Cvičení 3: Základní pojmy matematické statistiky

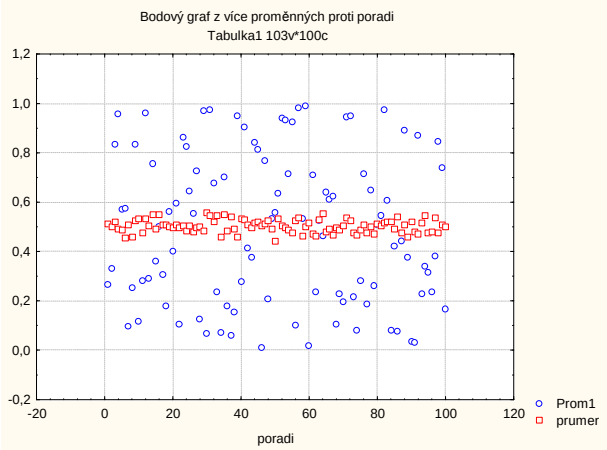
Úkol 1.: Průzkum chování výběrového průměru a výběrového rozptylu

1. Vytvořte nový datový soubor o 103 proměnných a 100 případech. Pomocí programu gener.svb, který si stáhnete z Učebních materiálů, se naplní prvních 100 proměnných 100 realizacemi náhodných veličin $X_i \sim Rs(0,1)$, i = 1, …, 100, do proměnné v101 se uloží pořadová čísla 1 až 100, do proměnné v102 (resp. v103) se uloží průměry (resp. rozptyly) proměnných v1 až v100.
Zdrojový text programu gener.svb:

```
Option Base 1
Sub Main
   Dim s As Spreadsheet
        Set s = ActiveSpreadsheet
        For i = 1 To 100
                s.Variable(i).FillRandomValues
                'do promennych v1 az v100 se ulozi nahodna cisla z intervalu(0,1)
        Next i
        s.VariableLongName(101) = "=v0"
        'do promenne v101 se ulozi poradova cisla 1 az 100
        s.VariableLongName(102) = "=mean(v1:v100)"
        'do promenne v102 se ulozi prumery promennych v1 az v100
        s.VariableLongName(103) = "=stdev(v1:v100)^2"
        'do do promenne v103 se ulozi rozptyly promennych v1 az v100
        s.Recalculate
End Sub
```

(Program se spouští pomocí modré šipky na panelu nástrojů.)
Proměnnou v101 přejmenujte na PORADI, v102 na PRUMER a v103 na ROZPTYL.

2. Graficky znázorněte hodnoty některé z proměnných v1, …, v100 (např. v1) a hodnoty proměnné PRUMER.
Návod: Grafy – Bodové grafy – Typ grafu Vícenásobný – vypneme Lineární proložení – Proměnné X PORADI, Y v1, PRUMER, OK, OK. Vidíme, že hodnoty proměnné v1 se nacházejí mezi 0 a 1, zatímco hodnoty proměnné PRUMER se koncentrují v úzkém pásu kolem 0,5. Znamená to, že průměr funguje jako těžiště dat - eliminuje příliš velké i příliš malé hodnoty.

