Analýza dat v neurologii

VIII. Binomické rozdělení

L. Dušek, T. Pavlík, J. Koptíková
Institut biostatistiky a analýz
Masarykova univerzita, Brno

✉
doc. RNDr. Ladislav Dušek, Dr.
Institut biostatistiky a analýz
Masarykova univerzita, Brno
e-mail: dusek@cba.muni.cz

Tento díl statistického seriálu věnujeme pragmatickému vysvětlení binomického rozdělení, které je využíváno pro modelování diskrétních znaků. Jen pro připomenutí uveďme, že rozdělení pravděpodobnosti vzniká tím, že jsme schopni výsledky opakovaného náhodného experimentu uspořádat na číselné ose a elementárním jevům přiřadíme pravděpodobnost jejich výskytu. Z předchozích dílů by mohl vzniknout nesprávný dojem, že rozdělení pravděpodobnosti se používají převážně pro spojité náhodné veličiny, tedy takové, které mohou nabývat všech reálných hodnot. Avšak naprosto stejný význam mají modelová rozdělení pravděpodobnosti pro diskrétní náhodné veličiny. Pouze formální vyjádření jejich distribuční funkce je odlišné, neboť tyto znaky mohou nabývat pouze diskrétních (od sebe oddělených) hodnot. Jejich distribuční funkce je tedy typicky schodovitá.

Význam pravděpodobnostních rozdělení je u diskrétních znaků stejný jako u znaků spojitých. Známe-li rozdělení, a tím i distribuční funkci, známe tak pravděpodobnost, s jakou může znak nabývat konkrétních hodnot. S využitím příslušného rozdělení můžeme tuto pravděpodobnost dokonce predikovat nebo můžeme simulovat různé situace. Popisem binomického rozdělení neprovádíme žádný virtuální výlet do světa matematiky, toto rozdělení reálně existuje kolem nás a není problém pro něj najít ilustrativní příklady „ze života".

Začněme poměrně strohou definicí. Binomické rozdělení popisuje četnost výskytu náhodného jevu v n nezávislých pokusech, v nichž má tento jev stále stejnou pravděpodobnost, že nastane. Smysl binomického rozdělení není těžké pochopit. Již jsme probrali, že u rozdělení normálního je hodnocena pravděpodobnost různých hodnot, např. výšky postavy. Výška postavy v cm je tedy na ose X. Binomické rozdělení podobně sleduje výskyt nějakého jevu a na ose X, bude tedy vyneseno tolikrát, kolikrát tento jev v opakovaných pokusech nastal. Učebnicovým příkladem je hod mincí, kde sledujeme, zda a kolikrát padne líc. Hodíme-li celkem 5krát *(n = 5)*, pak líc nemusí nutně padnout ani jednou a nejvíce může padnout právě pětkrát. Na ose X budou tedy diskrétní hodnoty 0, 1, 2, 3, 4, 5 a pravděpodobnost, že nastane konkrétní hodnota můžeme zjistit pomocí binomického rozdělení, pokud jsou splněny jeho předpoklady. Jednotlivé hody mincí musí být vzájemně zcela nezávislé a pravděpodobnost nastání sledovaného jevu se v opakovaných pokusech nesmí měnit. U běžné mince v běžných podmínkách je tato pravděpodobnost 0,5 a obecně ji označujeme p, při popisu cílové populace jako π. Hodnota p je parametrem binomického rozdělení a určuje pravděpodobnost nastání jevu v jednotlivých experimentech. Tyto musí být nastaveny tak, aby byla možná již jen jedna další možnost, tedy jev opačný nastávající s pravděpodobností $1 - p$ (někdy je označováno jako $1 - p = q$).

