

Téma 3: Využití systému STATISTICA při řešení příkladů na opakované pokusy a při stanovení prediktivní validity diagnostických testů

1. Opakované nezávislé pokusy

Opakované nezávisle provádíme týž náhodný pokus a sledujeme nastoupení jevu, kterému říkáme úspěch. V každém z těchto pokusů nastává úspěch s pravděpodobností ϑ , $0 < \vartheta < 1$.

a) Binomické rozložení pravděpodobností

Pravděpodobnost, že v prvních n pokusech úspěch nastane právě x -krát ($0 \leq x \leq n$):

$$P_n(x) = \binom{n}{x} \vartheta^x (1 - \vartheta)^{n-x}.$$

K výpočtu v systému STATISTICA slouží funkce Binom(x ; ϑ ; n)

Pravděpodobnost, že v prvních n pokusech úspěch nastane nejvýše x_1 -krát ($0 \leq x_1 \leq n$):

$$\sum_{x=0}^{x_1} P_n(x).$$

K výpočtu v systému STATISTICA slouží funkce IBinom(x_1 ; ϑ ; n)

Pravděpodobnost, že v prvních n pokusech úspěch nastane aspoň x_0 -krát ($0 \leq x_0 \leq n$):

$$\sum_{x=x_0}^n P_n(x).$$

Výpočet lze provést takto: $1 - \text{IBinom}(x_0 - 1; \vartheta; n)$

Pravděpodobnost, že v prvních n pokusech úspěch nastane aspoň x_0 -krát a nejvýše x_1 -krát:

$$\sum_{x=x_0}^{x_1} P_n(x).$$

Výpočet lze provést takto: $\text{IBinom}(x_1; \vartheta; n) - \text{IBinom}(x_0 - 1; \vartheta; n)$

Příklad na binomické rozložení pravděpodobností: Pojišťovna zjistila, že 12% pojistných událostí je způsobeno vloupáním. Jaká je pravděpodobnost, že mezi 30 náhodně vybranými pojistnými událostmi bude způsobeno vloupáním

- nejvýše 6,
- aspoň 6,
- právě 6,
- od dvou do pěti?

Řešení:

Počet pokusů: $n = 30$, pravděpodobnost úspěchu: $\vartheta = 0,12$

ad a)

$$\sum_{x=0}^{x_1} P_n(x) = \sum_{x=0}^6 P_{30}(x) = \sum_{x=0}^6 \binom{30}{x} 0,12^x 0,88^{30-x} = \text{IBinom}(6;0,12;30) = 0,9393$$

S pravděpodobností 93,93% bude mezi 30 náhodně vybranými pojistnými událostmi způsobeno vloupáním nejvýše 6 událostí.

ad b)

$$\sum_{x=x_0}^n P_n(x) = \sum_{x=6}^{30} P_{30}(x) = 1 - \sum_{x=0}^5 P_{30}(x) = 1 - \sum_{x=0}^5 \binom{30}{x} 0,12^x 0,88^{30-x} = 1 - \text{IBinom}(5;0,12;30) = 0,1431$$

S pravděpodobností 14,31% bude mezi 30 náhodně vybranými pojistnými událostmi způsobeno vloupáním aspoň 6 událostí.

ad c)

$$P_n(x) = P_{30}(6) = \binom{30}{6} 0,12^6 0,88^{24} = \text{Binom}(6;0,12;30) = 0,0825$$

S pravděpodobností 8,25% bude mezi 30 náhodně vybranými pojistnými událostmi způsobeno vloupáním právě 6 událostí.

ad d)

$$\sum_{x=x_0}^{x_1} P_n(x) = \sum_{x=2}^5 P_{30}(x) = \sum_{x=0}^5 P_{30}(x) - \sum_{x=0}^1 P_{30}(x) = \sum_{x=0}^5 \binom{30}{x} 0,12^x 0,88^{30-x} - \sum_{x=0}^1 \binom{30}{x} 0,12^x 0,88^{30-x} =$$

$$= \text{IBinom}(5;0,12;30) - \text{IBinom}(1;0,12;30) = 0,7469$$

S pravděpodobností 74,69% bude mezi 30 náhodně vybranými pojistnými událostmi způsobeno vloupáním od 2 do 5 událostí.

Návod: Otevřeme nový datový soubor se čtyřmi proměnnými a o jednom případě.

Do Dlouhého jména 1. proměnné napíšeme =IBinom(6;0,12;30).

Do Dlouhého jména 2. proměnné napíšeme =1-IBinom(5;0,12;30).

Do Dlouhého jména 3. proměnné napíšeme =Binom(6;0,12;30).

Do Dlouhého jména 4. proměnné napíšeme =IBinom(5;0,12;30)-IBinom(1;0,12;30).