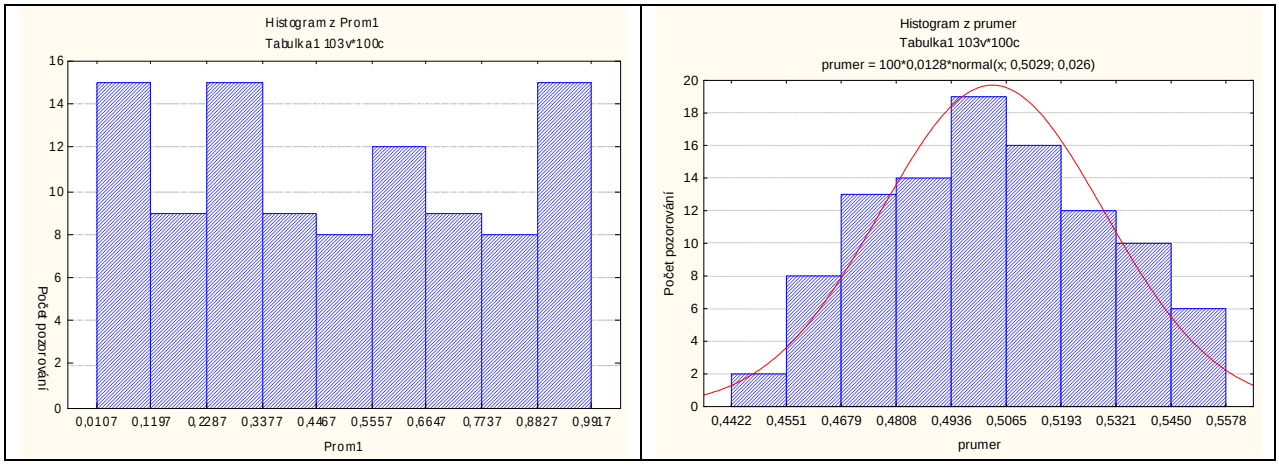


3. Vypočtěte průměr a rozptyl např. proměnné v1 a proměnné PRUMER. Průměr proměnné v1 by měl být blízký 0,5, rozptyl 1/12 = 0,083. Průměr proměnné PRUMER by se měl blížit 0,5, zatímco rozptyl by měl být 100 x menší než 1/12, tj. 0,00083. Dále vypočtěte průměr proměnné ROZPTYL. Měl by se blížit 1/12 = 0,083.

	Popisné statistiky (uniform)	
Proměnná	Průměr	Rozptyl
Prom1	0,536605	0,078676
PRUMER	0,503984	0,000783

	Popisné statistiky (uniform)
Proměnná	Průměr
ROZPTYL	0,083143

4. Nakreslete histogram pro proměnnou v1 a pro proměnnou PRUMER. První histogram se blíží úsečce, druhý Gaussově křivce.



Úkol 2.: Ilustrace nestrannosti výběrové distribuční funkce

1. Vytvořte nový datový soubor o třech proměnných a 1000 případech.

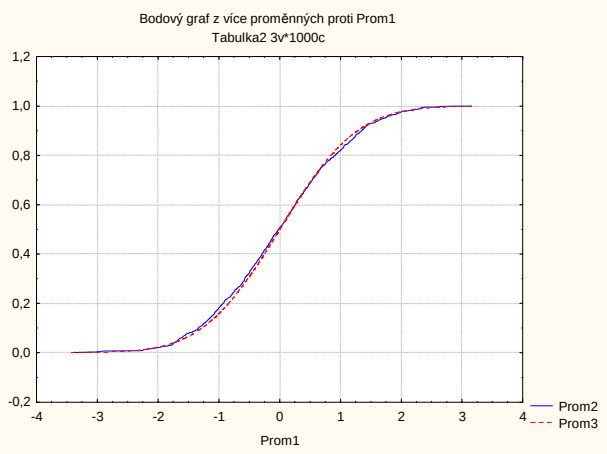
2. Do proměnné v1 uložte 1000 realizací náhodné veličiny s rozložením N(0,1) tak, že v Dlouhém jménu použijte příkaz =vnormal(rnd(1);0;1)

3. Hodnoty proměnné v1 setřiďte podle velikosti: Data - Setřídit.

4. Proměnnou v2 transformujte tak, že v Dlouhém jménu použijte příkaz =v0/1000. Tím získáme hodnoty výběrové distribuční funkce.

5. Do proměnné v3 uložte hodnoty distribuční funkce rozložení N(0,1). Do Dlouhého jména napište příkaz =INormal(v1;0;1)

6. Nakreslete dvourozměrný tečkový diagram, kde na osu x vyneste v1 a na osu y v2 a v3.



Vidíme, že průběh výběrové distribuční funkce $F_{1000}(x)$ (modrá čára) je velmi podobný průběhu distribuční funkce $\Phi(x)$ (červená čára).

7. Postup zopakujte pro rozsah výběru n = 100. Uvidíte, že průběh výběrové distribuční funkce $F_{100}(x)$ se od průběhu distribuční funkce $\Phi(x)$ liší výrazněji.