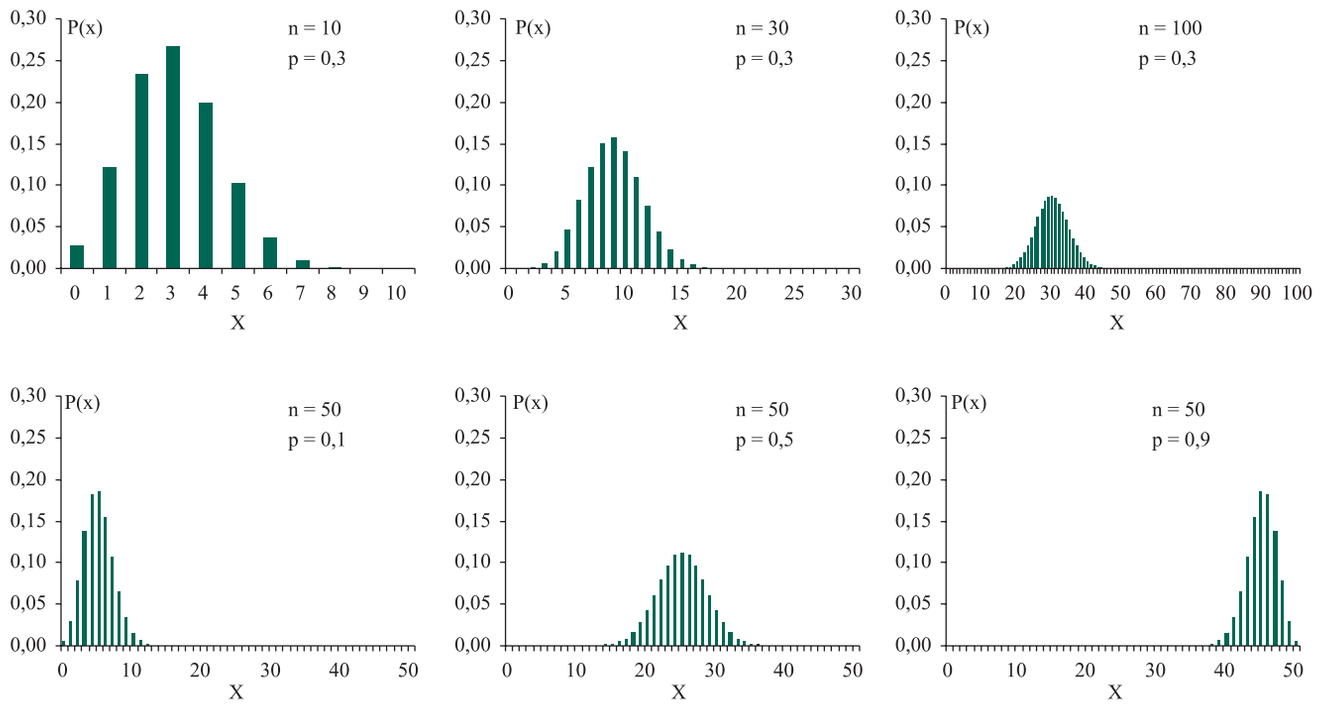
Střední hodnota znaku s binomickým rozdělením je $E(X) = n \times p$, rozptyl je $D(X) = n \times p \times (1 - p)$. Jelikož p může nabývat hodnot pouze od 0 do 1, snadno zjistíme, že D(X) je největší při $p = 0,5$. Paralelou E(X) u normálního rozdělení by byl aritmetický průměr jako střed rozdělení a paralelou D(X) rozptyl hodnot.

S pravděpodobností p nastání jevu někdy v běžných interpretacích ne zcela správně pracujeme. Hodíme-li mincí 10krát, potom střední hodnota binomického rozdělení jednoduše udává nejpravděpodobnější četnost sledovaného jevu (líc), což je při $p = 0,5$ celkem $n \times p = 5$. Ale pozor, to rozhodně neznamená, že nemůže padnou líc mince i 10krát a naopak třeba ani jednou. Každá hodnota znaku X může nastat, i když s jinou pravděpodobností. I kdyby měl sledovaný jev pravděpodobnost nastání v jednotlivém experimentu například 0,95, je s určitou, byť velmi malou, pravděpodobností možné, že se v 10krát opakovaném experimentu neobjeví ani jednou. Konkrétně tato pravděpodobnost je při $n = 10$ a $p = 0,95$ rovna $P(X = 0) = 9,8 \times 10^{-14}$.

A naopak relativně málo pravděpodobný jev může nastat i opakovaně za sebou. Abychom tyto pravděpodobnosti byli schopni spočítat, musíme použít vztah uvedený na obr. 1. Doufejme, že poněkud složitější vzorec čtenáře neodradí, jeho výpočet je jednoduchý a vyžaduje pouze aplikaci faktoriálů hodnot zahrnutých v kombinačním čísle. Výpočet je rozepsán pro příklad 5krát opakovaného experimentu v tab. 1. Obr. 1 dále dokládá tyto skutečnosti:

- binomický znak X vyjadřuje četnost sledovaného jevu v n opakováních a nabývá hodnot od 0 do n
- čím vícekrát opakujeme experiment, tím menší relativní podíl připadá na jednotlivé hodnoty X, neboť všechny dohromady musí nastat s pravděpodobností 1, nebo kumulativně 100 %

$$P(x)=\binom{n}{x}\cdot p^{x}\cdot(1-p)^{(n-x)}=\frac{n!}{x!(n-x)!}\cdot p^{x}\cdot(1-p)^{(n-x)}$$



Obr. 1. Binomické rozložení jako model.