Upozornění: Podobným způsobem postupujeme při řešení dalších příkladů

Příklady k samostatnému řešení

Příklad 1.: V rodině je 10 dětí. Za předpokladu, že chlapci i dívky se rodí s pravděpodobností 0,5 a pohlaví se formuje nezávisle na sobě, určete pravděpodobnost, že v této rodině je

- a) právě 5 chlapců
- b) nejméně 3 a nejvýše 8 chlapců.

$n = 10$, úspěch = narození chlapce, pravděpodobnost úspěchu $\vartheta = 0,5$

Výsledek: ad a) 0,246, ad b) 0,935

Příklad 2.: Na dvoukolejném železničním mostě se potkají během 24 hodin nejvýše dva vlaky, a to s pravděpodobností 0,2. Za předpokladu, že denní provoz jsou nezávislé, určete pravděpodobnost, že během týdne se dva vlaky na mostě potkají

- a) právě třikrát
- b) nejvýše třikrát
- c) alespoň třikrát.

$n = 7$, úspěch = setkání dvou vlaků během 24 hodin, pravděpodobnost úspěchu $\vartheta = 0,2$

Výsledek: ad a) 0,115, ad b) 0,967, ad c) 0,148

Příklad 3.: Je pravděpodobnější vyhrát se stejně silným soupeřem tři partie ze čtyř nebo pět partií z osmi, když nerozhodný výsledek je vyloučen a výsledky jsou nezávislé?

Úspěch je výhra partie se stejně silným soupeřem, když remíza je vyloučena, pravděpodobnost úspěchu $\vartheta = 0,5$.

- a) $n = 4$, $x = 3$
- b) $n = 8$, $x = 5$

Výsledek: ad a) 0,25, ad b) 0,219

Příklad 4.: Dvacetkrát nezávisle na sobě házíme třemi mincemi. Jaká je pravděpodobnost, že alespoň v jednom hodě padnou tři líce?

$n = 20$, úspěch je padnutí tří líců při hodu třemi mincemi, $\vartheta = 1/8 = 0,125$,

Výsledek: 0,931

b) Geometrické rozložení pravděpodobností

Pravděpodobnost, že prvnímu úspěchu bude předcházet x neúspěchů:

$$P(x) = (1 - \vartheta)^x \vartheta.$$

K výpočtu v systému STATISTICA slouží funkce $\text{Geom}(x; \vartheta)$

Pravděpodobnost, že prvnímu úspěchu bude předcházet nejvýše x_1 neúspěchů:

$$\sum_{x=0}^{x_1} P(x)$$

K výpočtu v systému STATISTICA slouží funkce $\text{IGeom}(x_1; \vartheta)$

Pravděpodobnost, že prvnímu úspěchu bude předcházet aspoň x_0 neúspěchů:

$$1 - \sum_{x=0}^{x_0-1} P(x)$$

Výpočet lze provést takto: $1 - \text{IGeom}(x_0-1; \vartheta)$

Příklad na geometrické rozložení pravděpodobností: Jaká je pravděpodobnost, že při hře „Člověče, nezlob se!“ nasadíme figurku nejpozději při třetím hoďu?

Řešení:

Počet neúspěchů: $x = 0, 1, 2$, pravděpodobnost úspěchu: $\vartheta = \frac{1}{6}$

$$\sum_{x=0}^2 P(x) = \sum_{x=0}^2 (1 - \vartheta)^x \vartheta = \sum_{x=0}^2 \left(\frac{5}{6}\right)^x \frac{1}{6} = \text{IGeom}(2; 1/6) = 0,4213$$

Pravděpodobnost, že figurku nasadíme nejpozději při třetím hoďu, je 42,13%.

Příklad k samostatnému řešení: Studenti biologie zkoumají barvu očí octomilek. Pravděpodobnost, že octomilka má bílou barvu očí, je 0,25, červenou 0,75. Jaká je pravděpodobnost, že až čtvrtá zkoumaná octomilka má bílou barvu očí?

Počet neúspěchů: $x = 3$, pravděpodobnost úspěchu: $\vartheta = 0,25$

Výsledek: Pravděpodobnost, že až 4. zkoumaná octomilka má bílou barvu očí, je 10,55 %.

2. Opakované závislé pokusy

Hypergeometrické rozložení pravděpodobností

Máme N objektů, mezi nimi je M objektů označeno $0 \leq M \leq N$. Náhodně bez vracení vybereme n objektů ($0 \leq n \leq N$).

Pravděpodobnost, že ve výběru je právě x označených objektů

$$(\max\{0, M - N + n\} \leq x \leq \min\{n, M\}): P_{N,M,n}(x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}.$$

Ve STATISTICE: =Combin(M;x)* Combin(N-M;n-x)/Combin(N;n)

Pravděpodobnost, že ve výběru je nejvýše x_1 označených objektů: $\sum_{x=\max\{0, M-N+n\}}^{x_1} P_{N,M,n}(x)$.

