# Téma 8.: Základní pojmy matematické statistiky, ověřování normality

**Příklad 1.:** Ve 12 náhodně vybraných prodejnách ve městě byly zjištěny následující ceny určitého výrobku (v Kč): 102, 99, 106, 103, 96, 98, 100, 105, 103, 98, 104, 107. Těchto 12 hodnot považujeme za realizace náhodného výběru $X_1$, ..., $X_{12}$ z rozložení, které má střední hodnotu $\mu$ a rozptyl $\sigma^2$.
a) Určete nestranné bodové odhady neznámé střední hodnoty $\mu$ a neznámého rozptylu $\sigma^2$.
b) Najděte výběrovou distribuční funkci $F_{12}(x)$ a nakreslete její graf.
**Řešení:**
Vypočteme realizaci výběrového průměru

$$m = \frac{1}{12}(102 + 99 + \ldots + 107) = 101{,}75\,\text{Kč}$$

Vypočteme realizaci výběrového rozptylu:

$$s^2 = \frac{1}{11}\left[(102 - 101{,}75)^2 + (99 - 101{,}75)^2 + \ldots + (107 - 101{,}75)^2\right] = 12{,}39\,\text{Kč}^2$$

Pro usnadnění výpočtu hodnot výběrové distribuční funkce $F_{12}(x)$ uspořádáme ceny podle velikosti: 96, 98, 98, 99, 100, 102, 103, 103, 104, 105, 106, 107.
Číselnou osu rozdělíme na 11 intervalů a v každém intervalu stanovíme hodnotu výběrové distribuční funkce.

$$x < 96 : F_{12}(x) = 0$$

$$96 \le x < 98 : F_{12}(x) = \frac{1}{12} = 0{,}08\overline{3}$$

$$98 \le x < 99 : F_{12}(x) = \frac{3}{12} = 0{,}25$$

$$99 \le x < 100 : F_{12}(x) = \frac{4}{12} = 0{,}\overline{3}$$

$$100 \le x < 102 : F_{12}(x) = \frac{5}{12} = 0{,}41\overline{6}$$

$$102 \le x < 103 : F_{12}(x) = \frac{6}{12} = 0{,}5$$

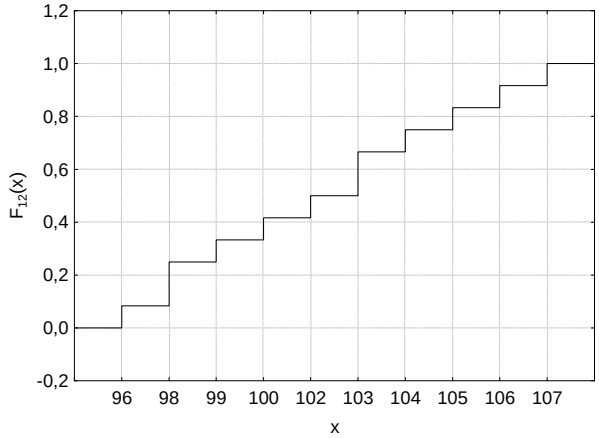
$$103 \le x < 104 : F_{12}(x) = \frac{8}{12} = 0{,}\overline{6}$$

$$104 \le x < 105 : F_{12}(x) = \frac{9}{12} = 0{,}75$$

$$105 \le x < 106 : F_{12}(x) = \frac{10}{12} = 0{,}8\overline{3}$$

$$106 \le x < 107 : F_{12}(x) = \frac{11}{12} = 0{,}91\overline{6}$$

$$x \ge 107 : F_{12}(x) = 1$$




**Výpočet pomocí systému STATISTICA:**
Otevřeme datový soubor ceny_vyrobku.sta.
Výpočet realizace výběrového průměru a výběrového rozptylu:
Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – vybereme Průměr a Rozptyl – Výpočet. Dostaneme tabulku:

| | Popisné statistiky (Tabulka15) | |
|---|---|---|
| Proměnná | Průměr | Rozptyl |
| X | 101,7500 | 12,38636 |

Výpočet hodnot výběrové distribuční funkce:
Statistiky – Základní statistiky/tabulky – Tabulky četností – OK – Proměnné X – OK – Možnosti – ponecháme zaškrtnuté pouze Kumulativní relativní četnosti – Výpočet.
Ke vzniklé tabulce přidáme jeden případ před první případ (do sloupce Kategorie napíšeme 95, do sloupce Kumulativní rel. četnost napíšeme 0 ) a jeden případ za poslední případ (do sloupce Kategorie napíšeme 107, do sloupce Kumulativní rel. četnost napíšeme 100). Proměnnou Kumulativní rel. četnost podělíme 100: do jejího Dlouhého jména napíšeme = v2/100.

