9}{8}\right) = 1-\Phi(1{,}125) = 1-0{,}87076 = 0{,}12924$$

Pravděpodobnost, že podnik bude platit pokutu, je asi 12,9%.

Řešení pomocí systému STATISTICA:

Využijeme toho, že STATISTICA pomocí funkce INormal(x;mu;sigma) umí vypočítat hodnotu distribuční funkce normálního rozložení se střední hodnotou mu a směrodatnou odchylkou sigma. Tedy $P(M \le 999) = \Phi(999)$, kde Φ je distribuční funkce rozložení N(1002, 64/9).

Otevřeme nový datový soubor o jedné proměnné a jednom případu. Dvakrát klikneme na název proměnné Prom1. Do Dlouhého jména této proměnné napíšeme = INormal(999;1002;8/3).

V proměnné Prom1 se objeví hodnota 0,130295.

Vzorce pro meze 100(1-α)% empirických intervalů spolehlivosti pro μ a σ^2

a) Interval spolehlivosti pro μ, když σ^2 známe (využití pivotové statistiky U)

Oboustranný: $(d, h) = \left(m - \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha/2},\ m + \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha/2}\right)$

Levostranný: $(d, \infty) = \left(m - \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha},\ \infty\right)$

Pravostranný: $(-\infty, h) = \left(-\infty,\ m + \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha}\right)$

b) Interval spolehlivosti pro μ, když σ^2 neznáme (využití pivotové statistiky T)

Oboustranný: $(d, h) = \left(m - \frac{s}{\sqrt{n}}\, t_{1-\alpha/2}(n-1),\ m + \frac{s}{\sqrt{n}}\, t_{1-\alpha/2}(n-1)\right)$

Levostranný: $(d, \infty) = \left(m - \frac{s}{\sqrt{n}}\, t_{1-\alpha}(n-1),\ \infty\right)$

Pravostranný: $(-\infty, h) = \left(-\infty,\ m + \frac{s}{\sqrt{n}}\, t_{1-\alpha}(n-1)\right)$

c) Interval spolehlivosti pro σ^2, když μ neznáme (využití pivotové statistiky K)

Oboustranný: (d, h) = $\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\right)$

Levostranný: (d, ∞) = $\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha}(n-1)}, \infty\right)$

Pravostranný: (-∞, h) = $\left(0, \frac{(n-1)s^2}{\chi^2_{\alpha}(n-1)}\right)$

d) Interval spolehlivosti pro σ^2, když μ známe (využití pivotové statistiky $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}$)

Oboustranný: (d, h) = $\left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}, \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{\alpha/2}(n)}\right)$

Levostranný: (d, ∞) = $\left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{1-\alpha}(n)}, \infty\right)$

Pravostranný: (-∞, h) = $\left(0, \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{\alpha}(n)}\right)$

Příklad: 10 krát nezávisle na sobě byla změřena jistá konstanta μ. Výsledky měření byly: 2 1,8 2,1 2,4 1,9 2,1 2 1,8 2,3 2,2. Tyto výsledky považujeme za číselné realizace náhodného výběru $X_1, ..., X_{10}$ z rozložení $N(\mu, \sigma^2)$, kde parametry μ, σ^2 neznáme. Najděte 95% empirický interval spolehlivosti jak pro μ, tak pro σ^2 a to

a) oboustranný,
b) levostranný,
c) pravostranný.

Řešení: $m = 2,06$, $s^2 = 0,0404$, $s = 0,2011$, $\alpha = 0,05$, $t_{0,975}(9) = 2,2622$, $t_{0,95}(9) = 1,8331$, $\chi^2_{0,975}(9) = 19,023$, $\chi^2_{0,025}(9) = 2,7$, $\chi^2_{0,95}(9) = 16,919$, $\chi^2_{0,05}(9) = 3,325$

ad a) Oboustranný interval spolehlivosti pro střední hodnotu μ

$$d = m - \frac{s}{\sqrt{n}}\, t_{1-\alpha/2}(n-1) = 2,06 - \frac{0,2011}{\sqrt{10}}\, 2,2622 = 1,92$$

$$h = m + \frac{s}{\sqrt{n}}\, t_{1-\alpha/2}(n-1) = 2,06 + \frac{0,2011}{\sqrt{10}}\, 2,2622 = 2,20$$

$1,92 < \mu < 2,20$ s pravděpodobností aspoň 0,95.