Pravděpodobnost, že ve výběru je aspoň x_0 označených objektů: $\sum_{x=x_0}^{\min\{n, M\}} P_{N,M,n}(x)$.

Příklad na hypergeometrické rozložení pravděpodobností: Koupili jsme 10 cibulek červených tulipánů a 5 cibulek žlutých tulipánů. Zasadili jsme 8 náhodně vybraných cibulek.

- Jaká je pravděpodobnost, že žádná nebude cibulka žlutých tulipánů?
- Jaká je pravděpodobnost, že jsme zasadili všech 5 cibulek žlutých tulipánů?
- Jaká je pravděpodobnost, že aspoň dvě budou cibulky žlutých tulipánů?

Řešení:

Počet objektů: $N = 15$, počet označených objektů: $M = 5$, počet vybraných objektů: $n = 8$

$$\text{ad a) } P_{15,5,8}(0) = \frac{\binom{5}{0} \binom{10}{8}}{\binom{15}{8}} = \frac{\binom{10}{8}}{\binom{15}{8}} = \text{Combin}(10;8) / \text{Combin}(15;8) = 0,007$$

Mezi 8 náhodně vybranými cibulkami se s pravděpodobností 0,7 % nevyskytne žádná cibulka žlutých tulipánů.

$$\text{ad b) } P_{15,5,8}(5) = \frac{\binom{5}{5} \binom{10}{3}}{\binom{15}{8}} = \frac{\binom{10}{3}}{\binom{15}{8}} = \text{Combin}(10;3) / \text{Combin}(15;8) = 0,0186$$

S pravděpodobností 1,86 % bude mezi 8 náhodně vybranými cibulkami právě 5 cibulek žlutých tulipánů.

ad c)

$$1 - P_{15,5,8}(0) - P_{15,5,8}(1) = 1 - \frac{\binom{5}{0} \binom{10}{8}}{\binom{15}{8}} - \frac{\binom{5}{1} \binom{10}{7}}{\binom{15}{8}} = 1 - \frac{\binom{10}{8}}{\binom{15}{8}} - \frac{5 \cdot \binom{10}{7}}{\binom{15}{8}} =$$

$$= 1 - \text{Combin}(10;8) / \text{Combin}(15;8) - 5 * \text{Combin}(10;7) / \text{Combin}(15;8) = 0,8998$$

S pravděpodobností 89,98 % budou mezi 8 náhodně vybranými cibulkami aspoň dvě cibulky žlutých tulipánů.

Příklad k samostatnému řešení:

Dítě dostalo sáček, v němž bylo 5 červených a 5 žlutých bonbónů. Dítě náhodně vybralo ze sáčku 6 bonbónů. Jaká je pravděpodobnost, že mezi vybranými bonbóny budou právě 2 červené?

Výsledek: Pravděpodobnost, že mezi 6 vybranými bonbóny budou právě 2 červené, je 23,8%.

3. Diagnostické testy

Příklad: Provádělo se ověřování kvality nového testu pro diagnostikování jisté poruchy sluchu, která se vyskytuje u 12 % osob v populaci. Test byl ověřován u 1245 osob, u nichž byl stav sluchu vyšetřen již dříve podrobnými klinickými postupy.

Výsledky máme v tabulce:

výsledek testu	skutečnost		celkem
	H (pozitivní)	\bar{H} (negativní)	
A (pozitivní)	a=50	b=300	350
\bar{A} (negativní)	c=25	d=870	895
celkem	75	1170	1245

Vypočtete prediktivní validitu pozitivního i negativního testu.

Výpočet prediktivní validity diagnostického testu provedeme pomocí makra diagnosticky_test.svb.

Otevřeme nový datový soubor o sedmi proměnných a jednom případě.

Proměnná prevalence ... obsahuje pravděpodobnost výskytu sledované vlastnosti.

Proměnná a ... počet pozitivních objektů, které test označil za pozitivní.

Proměnná b ... počet negativních objektů, které test označil za pozitivní.

Proměnná c ... počet pozitivních objektů, které test označil za negativní.

Proměnná d ... počet negativních objektů, které test označil za negativní.

Proměnná PPV ... vypočtená prediktivní validita pozitivního testu.

Proměnná NPV ... vypočtená prediktivní validita negativního testu.

Do proměnné prevalence napíšeme hodnotu 0,12 a do proměnných a, b, c, d postupně 50, 300, 25, 870.

Dále otevřeme makro diagnosticky_test.svb a spustíme ho klávesou F5. V proměnných PPV a NPV se objeví hodnoty 0,2617 a 0,9424.

Znamená to, že pouze u 26,17 % osob, které nový diagnostický test označil za osoby s danou poruchou sluchu, můžeme tuto poruchu očekávat.

Naopak, 94,24 % osob, o nichž test prohlásil, že danou poruchu nemají, danou poruchu skutečně nemá.