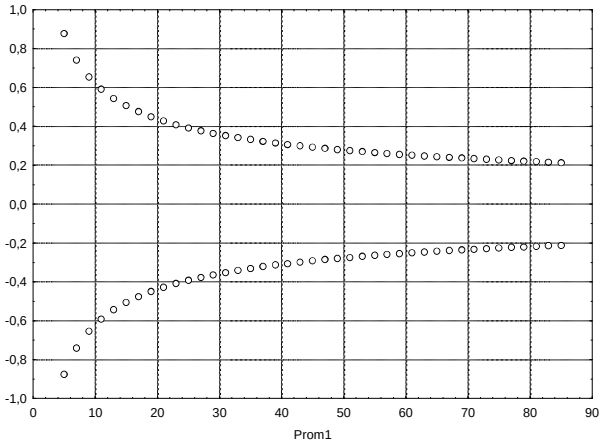
Úkol 3.: Sledování vlivu rozsahu výběru na šířku intervalu spolehlivosti (při α=0,05)

Pro hypotetické náhodné výběry rozsahu n (n = 5, 7, 9, …, 85) z rozložení N(0,1), jejichž výběrové průměry se vždy realizovaly hodnotou 0, vypočtěte dolní a horní meze 95% intervalů spolehlivosti pro μ a graficky znázorněte závislost těchto mezí na rozsahu n.
Upozornění: Meze 100(1-α)% empirického intervalu spolehlivosti pro střední hodnotu při známém rozptylu se počítají podle vzorců: $d = m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}$, $h = m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}$

Návod: Z Učebních materiálů stáhněte program intsp1.svb a otevřete ho v programovacím okně.
Zdrojový text programu intsp1.svb:

```
Option Base 1
Dim s As Spreadsheet
Sub Main
        alfa = 0.05
        'pevně zvolené riziko
        m = 0
        'pevně zvolený průměr
        sigma = 1
        'pevně zvolená směrodatná odchylka
        n = 3
        'počáteční rozsah výběru
        Set s = ActiveSpreadsheet
        For I = 1 To 41
                s.Cells(I, 2) = m - VNormal(1 - alfa / 2, 0, 1) / Sqrt(n + 2 * I)
                'dolní mez intervalu spolehlivosti
                s.Cells(I, 3) = m + VNormal(1 - alfa / 2, 0, 1) / Sqrt(n + 2 * I)
                'horní mez intervalu spolehlivosti
                s.Cells(I, 1) = n + 2 * I
                'zvětšení rozsahu výběru o 2
        Next I
End Sub
```

Vytvořte nový datový soubor o 3 proměnných a 41 případech. Po spuštění programu intsp1 se do proměnné v1 uloží rozsahy výběrů 5, 7, ..., 85, do v2 (resp. v3) dolní (resp. horní) meze 95% intervalů spolehlivosti pro μ. Vytvoření grafu: Grafy – Bodové grafy – Typ grafu Vícenásobný – vypneme Lineární proložení – Proměnné X v1, Y v2, v3 OK, OK.



Vidíme, že šířka intervalu spolehlivosti klesá se zvětšujícím se rozsahem náhodného výběru, zprvu rychle a pak stále pomaleji.

Úkol 4.: Sledování vlivu rizika na šířku intervalu spolehlivosti (při konstantním rozsahu výběru)

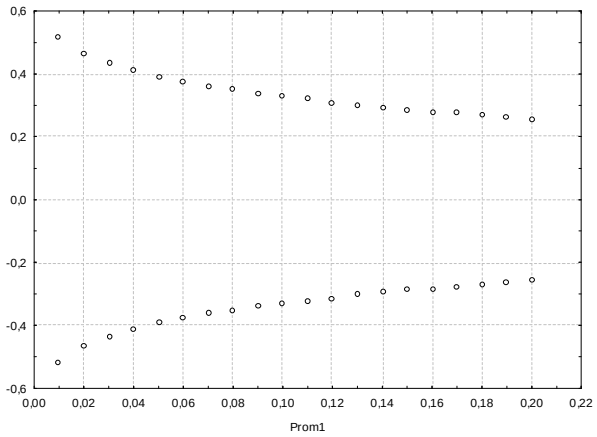
Pro hypotetický náhodný výběr rozsahu n=25 z rozložení N(0,1), jehož výběrový průměr se realizoval hodnotou 0, vypočtěte dolní a horní meze 100(1-α)% intervalů spolehlivosti (α=0,20, 0,19, …, 0,01) pro μ a graficky znázorněte závislost těchto mezí na riziku α.