Oboustranný interval spolehlivosti pro rozptyl σ^2

$$d = \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} = \frac{9 \cdot 0,0404}{19,023} = 0,0191$$

$$h = \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)} = \frac{9 \cdot 0,0404}{2,7} = 0,1347$$

$0,0191 < \sigma^2 < 0,1347$ s pravděpodobností aspoň 0,95.

ad b) Levostranný interval spolehlivosti pro střední hodnotu μ

$$d = m - \frac{s}{\sqrt{n}} t_{1-\alpha}(n-1) = 2{,}06 - \frac{0{,}2011}{\sqrt{10}} 1{,}8331 = 1{,}94$$

$1{,}94 < \mu$ s pravděpodobností aspoň 0,95.

Levostranný interval spolehlivosti pro rozptyl σ^2

$$d = \frac{(n-1)s^2}{\chi^2_{1-\alpha}(n-1)} = \frac{9 \cdot 0{,}0404}{16{,}919} = 0{,}0215$$

$\sigma^2 > 0{,}0215$ s pravděpodobností aspoň 0,95.

ad c) Pravostranný interval spolehlivosti pro střední hodnotu μ

$$h = m + \frac{s}{\sqrt{n}} t_{1-\alpha}(n-1) = 2{,}06 + \frac{0{,}2011}{\sqrt{10}} 1{,}8331 = 2{,}18$$

$\mu < 2{,}18$ s pravděpodobností aspoň 0,95.

Pravostranný interval spolehlivosti pro rozptyl σ^2

$$h = \frac{(n-1)s^2}{\chi^2_{\alpha}(n-1)} = \frac{9 \cdot 0{,}0404}{3{,}325} = 0{,}1094$$

$\sigma^2 < 0{,}1094$ s pravděpodobností aspoň 0,95.

Řešení pomocí systému STATISTICA:
Vytvoříme nový datový soubor o jedné proměnné X a 10 případech. Do proměnné X napíšeme dané hodnoty.
Statistika – Základní statistiky a tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – zaškrtneme Meze spolehl. prům. a Meze sp. směr. odch. (ostatní volby zrušíme) – pro oboustranný 95% interval spolehlivosti ponecháme implicitní hodnotu pro Interval 95,00, pro jednostranné intervaly změníme hodnotu na 90,00.

Výsledky pro oboustranné 95% intervaly spolehlivosti pro střední hodnotu μ, pro směrodatnou odchylku σ a rozptyl σ^2:

Proměnná	Int. spolehl. -95,000%	Int. spolehl. 95,000	Spolehlivost Sm.Odch. -95,000%	Spolehlivost Sm.Odch. +95,000%	NProm1 =v3 ^2	NProm2 =v4 ^2
X	1,916136	2,203864	0,138329	0,367145	0,019135	0,134795

Vidíme, že
$1,92 < \mu < 2,20$ s pravděpodobností aspoň 0,95,
$0,1383 < \sigma < 0,3671$ s pravděpodobností aspoň 0,95.
$0,0191 < \sigma^2 < 0,1348$ s pravděpodobností aspoň 0,95.

Výsledky pro jednostranné 95% intervaly spolehlivosti pro střední hodnotu μ, pro směrodatnou odchylku σ a rozptyl σ^2:

Proměnná	Int. spolehl. -90,000%	Int. spolehl. 90,000	Spolehlivost Sm.Odch. -90,000%	Spolehlivost Sm.Odch. +90,000%	NProm1 =v3^2	NProm2 =v4^2
X	1,943421	2,176579	0,146678	0,330862	0,021514	0,10947

Vidíme, že

$\mu > 1{,}94$ s pravděpodobností aspoň 0,95,

$\mu < 2{,}20$ s pravděpodobností aspoň 0,95,

$\sigma > 0{,}1467$ s pravděpodobností aspoň 0,95,

$\sigma < 0{,}3309$ s pravděpodobností aspoň 0,95,

$\sigma^2 > 0{,}0215$ s pravděpodobností aspoň 0,95,

$\sigma^2 < 0{,}1095$ s pravděpodobností aspoň 0,95,

Jednotlivé typy testů pro parametry normálního rozložení

a) Nechť $X_1, \ldots, X_n$ je náhodný výběr $N(\mu, \sigma^2)$, kde σ^2 známe. Nechť $n \geq 2$ a c je konstanta. Test H_0: $\mu = c$ proti H_1: $\mu \neq c$ se nazývá jednovýběrový z-test.

b) Nechť $X_1, \ldots, X_n$ je náhodný výběr $N(\mu, \sigma^2)$, kde σ^2 neznáme. Nechť $n \geq 2$ a c je konstanta. Test H_0: $\mu = c$ proti H_1: $\mu \neq c$ se nazývá jednovýběrový t-test.

c) Nechť $X_1, \ldots, X_n$ je náhodný výběr $N(\mu, \sigma^2)$, kde μ neznáme. Nechť $n \geq 2$ a c je konstanta. Test H_0: $\sigma^2 = c$ proti H_1: $\sigma^2 \neq c$ se nazývá test o rozptylu.

