Téma 7.: Aplikace Moivreovy - Laplaceovy věty. Základní pojmy matematické statistiky.

Moivreova - Laplaceova věta tvrdí, že za určitých podmínek lze binomické rozložení aproximovat standardizovaným normálním rozložením. Aproximace se považuje za vyhovující, když jsou splněny podmínky: $\frac{1}{n+1} < \vartheta < \frac{n}{n+1}$ a $n\vartheta(1-\vartheta) > 9$.
Na základě Moivreovy – Laplaceovy věty se používá aproximativní vzorec, který složitý výpočet distribuční funkce binomického rozložení nahrazuje jednoduchým hledáním v tabulkách hodnot distribuční funkce standardizovaného normálního rozložení.

Máme náhodnou veličinu $Y_n \sim Bi(n, \vartheta)$. Pak

pravděpodobnostní funkce $P(Y_n = y) = \binom{n}{y}\vartheta^y(1-\vartheta)^{n-y}$ pro y = 0, 1, …, n,

distribuční funkce $P(Y \le y) = \sum_{t=0}^{y} P(Y = t) = \sum_{t=0}^{y}\binom{n}{t}\vartheta^t(1-\vartheta)^{n-t}$ - složitý výpočet

Aproximativní vzorec: $P(Y_n \le y) \approx \Phi\left(\frac{y-n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}\right)$.

Příklad 1.: Pravděpodobnost úspěchu při jednom pokusu je 0,3. S jakou pravděpodobností lze tvrdit, že počet úspěchů ve 100 pokusech bude v mezích od 20 do 40?
Řešení:
Y_{100} – počet úspěchů v posloupnosti n = 100 opakovaných nezávislých pokusů, pravděpodobnost úspěchu $\vartheta = 0{,}3$, $Y_{100} \sim Bi(100, 0{,}3)$, $E(Y_{100}) = n\vartheta = 30$, $D(Y_{100}) = n\vartheta(1-\vartheta) = 21$. Vidíme, že podmínky dobré aproximace jsou splněny, protože $\frac{1}{101} < 0{,}3 < \frac{100}{101}$ a $n\vartheta(1-\vartheta) = 21 > 9$.)

Aproximativní výpočet:

$$P(20 \le Y_{100} \le 40) = P(19 < Y_{100} \le 40) = P\left(\frac{19-30}{\sqrt{21}} < \frac{Y_{100}-30}{\sqrt{21}} \le \frac{40-30}{\sqrt{21}}\right) \approx$$
$$\approx \Phi\left(\frac{10}{\sqrt{21}}\right) - \Phi\left(-\frac{11}{\sqrt{21}}\right) = \Phi(2{,}18) - \Phi(-2{,}40) = \Phi(2{,}18) - [1 - \Phi(2{,}40)] =$$
$$= \Phi(2{,}18) + \Phi(2{,}40) - 1 = 0{,}98537 + 0{,}99180 - 1 = 0{,}97717$$

Postup ve STATISTICE:
Otevřeme nový datový soubor o jedné proměnné a jednom případu. Do dlouhého jména této proměnné napíšeme =INormal(10/sqrt(21);0;1)- INormal(-11/sqrt(21);0;1) OK. (Funkce INormal(x;mu;sigma) poskytuje hodnotu distribuční funkce v bodě x normálního rozložení se střední hodnotou mu a směrodatnou odchylkou sigma.) Dostaneme výsledek 0,977263.

Přesný výpočet:

$$P(20 \le Y_{100} \le 40) = \sum_{t=20}^{40}\binom{100}{t}0{,}3^t(1-0{,}3)^{100-t} = 0{,}978614$$

Postup ve STATISTICE:
Otevřeme nový datový soubor o jedné proměnné a jednom případu. Do Dlouhého jména této proměnné napíšeme =IBinom(40;0,3;100)- IBinom(19;0.3;100) – viz Téma 4, 1a) 4. položka. (Funkce IBinom(x;p;n) poskytuje hodnotu distribuční funkce v bodě x binomického rozložení s parametry p a n.)

Příklad 2.: Pravděpodobnost, že zakoupený elektrospotřebič bude vyžadovat opravu během záruční doby, je rovna 0,2. Jaká je pravděpodobnost, že během záruční doby bude nutno ze 400 prodaných spotřebičů opravit více než 96?
Řešení:
$n = 400$, $\vartheta = 0{,}2$, úspěch je nutnost opravy v záruční době
$n\vartheta = 80$, $n\vartheta(1-\vartheta) = 64$
aproximativní výpočet: $P(Y_{400} > 96) \approx$ 1 - INormal(16/8;0;1) = 0,022750
přesný výpočet: $P(Y_{400} > 96) =$ 1 – IBinom(96;0,2;400) = 0,021389

Příklad 3.: Pravděpodobnost narození chlapce je 0,515. Jaká je pravděpodobnost, že mezi 10 000 novorozenci bude
a) stejně nebo více děvčat než chlapců
b) chlapců od 5000 do 5300?