Návod: Z Učebních materiálů stáhněte program intsp2.svb a otevřete ho v programovacím okně.

Zdrojový text programu intsp2.svb:

```
Option Base 1
Dim s As Spreadsheet
Sub Main
	alfa = 0.21
	'počáteční hodnota rizika
	m = 0
	'pevně zvolený průměr
	sigma = 1
	'pevně zvolená směrodatná odchylka
	n = 25
	'pevně zvolený rozsah výběru
	Set s = ActiveSpreadsheet
	For I = 1 To 20
		s.Cells(I, 2) = m - VNormal(1 - (alfa - I / 100) / 2, 0, 1) / Sqrt(n)
		'dolní mez intervalu spolehlivosti
		s.Cells(I, 3) = m + VNormal(1 - (alfa - I / 100) / 2, 0, 1) / Sqrt(n)
		'horní mez intervalu spolehlivosti
		s.Cells(I, 1) = alfa - I / 100
		'zmenšení rizika o 1/100
	Next I
End Sub
```

Vytvořte nový datový soubor o 3 proměnných a 20 případech. Po spuštění programu intsp2 se do proměnné v1 uloží rizika 0,20, 0,19, ..., 0,01, do v2 (resp. v3) dolní (resp. horní) meze 100(1-α)% intervalů spolehlivosti pro μ. Vytvoření grafu: stejným způsobem jako v předešlém případě.



Vidíme, že šířka intervalu spolehlivosti s rostoucím rizikem klesá.

Úkol 5.: Testování normality

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 A-D testu testujte na hladině významnosti 0,05 hypotézu, že data pocházejí z normálního rozložení. Pomocí N-P grafu posuďte vizuálně předpoklad normality.

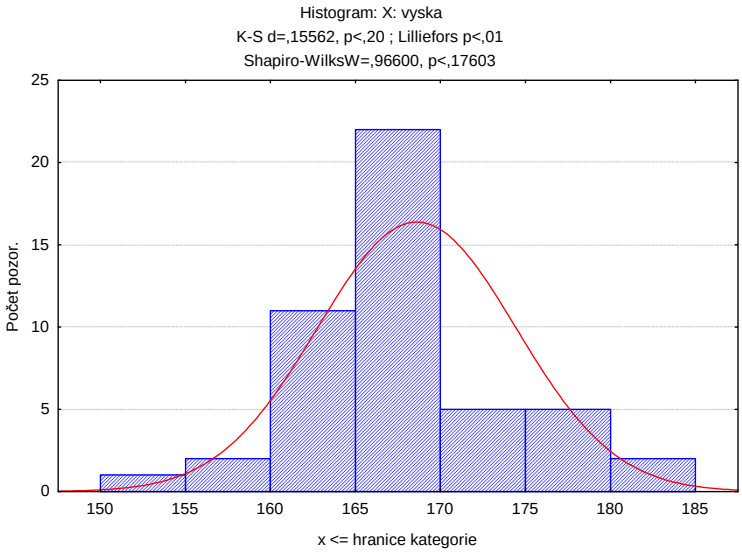
Návod:
1. způsob provedení Lilieforsova a S-W testu: Statistiky – Základní statistiky/tabulky – Tabulky četností – OK – Proměnné X – OK – Normalita – zaškrtneme Lilieforsův test a S-W test – Testy normality.

	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



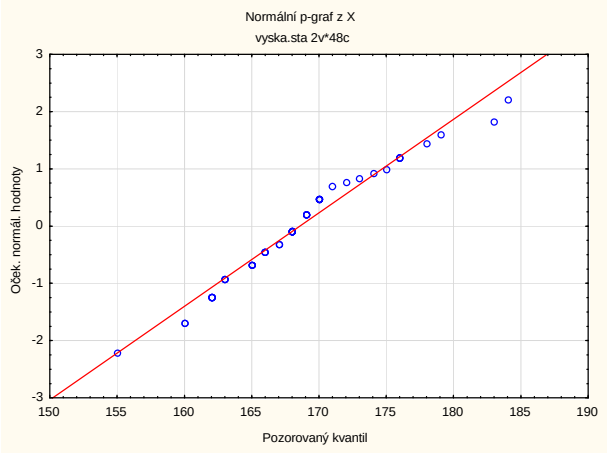
V tomto případě dostaneme v záhlaví tabulky či histogramu stejné informace jako pomocí předešlého způsobu.