Kreslení grafu výběrové distribuční funkce:
Nastavíme se kurzorem na proměnnou Kumulativní rel. četnost, klikneme pravým tlačítkem – Grafy bloku dat – Spojnicový graf: celé sloupce. Ve vytvořeném grafu odstraníme značky, spojnici změníme na schodovitou a upravíme měřítko na vodorovné ose od 1 do 12.

**Příklad 2.:** Přírůstky cen akcií v % na burze v New Yorku u 10 náhodně vybraných společností dosáhly těchto hodnot: 10, 16, 5, 10, 12, 8, 4, 6, 5, 4.
Odhadněte střední hodnotu a směrodatnou odchylku růstu cen akcií a dále odhadněte pravděpodobnost růstu cen akcií aspoň o 8,5 %. Data jsou uložena v souboru akcie_NY.sta.
**Výsledky:** Průměrný růst cen akcií odhadujeme na 8 % se směrodatnou odchylkou 3,97 %. Dále, u 40 % akcií vzrostla cena aspoň o 8,5 %.

**Příklad 3.:** Bylo zkoumáno 9 vzorků půdy s různým obsahem fosforu (veličina X). Hodnoty veličiny Y označují obsah fosforu v obilných klíčcích (po 38 dnech), jež vyrostly na těchto vzorcích půdy.

| číslo vzorku | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| X | 1 | 4 | 5 | 9 | 11 | 13 | 23 | 23 | 28 |
| Y | 64 | 71 | 54 | 81 | 76 | 93 | 77 | 95 | 109 |

Těchto 9 dvojic hodnot považujeme za realizace náhodného výběru $(X_1,Y_1)$, ..., $(X_9,Y_9)$ z dvourozměrného rozložení s kovariancí $\sigma_{12}$ a koeficientem korelace $\rho$. Najděte bodové odhady kovariance $\sigma_{12}$ a koeficientu korelace $\rho$.
**Výpočet pomocí systému STATISTICA:**
Otevřeme datový soubor fosfor_v_kliccich.sta.
Výpočet výběrové kovariance: Statistiky – Vícerozměrná regrese – Proměnné – Závisle proměnná Y, nezávisle proměnná X – OK – OK – Residua/předpoklady/předpovědi – Popisné statistiky – Další statistiky – Kovariance. Dostaneme tabulku:

| | Kovariance (Tabulka18 | |
|---|---|---|
| Proměnná | X | Y |
| X | 91,7500 | 130,0000 |
| Y | 130,0000 | 284,2500 |

Vidíme, že výběrová kovariance veličin X, Y se realizuje hodnotou 130. (Výběrový rozptyl proměnné X resp. Y nabyl hodnoty 91,75 resp. 284,25.)
Výpočet výběrového koeficientu korelace: V menu Další statistiky vybereme Korelace.

| Proměnná | Korelace (Tabulka18 | |
|---|---|---|
| | X | Y |
| X | 1,000000 | 0,804989 |
| Y | 0,804989 | 1,000000 |

Výběrový koeficient korelace veličin X, Y nabyl hodnoty 0,805, tedy mezi veličinami x, Y existuje silná přímá lineární závislost.
Upozornění: Výběrový koeficient korelace lze pomocí systému STATISTICA vypočítat i jiným způsobem: Statistika – Základní statistiky/tabulky – Korelační matice – OK – 1 seznam proměnných – X, Y – OK – Výpočet. Ve výsledné tabulce máme též realizace výběrových průměrů a směrodatných odchylek.

| Proměnná | Korelace (Tabulka18)<br>Označ. korelace jsou významné na hlad. p < ,05000<br>N=9 (Celé případy vynechány u ChD) | | | |
|---|---|---|---|---|
| | Průměry | Sm.odch. | X | Y |
| X | 13,00000 | 9,57862 | 1,000000 | 0,804989 |
| Y | 80,00000 | 16,85972 | 0,804989 | 1,000000 |

**Příklad 4.:** Pět mužů zjistilo a zapsalo svou hmotnost (v kg) a výšku (v cm):

| Číslo muže | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Hmotnost | 76 | 86 | 73 | 84 | 79 |
| Výška | 170 | 177 | 169 | 174 | 175 |

Najděte nestranný bodový odhad rozptylu hmotnosti, rozptylu výšky a kovariance hmotnosti a výšky. Vypočtěte rovněž realizaci výběrového koeficientu korelace hmotnosti a výšky.
**Výsledky:** Výběrový rozptyl hmotnosti se realizuje hodnotou 29,3, výběrový rozptyl výšky 11,5 a výběrová kovariance hmotnost a výšky se realizuje hodnotou 16,5.
Výběrový koeficient korelace hmotnosti a výšky nabývá hodnoty 0,8989.