Provedení testů o parametrech μ, σ^2 pomocí kritického oboru

a) Provedení jednovýběrového z-testu

Vypočteme realizaci testového kritéria $t_0 = \dfrac{m-c}{\dfrac{\sigma}{\sqrt{n}}}$. Stanovíme kritický obor W. Pokud $t_0 \in$ W, H_0 zamítáme na hladině významnosti α a přijímáme H_1.

Oboustranný test: Testujeme H_0: $\mu = c$ proti H_1: $\mu \neq c$. Kritický obor má tvar: $W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right)$.

Levostranný test: Testujeme H_0: $\mu = c$ proti H_1: $\mu < c$. Kritický obor má tvar: $W = \left(-\infty, -u_{1-\alpha}\right\rangle$.

Pravostranný test: Testujeme H_0: $\mu = c$ proti H_1: $\mu > c$. Kritický obor má tvar: $W = \left\langle u_{1-\alpha}, \infty\right)$.

b) Provedení jednovýběrového t-testu

Vypočteme realizaci testového kritéria $t_0 = \dfrac{m-c}{\dfrac{s}{\sqrt{n}}}$. Stanovíme kritický obor W. Pokud $t_0 \in$ W, H_0 zamítáme na hladině významnosti α a přijímáme H_1.

Oboustranný test: Testujeme H_0: $\mu = c$ proti H_1: $\mu \neq c$. Kritický obor má tvar: $W = \left(-\infty, -t_{1-\alpha/2}(n-1)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-1), \infty\right)$.

Levostranný test: Testujeme H_0: $\mu = c$ proti H_1: $\mu < c$. Kritický obor má tvar: $W = \left(-\infty, -t_{1-\alpha}(n-1)\right\rangle$.

Pravostranný test: Testujeme H_0: $\mu = c$ proti H_1: $\mu > c$. Kritický obor má tvar: $W = \left\langle t_{1-\alpha}(n-1), \infty\right)$.

c) Provedení testu o rozptylu

Vypočteme realizaci testového kritéria $t_0 = \frac{(n-1)s^2}{c}$. Stanovíme kritický obor W. Pokud $t_0 \in$ W, H_0 zamítáme na hladině významnosti α a přijímáme H_1.

Oboustranný test: Testujeme H_0: $\sigma^2 = c$ proti H_1: $\sigma^2 \neq c$. Kritický obor má tvar:.

$$W = \langle 0, \chi^2{}_{\alpha/2}(n-1)\rangle \cup \langle \chi^2{}_{1-\alpha/2}(n-1), \infty)$$

Levostranný test: Testujeme H_0: $\sigma^2 = c$ proti H_1: $\sigma^2 < c$. Kritický obor má tvar: $W = \langle 0, \chi^2{}_{\alpha}(n-1)\rangle$.

Pravostranný test: Testujeme H_0: $\sigma^2 = c$ proti H_1: $\sigma^2 > c$. Kritický obor má tvar: $W = \langle \chi^2{}_{1-\alpha}(n-1), \infty)$.

Příklad: Podle údajů na obalu čokolády by její čistá hmotnost měla být 125 g. Výrobce dostal několik stížností od kupujících, ve kterých tvrdili, že hmotnost čokolád je nižší než deklarovaných 125 g. Z tohoto důvodu oddělení kontroly náhodně vybralo 50 čokolád a zjistilo, že jejich průměrná hmotnost je 122 g a směrodatná odchylka 8,6 g. Za předpokladu, že hmotnost čokolád se řídí normálním rozložením, můžeme na hladině významnosti 0,01 považovat stížnosti kupujících za oprávněné?

Řešení: $X_1, \ldots, X_{50}$ je náhodný výběr z $N(\mu, \sigma^2)$. Testujeme hypotézu
H_0: $\mu = 125$ proti levostranné alternativě H_1: $\mu < 125$. Protože neznáme rozptyl σ^2, použijeme jednovýběrový t-test.

Testové kritérium $\dfrac{m-c}{\frac{s}{\sqrt{n}}} = \dfrac{122-125}{\frac{8,6}{\sqrt{50}}} = -2,4667$.

Kritický obor $W = \left(-\infty, -t_{1-\alpha}(n-1)\right\rangle = \left(-\infty, -t_{0,99}(49)\right\rangle = \left(-\infty, -2,4049\right\rangle$.