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 (vyska.sta)			
	K-S d	K-S p-hodn.	AD stat.	AD p-hodn.
Normální (poloha,měřítko)	0,155621	0,175802	0,660990	0,591425

Testová statistika A – D testu je 0,66099, odpovídající p-hodnota je 0,5914, tedy hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

Grafické ověření normality pomocí N-P grafu – viz úkol 8 ve cvičení 1.



Upozornění: Při vytváření N-P plotu lze zaškrtnout volbu Shapiro- Wilkův test a pak současně s grafem obdržíme i hodnotu testové statistiky a p-hodnotu.

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

Pro kontrolu:

Výsledky Lilieforsova testu a S-W testu pro obor národní hospodářství:

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

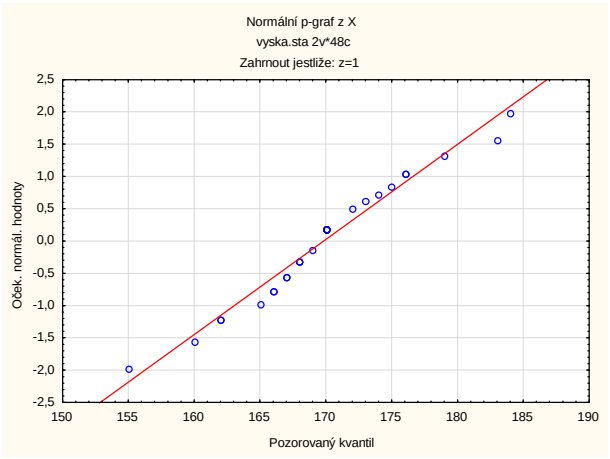
Vidíme, že Lilieforsova varianta K-S testu zamítá hypotézu o normalitě na hladině významnosti 0,05 (p-hodnota je menší než 0,05), zatímco S-W test hypotézu o normalitě nezamítá (p-hodnota je větší než 0,05).

Výsledky A-D testu pro obor národní hospodářství:

	Souhrn rozdělení for Proměnná: X (vyska.sta) Zhrnout podmínku: z=1			
	K-S d	K-S p-hodn.	AD stat.	AD p-hodn.
Normální (poloha,měřítko)	0,167473	0,370570	0,419238	0,828398

Testová statistika A – D testu je 0,4193, odpovídající p-hodnota je 0,8284, tedy hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

N-P plot pro obor národní hospodářství:



Výsledky Lilieforsova testu a S-W testu 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

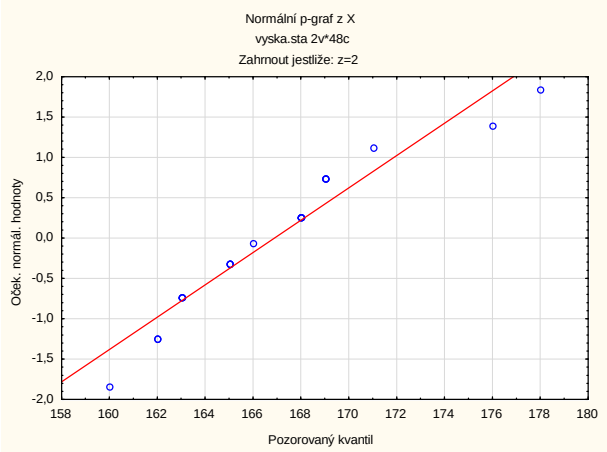
Vidíme, že Lilieforsova varianta K-S testu ani S-W test nezamítají hypotézu o normalitě na hladině významnosti 0,05 (v obou případech je p-hodnota je větší než 0,05).