**Příklad 5.**: Při kontrolních zkouškách životnosti 16 žárovek byl stanoven odhad m = 3000 h střední hodnoty jejich životnosti. Z dřívějších zkoušek je známo, že životnost žárovky se řídí normálním rozložením se směrodatnou odchylkou σ = 20 h. Vypočtěte
a) 99% empirický interval spolehlivosti pro střední hodnotu životnosti
b) 90% levostranný empirický interval spolehlivosti pro střední hodnotu životnosti
c) 95% pravostranný empirický interval spolehlivosti pro střední hodnotu životnosti.
**Upozornění**: Výsledek zaokrouhlete na jedno desetinné místo a vyjádřete v hodinách a minutách.
**Řešení**:
ad a)

$$d = m - \frac{\sigma}{\sqrt{n}} u_{0,995} = 3000 - \frac{20}{\sqrt{16}} 2{,}57583 = 2987{,}1,$$

$$h = m + \frac{\sigma}{\sqrt{n}} u_{0,995} = 3000 + \frac{20}{\sqrt{16}} 2{,}57583 = 3012{,}9$$

2987 h a 6 min < μ < 3012 h a 54 min s pravděpodobností 0,99

**Výpočet pomocí systému STATISTICA**
Otevřeme nový datový soubor o dvou proměnných d, h a jednom případu.
Do Dlouhého jména proměnné d napíšeme vzorec =3000-20/sqrt(16)*VNormal(0,995;0;1)
Do Dlouhého jména proměnné h napíšeme vzorec =3000+20/sqrt(16)*VNormal(0,995;0;1)

ad b)

$$d = m - \frac{\sigma}{\sqrt{n}} u_{0,9} = 3000 - \frac{20}{\sqrt{16}} 1{,}28155 = 2993{,}6$$

2993 h a 36 min < μ s pravděpodobností 0,9

**Výpočet pomocí systému STATISTICA**
Otevřeme nový datový soubor o jedné proměnné d a jednom případu.
Do Dlouhého jména proměnné d napíšeme vzorec =3000-20/sqrt(16)*VNormal(0,9;0;1)

ad c)

$$h = m + \frac{\sigma}{\sqrt{n}} u_{0,975} = 3000 + \frac{20}{\sqrt{16}} 1{,}95996 = 3009{,}8$$

3009 h a 48 min > μ s pravděpodobností 0,95

**Výpočet pomocí systému STATISTICA**
Otevřeme nový datový soubor o jedné proměnné h a jednom případu.
Do Dlouhého jména proměnné h napíšeme vzorec =3000+20/sqrt(16)*VNormal(0,975;0;1)

**Užitečný odkaz:** na adrese http://www.prevody-jednotek.cz je program, s jehož pomocí lze převádět různé fyzikální jednotky, v našem případě hodiny na minuty.

**Příklad 6.:** Při nanášení tenkých kovových vrstev stříbra na polymerní materiál se vyžaduje, aby tloušťka vrstvy byla 0,020 μm. Pomocí atomové absorpční spektroskopie se zjistily hodnoty, jež jsou uvedeny v tabulce a uloženy v souboru vrstva_stribra.sta. Posuďte N-P grafem a Q-Q grafem, zda výsledky měření se řídí normálním rozložením.

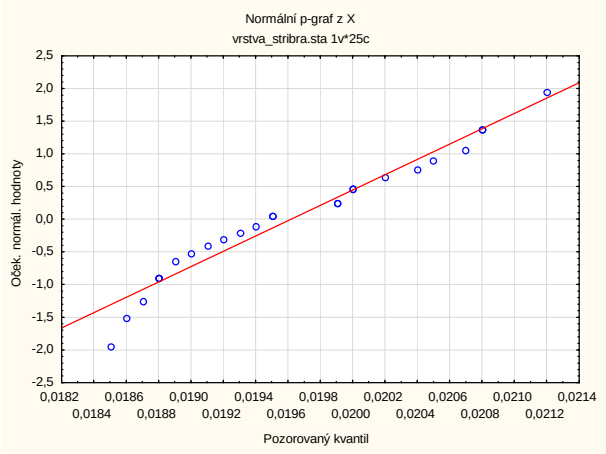
| tloušťka vrstvy | | | | |
|---|---|---|---|---|
| 0,0212 | 0,0186 | 0,0192 | 0,0207 | 0,0200 |
| 0,0200 | 0,0190 | 0,0188 | 0,0208 | 0,0194 |
| 0,0188 | 0,0193 | 0,0204 | 0,0185 | 0,0187 |
| 0,0195 | 0,0191 | 0,0195 | 0,0199 | 0,0205 |
| 0,0189 | 0,0188 | 0,0199 | 0,0202 | 0,0208 |