Jelikož testové kritérium se realizuje v kritickém oboru, zamítáme nulovou hypotézu na hladině významnosti 0,01. Stížnosti kupujících tedy lze považovat za oprávněné.

Výpočet pomocí systému STATISTICA:
Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma průměry (normální rozdělení) – zaškrtneme Výběrový průměr vs. Střední hodnota a zvolíme jednostr. – do políčka Pr1 napíšeme 122, do políčka SmOd1 napíšeme 8,6, do políčka N1 napíšeme 50, do políčka Pr2 napíšeme 125 - Výpočet. Dostaneme p-hodnotu 0,0086, tedy zamítáme nulovou hypotézu na hladině významnosti 0,01

Náhodný výběr z dvourozměrného rozložení

Nechť $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ je náhodný výběr z dvourozměrného rozložení, přičemž $n \geq 2$. Označíme $\mu = \mu_1 - \mu_2$ a zavedeme rozdílový náhodný výběr $Z_1 = X_1 - Y_1, \ldots, Z_n = X_n - Y_n$, o němž předpokládáme, že se řídí normálním rozložením.

Vypočteme $M = \frac{1}{n}\sum_{i=1}^{n} Z_i$, $S^2 = \frac{1}{n}\sum_{i=1}^{n} (Z_i - M)^2$.

Vzorec pro meze 100(1-α)% empirického intervalu spolehlivosti pro střední hodnotu rozdílového náhodného výběru

Oboustranný: $(d, h) = \left(m - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1),\ m + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1)\right)$

Levostranný: $(d, \infty) = \left(m - \frac{s}{\sqrt{n}} t_{1-\alpha}(n-1),\ \infty\right)$

Pravostranný: $(-\infty, h) = \left(-\infty,\ m + \frac{s}{\sqrt{n}} t_{1-\alpha}(n-1)\right)$

Příklad: Dvěma rozdílnými laboratorními metodami se zjišťoval obsah chemické látky v roztoku (v procentech). Bylo vybráno 5 vzorků a proměřeno oběma metodami. Výsledky měření jsou obsaženy v tabulce:

číslo vzorku	1	2	3	4	5
1. metoda	2,3	1,9	2,1	2,4	2,6
2. metoda	2,4	2,0	2,0	2,3	2,5

Za předpokladu, že data mají normální rozložení, sestrojte 90% empirický interval spolehlivosti pro rozdíl středních hodnot výsledků obou metod.

Řešení:

Přejdeme k rozdílovému náhodnému výběru, jehož realizace jsou: -0,1 -0,1 0,1 0,1 0,1. Vypočteme m = 0,02, s^2 = 0,012, s = 0,109545. Předpokládáme, že tato data pocházejí z normálního rozložení N(μ, σ^2). Vypočteme meze 90% oboustranného intervalu spolehlivosti pro μ při neznámém σ:

$$d = m - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) = 0{,}02 - \frac{0{,}109545}{\sqrt{5}} t_{0{,}95}(4) = 0{,}02 - \frac{0{,}109545}{\sqrt{5}} 2{,}1318 = -0{,}0844$$

$$h = m + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) = 0{,}02 + \frac{0{,}109545}{\sqrt{5}} t_{0{,}95}(4) = 0{,}02 + \frac{0{,}109545}{\sqrt{5}} 2{,}1318 = 0{,}1244$$

-0,0844 < μ < 0,1244 s pravděpodobností aspoň 0,9.

Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor o 3 proměnných a 5 případech. Do 1. proměnné X napíšeme hodnoty pro 1. metodu, do 2. proměnné Y hodnoty pro 2. metodu a do 3. proměnné Z rozdíly mezi X a Y.

Statistiky – Základní statistiky a tabulky – Popisné statistiky, OK - Proměnné Z, Detailní výsledky – zaškrtneme Meze spolehl. Prům. – Interval 90% - Výpočet. Dostaneme tabulku:

	Popisné statistiky (chemicka latka)	
Proměnná	Int. spolehl. -90,000%	Int. spolehl. 90,000
Z	-0,084439	0,124439

Vidíme tedy, že -0,0844 < μ < 0,1244 s pravděpodobností aspoň 0,9.

Párový t-test

Nechť $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ je náhodný výběr z rozložení $N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right)$, přičemž $n \geq 2$. Testujeme H_0: $\mu_1 - \mu_2 = c$ (tj. $\mu = c$) proti H_1: $\mu_1 - \mu_2 \neq c$ (tj. $\mu \neq c$) nebo testujeme nulovou hypotézu proti jedné z jednostranných alternativ. Tento test se nazývá párový t-test.