Řešení:
$n = 10000$, $\vartheta = 0{,}515$, úspěch je narození chlapce
$n\vartheta = 5150$, $n\vartheta(1-\vartheta) = 2497{,}75$
Úkol (a)
aproximativní výpočet: $P(Y_{10000} \le 5000) \approx$ INormal(-150/sqrt(2497,75);0;1) = 0,001344
přesný výpočet: $P(Y_{10000} \le 5000) =$ IBinom(5000;0,515;10000) = 0,001391
Úkol (b)
aproximativní výpočet: $P(4999 < Y_{10000} \le 5300) \approx$ INormal(150/sqrt(2497,75);0;1) - INormal(-151/sqrt(2497,75);0;1) = 0,997399
přesný výpočet: $P(4999 < Y_{10000} \le 5300) =$ IBinom(5300;0;1) – IBinom(4999;0;1) = 0,997400

Příklad 4.: Pravděpodobnost, že určitý typ výrobku má výrobní vadu, je 0,05. Jaká je pravděpodobnost, že ze série 1000 výrobků bude mít výrobní vadu nejvýše 70?
Řešení:
$n = 1000$, $\vartheta = 0{,}05$, úspěch je zhotovení vadného výrobku
$n\vartheta = 50$, $n\vartheta(1-\vartheta) = 47{,}5$
aproximativní výpočet: $P(Y_{1000} \le 70) \approx$ INormal(20/sqrt(47,5);0;1) = 0,998145
přesný výpočet: $P(Y_{1000} \le 70) =$ IBinom(70;0,05;1000) = 0,997670

Vlastnosti důležitých statistik odvozených z jednorozměrného náhodného výběru:
Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení se střední hodnotou μ, rozptylem σ^2 a distribuční funkcí $\Phi(x)$. Nechť $n \ge 2$. Označme

$M = \frac{1}{n}\sum_{i=1}^{n} X_i$ výběrový průměr,

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - M)^2$ výběrový rozptyl,

pro libovolné, ale pevně dané $x \in \mathbb{R}$ označme

$F_n(x) = \frac{1}{n}$ počet těch veličin $X_1, \ldots, X_n$, které jsou $\leq x$

hodnotu výběrové distribuční funkce.
Pak pro libovolné hodnoty parametrů μ , σ^2 a libovolné, ale pevně dané reálné číslo x platí:
$E(M) = \mu$,
$E(S_n^2) = \sigma^2$,
$E(F_n(x)) = \Phi(x)$,
Znamená to, že M je nestranným odhadem μ, S^2 je nestranným odhadem σ^2, pro libovolné, ale pevně dané $x \in \mathbb{R}$ je výběrová distribuční funkce $F_n(x)$ nestranným odhadem $\Phi(x)$.

Příklad 5.: 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 $X_1, \ldots, X_{12}$ z rozložení, které má střední hodnotu μ a rozptyl σ^2. Určete nestranné bodové odhady neznámé střední hodnoty μ a neznámého rozptylu σ^2.

Řešení:

Vypočteme realizaci výběrového průměru

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

Vypočteme realizaci výběrového rozptylu:

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

Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o jedné proměnné (nazveme ji X) a 12 případech. Do proměnné X napíšeme zjištěné ceny.
Statistika – 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

Vlastnosti důležitých statistik odvozených z dvourozměrného náhodného výběru:

Nechť $(X_1,Y_1), \ldots, (X_n,Y_n)$ je náhodný výběr z dvourozměrného rozložení s kovariancí σ_{12} a koeficientem korelace ρ. Označme $S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2)$ výběrovou kovarianci a

$R_{12} = \frac{S_{12}}{S_1 S_2}$ výběrový koeficient korelace. Pak pro libovolné hodnoty parametrů σ_{12} a ρ platí:
$E(S_{12}) = \sigma_{12}$,
$E(R_{12}) \approx \rho$ (shoda je vyhovující pro $n \geq 30$).
Znamená to, že výběrová kovariance S_{12} je nestranným odhadem kovariance σ_{12}, avšak výběrový koeficient korelace R_{12} je vychýleným odhadem koeficientu korelace ρ.

Příklad 6.: 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 σ_{12} a koeficientem korelace ρ. Najděte bodové odhady kovariance σ_{12} a koeficientu korelace ρ.

Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o dvou proměnných X a Y 9 případech. Do proměnných X a Y zapíšeme zjištěné hodnoty obsahu fosforu v půdě a v obilných klíčcích.

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 kovarince veličn 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.

	Korelace (Tabulka18)	
Proměnná	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 ořímá 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.

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

Vzorce pro meze 100(1-α)% empirického intervalu spolehlivosti pro střední hodnotu μ při známém rozptylu σ^2:

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

Levostranný: $d = m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}$.

Pravostranný: $h = m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}$.

Příklad 7.: 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 aspoň 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 aspoň 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 aspoň 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.