Výsledky A-D testu pro obor informatika:

	Souhrn rozdělení for Proměnná: X (vyska.sta) Zhrnout podmínku: z=2			
	K-S d	K-S p-hodn.	AD stat.	AD p-hodn.
Normální (poloha,měřítko)	0,172301	0,536360	0,566019	0,678546

Testová statistika A – D testu je 0,5660, odpovídající p-hodnota je 0,6785, tedy hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

N-P plot pro obor informatika:



Úkol 9: Jednovýběrový z-test

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 s neznámou střední hodnotou μ a známou směrodatnou odchylkou $\sigma = 0{,}04$.
Na hladině významnosti 0,1 testujte nulovou hypotézu, že střední hodnota délky válečků je 5,35 mm. Proti nulové hypotéze postavte

a) oboustrannou alternativu
b) levostrannou alternativu
c) pravostrannou alternativu.

Test proveďte pomocí

a) kritického oboru
b) intervalu spolehlivosti
c) p-hodnoty.

Systém STATISTICA použijte jako inteligentní kalkulačku.

Návod:
Formulace nulové hypotézy: H_0: $\mu = 5{,}35$, formulace alternativní hypotézy:
ad a) H_1: $\mu \neq 5{,}35$, ad b) H_1: $\mu < 5{,}35$, ad c) H_1: $\mu > 5{,}35$
Jedná se o jednovýběrový z-test.

Testová statistika $T_0 = \frac{M - c}{\frac{\sigma}{\sqrt{n}}}$ bude mít rozložení N(0, 1), pokud je nulová hypotéza pravdivá.

Provedení testu:

Ad a) Pomocí kritického oboru
Kritický obor pro oboustrannou alternativu:
$W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right)$.
Kritický obor pro levostrannou alternativu:
$W = \left\langle -\infty, -u_{1-\alpha}\right)$.
Kritický obor pro pravostrannou alternativu:
$W = \left\langle u_{1-\alpha}, \infty\right)$.
Pokud $t_0 \in W$, H_0 zamítáme na hladině významnosti α.

Ad b) Pomocí intervalu spolehlivosti
Oboustranný interval spolehlivosti pro μ při známém σ:
$(d, h) = (m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2})$.
Pravostranný interval spolehlivosti pro μ při známém σ:
$(-\infty, h) = (-\infty, m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha})$.
Levostranný interval spolehlivosti pro μ při známém σ:
$(d, \infty) = (m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \infty)$.
Pokud číslo c (v našem případě 5,35) nepatří do 100(1-α)% intervalu spolehlivosti pro μ, H_0 zamítáme na hladině významnosti α.

Ad c) Pomocí p-hodnoty
Vzhledem k tomu, že testová statistika T_0 je spojitá náhodná veličina, můžeme použít úpravu $P(T_0 \geq t_0) = P(T_0 > t_0) = 1 - \Phi(t_0)$.
Vzorec pro výpočet p-hodnoty pro oboustrannou alternativu:
$p = 2 \min\{P(T_0 \leq t_0), P(T_0 \geq t_0)\} = 2 \min\{\Phi(t_0), 1 - \Phi(t_0)\}$.
Vzorec pro výpočet p-hodnoty pro levostrannou alternativu:
$p = P(T_0 \leq t_0) = \Phi(t_0)$.
Vzorec pro výpočet p-hodnoty pro pravostrannou alternativu:
$p = P(T_0 \geq t_0) = 1 - \Phi(t_0)$.
Pokud $p \leq \alpha$, H_0 zamítáme na hladině významnosti α.

Provedení jednovýběrového z-testu v systému STATISTICA

Zjištěné hodnoty zapíšeme do nového datového souboru o 10 případech a jedné proměnné, kterou nazveme X.
Pomocí Popisných statistik spočteme realizaci výběrového průměru: m = 5,37.
Pro pomocné výpočty otevřeme nový datový soubor o jednom případu a deseti proměnných, které nazveme t0, p1, p2, p3, kv1, kv2, d, h, d1, h2.

Do proměnné t_0 uložíme realizaci testové statistiky, a to tak, že do jejího Dlouhého jména napíšeme vzorec pro výpočet testové statistiky:
= (5,37-5,35)/(0,04/sqrt(10)).
Zjistíme, že $t_0 = 1,5811$.