**Výpočet pomocí systému STATISTICA:**

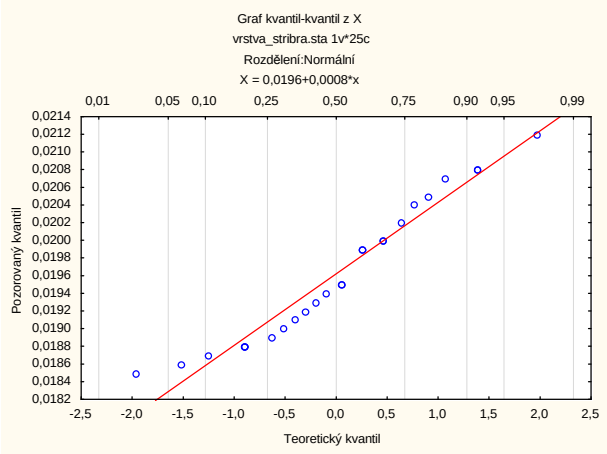
**Vytvoření N-P plotu:** Grafy – 2D Grafy – Normální pravděpodobnostní grafy – Proměnná X – OK - odškrtneme Neurčovat průměrnou pozici svázaných pozorování - OK.

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

N-P plot




Q-Q plot




Dle vzhledu obou diagramů lze soudit, že data vykazují jen lehké odchylky od normality.

**Příklad 7. :** U 48 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:**
**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 |

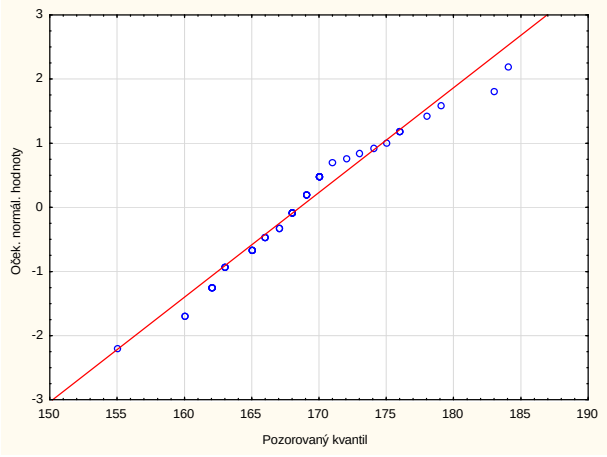
Výstupní tabulka obsahuje počet pozorování, hodnotu testové statistiky Lilieforsovy modifikace K-S testu (max D = 0,155621), p-hodnotu (p < 0,01), testovou statistiku S-W testu (W = 0,965996) a odpovídající p-hodnotu (p = 0,176031). Vidíme, že Lilieforsův test zamítá hypotézu o normalitě na hladině významnosti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

**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. | Chí-kvadrát | Chí-kvadr. p-hodn. | Chí-kvadr. SV | Posun (práh/poloha) |
| Normální (poloha,měřítko) | 0,155621 | 0,175802 | 0,660990 | 0,591425 | 15,37500 | 0,017532 | 6,000000 | |

Vidíme, že Testová statistika A – D testu je 0,661, odpovídající p-hodnota je 0,5914, tedy hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

**Vytvoření N-P grafu:**
**Návod**: Grafy – 2D Grafy – Normální pravděpodobnostní grafy – Proměnné X – OK.




Tečky se řadí podél ideální přímky, normalita je jen lehce porušena.

**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 pro obor národní hospodářství:

| Proměnná | Testy normality (vyska.sta)<br>Zhrnout podmínku: z=1<br>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).

| | Souhrn rozdělení for Proměnná: X (vyska.sta)<br>Zhrnout podmínku: z=1<br>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) | 0,167473 | 0,370570 | 0,419238 | 0,828398 | 2,000000 | 0,157299 | 1,000000 | |

A-D test poskytne hodnotu testové statistiky 0,4192, odpovídající p-hodnota je 0,8284, tedy A-D test nezamítá hypotézu o normalitě na hladině významnosti 0,05.

Výsledky pro obor informatika:

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

| | Souhrn rozdělení for Proměnná: X (vyska.sta)<br>Zhrnout podmínku: z=2<br>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) | 0,172301 | 0,536360 | 0,566019 | 0,678546 | | | | |

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