Provedení párového t-testu

Vypočteme realizaci testového kritéria $t_0 = \dfrac{m - c}{\frac{s}{\sqrt{n}}}$. Stanovíme kritický obor W. Pokud $t_0 \in W$, H_0 zamítáme na hladině významnosti α a přijímáme H_1.

Oboustranný test: Testujeme H_0: $\mu = c$ proti H_1: $\mu \neq c$. Kritický obor má tvar: $W = \left(-\infty, -t_{1-\alpha/2}(n-1)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-1), \infty\right)$.

Levostranný test: Testujeme H_0: $\mu = c$ proti H_1: $\mu < c$. Kritický obor má tvar: $W = \left(-\infty, -t_{1-\alpha}(n-1)\right\rangle$.

Pravostranný test: Testujeme H_0: $\mu = c$ proti H_1: $\mu > c$. Kritický obor má tvar: $W = \left\langle t_{1-\alpha}(n-1), \infty\right)$.

Příklad: V následující tabulce jsou údaje o výnosnosti dosažené 12 náhodně vybranými firmami při investování do mezinárodního podnikání (veličina X) a do domácího podnikání (veličina Y):

č.firmy	1	2	3	4	5	6	7	8	9	10	11	12
X	10	12	14	12	12	17	9	15	9	11	7	15
Y	11	14	15	11	13	16	10	13	11	17	9	19

(Výnosnost je vyjádřena v procentech a představuje podíl na zisku vložených investic za rok.)
Za předpokladu, že data pocházejí z dvourozměrného rozložení a jejich rozdíl se řídí normálním rozložením, na hladině významnosti 0,1 testujte hypotézu, že neexistuje rozdíl mezi střední hodnotou výnosnosti investic do mezinárodního a domácího podnikání proti oboustranné alternativě.
Testování proveďte
a) pomocí intervalu spolehlivosti, b) pomocí kritického oboru.
(Pro úsporu času známe realizace výběrového průměru m = $-1,\bar{3}$ a výběrového rozptylu $s^2 = 4,\overline{78}$ rozdílového náhodného výběru $Z_i = X_i - Y_i$, i = 1, …, 12.)

Řešení:

Testujeme H_0: $\mu = 0$ proti H_1: $\mu \neq 0$

ad a) 90% interval spolehlivosti pro střední hodnotu μ při neznámém rozptylu σ^2 má meze:

$$d = m - \frac{s}{\sqrt{n}} t_{0,95}(n-1) = -1,\bar{3} - \frac{\sqrt{4,\overline{78}}}{\sqrt{12}} 1,7959 = -2,4677$$

$$h = m + \frac{s}{\sqrt{n}} t_{0,95}(n-1) = -1,\bar{3} + \frac{\sqrt{4,\overline{78}}}{\sqrt{12}} 1,7959 = -0,1989$$

Protože číslo c = 0 neleží v intervalu (-2,4677; -0,1989), H_0 zamítáme na hladině významnosti 0,1.

ad b) Vypočítáme realizaci testové statistiky $t_0 = \frac{m-c}{\frac{s}{\sqrt{n}}} = \frac{-1,\bar{3}}{\frac{\sqrt{4,\overline{78}}}{\sqrt{12}}} = -2,11085$

Stanovíme kritický obor $W = \left(-\infty, -t_{0,95}(11)\right\rangle \cup \left\langle t_{0,95}(11), \infty\right) = \left(-\infty, -1,7959\right\rangle \cup \left\langle 1,7959, \infty\right)$

Protože testová statistika se realizuje v kritickém oboru, H_0 zamítáme na hladině významnosti 0,1.

Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor o 2 proměnných a 12 případech. Do 1. proměnné X napíšeme hodnoty pro mezinárodní podnikání, do 2. proměnné hodnoty pro domácí podnikání.

Statistiky – Základní statistiky a tabulky – t-test pro závislé vzorky, OK - Proměnné X, Y – OK – Výpočet. Dostaneme tabulku:

Proměnná	t-test pro závislé vzorky (investovani) Označ. rozdíly jsou významné na hlad. p < ,05000							
	Průměr	Sm.odch.	N	Rozdíl	Sm.odch. rozdílu	t	sv	p
X	11,91667	2,937480						
Y	13,25000	3,048845	12	-1,33333	2,188122	-2,11085	11	0,058490

Vypočtenou p-hodnotu 0,05849 porovnáme se zvolenou hladinou významnosti $\alpha = 0,1$. Protože $p \leq \alpha$, zamítáme nulovou hypotézu na hladině významnosti 0,1.