Nyní již můžeme provést test pomocí p-hodnoty.
Do Dlouhého jména proměnné p1 napíšeme vzorec pro výpočet p-hodnoty pro oboustrannou alternativu:
=2*min(INormal(t0;0;1);1-INormal(t0;0;1))
Vypočtená p-hodnota je 0,1138, což je větší než hladina významnosti 0,1 a nulovou hypotézu nelze na této hladině významnosti zamítnout ve prospěch oboustranné alternativy.
Do Dlouhého jména proměnné p2 napíšeme vzorec pro výpočet p-hodnoty pro levostrannou alternativu:
=INormal(t0;0;1)
I tato p-hodnota (0,9431) je větší než 0,1, což znamená, že nulovou hypotézu nelze na hladině významnosti 0,1 zamítnout ve prospěch levostranné alternativy.
Do Dlouhého jména proměnné p3 napíšeme vzorec pro výpočet p-hodnoty pro pravostrannou alternativu:
=1-INormal(t0;0;1)
Vyjde nám 0,0569, tedy na hladině významnosti 0,1 zamítáme nulovou hypotézu ve prospěch pravostranné alternativy. S rizikem omylu nejvýše 10 % jsme prokázali, že střední hodnota délky válečků je větší než 5,35 mm.

Dále provedeme test pomocí kritického oboru, nejprve pro oboustrannou alternativu.
Do proměnné kv1 uložíme kvantil $u_{1-\alpha/2} = u_{0,95}$:
= VNormal(0,95;0;1).
Vyjde nám 1,6449.
Kritický obor pro oboustrannou alternativu je tedy $W = (-\infty, -1,6449\rangle \cup \langle 1,6449, \infty)$.
Vidíme, že testová statistika nepatří do W, což znamená, že H_0 nezamítáme na hladině významnosti 0,1 ve prospěch oboustranné alternativy.
Pro testování nulové hypotézy proti jednostranným alternativám musíme znát kvantil $u_{1-\alpha} = u_{0,9}$. Uložíme ho do proměnné kv2:
= VNormal(0,9;0;1).
Vyjde nám 1,2816.
Kritický obor pro levostrannou alternativu je tedy $W = \langle -\infty, -1,2816)$.
Vidíme, že testová statistika 1,5811 nepatří do W, což znamená, že H_0 nezamítáme na hladině významnosti 0,1 ve prospěch levostranné alternativy.
Kritický obor pro pravostrannou alternativu je tedy $W = \langle 1,2816, \infty)$
Vidíme, že testová statistika1,5811 patří do W, což znamená, že H_0 zamítáme na hladině významnosti 0,1 ve prospěch pravostranné alternativy.

Nakonec provedeme test pomocí intervalu spolehlivosti.
Pro oboustrannou alternativu:
Do Dlouhého jména proměnné d (resp. h) napíšeme vzorec pro dolní (resp. horní) mez oboustranného 90% intervalu spolehlivosti pro μ při známém σ:
=5,37-0,04*kv1/sqrt(10) (resp. =5,37+0,04*kv1/sqrt(10))
Zjistíme, že číslo c = 5,35 patří do intervalu (5,3492; 5,3908), tedy H_0 nezamítáme na hladině významnosti 0,1 ve prospěch oboustranné alternativy.
Pro levostrannou alternativu:

Do Dlouhého jména proměnné h2 napíšeme vzorec pro horní mez pravostranného 90% intervalu spolehlivosti pro μ při známém σ:
=5,37+0,04*kv2/sqrt(10)
Protože 5,35 patří do intervalu (-∞; 5,3862), H_0 nezamítáme na hladině významnosti 0,1 ve prospěch levostranné alternativy.
Pro pravostrannou alternativu:
Do Dlouhého jména proměnné d2 napíšeme vzorec pro dolní mez levostranného 90% intervalu spolehlivosti pro μ při známém σ:
=5,37-0,04*kv2/sqrt(10)
Protože 5,35 nepatří do intervalu (5,3538; ∞), H_0 zamítáme na hladině významnosti 0,1 ve prospěch pravostranné alternativy.