- rozdělení s *p = 0,5* je symetrické kolem středu osy X, menší nebo větší *p* posouvá střed rozdělení směrem k limitním hodnotám, tedy směrem k hodnotě 0 nebo *N*
- hodnota *p = 0,1* neznamená automaticky 10% výskyt daného jevu, všechny hodnoty od 0 do *n* mohou nastat, i když s různou pravděpodobností.

Příklad na obr. 1 usnadní i pochopení dvou extrémních případů. První je v praxi klinického výzkumu málo využitelný a jde o situaci, kdy experiment opakujeme pouze jednou. Pokud je speciálně *n = 1*, jde o tzv. alternativní rozdělení a daný jev buď nenastane, nebo nastane jednou. Druhým extrémem je, že náhodný experiment opakujeme mnohokrát. Tedy máme velmi velký počet jedinců, objektů nebo situací, u kterých hodnotíme (vždy u jednoho po druhém), zda jev nastal nebo ne. Na obr. 1 vidíme, že rozložení za této situace začne vytvářet téměř spojitý obrázek. V takovém případě práce s výsledky vyžaduje aproximaci na vhodný typ spojitého rozdělení, např. při *p = 0,5* aproximaci na normální rozdělení. Při *n = 1 000* jistě nemá smysl ptát se na pravděpodobnost, že se jev objeví přesně například u 897 jedinců. Osa X bude přirozeně tříděna do intervalů

Tab. 1. Pravděpodobnost četnosti jevu dle binomického rozdělení.

X: četnost jevu	Tři příklady výpočtu P(X)* pro různé hodnoty p při n = 5 (5krát opakovaný experiment)		
	p = 0,1	p = 0,5	p = 0,9
0	0,59049	0,03125	0,00001
1	0,32805	0,15625	0,00045
2	0,07290	0,31250	0,00810
3	0,00810	0,31250	0,07290
4	0,00045	0,15625	0,32805
5	0,00001	0,03125	0,59049
Celkem	1	1	1

** Vztah pro výpočet P(X) je uveden na obr. 1.*

Příklad: sledování výskytu jevu v anamnéze u n = 100 jedinců. Přítomnost jevu byla zachycena u 60 jedinců, tedy u 60 %.

Bodový odhad hodnoty p:

$$\hat{p} = 0{,}6$$

95% interval spolehlivosti:

$$\hat{p} - Z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} \le \pi \le \hat{p} + Z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}$$

$$0{,}6 - 1{,}96 \cdot 0{,}049 \le \pi \le 0{,}6 + 1{,}96 \cdot 0{,}049$$

$$0{,}503 \le \pi \le 0{,}697$$

Obr. 2. Výpočet intervalu spolehlivosti pro parametr π binomického rozložení.

Spodní hranice intervalu:

$$L_1 = \frac{r}{r + (n - r + 1) \cdot F_{1-\alpha/2}^{(\nu_1;\nu_2)}} \qquad \text{.....kde:} \quad \nu_1 = 2(n - r + 1);\ \nu_2 = 2r$$

Horní hranice intervalu:

$$L_2 = \frac{(r+1) \cdot F_{1-\alpha/2}^{(\nu_1';\nu_2')}}{n - r + (r+1) \cdot F_{1-\alpha/2}^{(\nu_1';\nu_2')}} \qquad \text{.....kde:} \quad \nu_1' = 2(r+1) = \nu_2 + 2;\ \nu_2' = 2(n-r) = \nu_1 - 2$$

Obr. 3. Výpočet intervalu spolehlivosti pro parametr binomického rozdělení bez aproximace na normální rozdělení – vztahy.

Příklad: náhodný vzorek n = 200 jedinců. Zjištěni 4 jedinci se sledovaným znakem.

Bodový odhad hodnoty π:

$$p = {}^{4}\!/_{200} = \underline{\underline{0{,}02}}$$

95% interval spolehlivosti:

Spodní hranice	Horní hranice
$\nu_1 = 2(n - r + 1) = 2(200 - 4 + 1) = 394$	$\nu_1' = 2(r+1) = 10$
$\nu_2 = 2r = 2 \cdot 4 = 8$	$\nu_2' = 2(n-r) = 2(200-4) = 392$
$F_{1-\alpha/2}^{(394;8)} = \underline{\underline{3{,}67}}$	$F_{1-\alpha/2}^{(10;392)} = \underline{2{,}08}$
$L_1 = \frac{4}{4 + (200 - 4 + 1) \cdot 3{,}67} = \underline{\underline{0{,}0055}}$	$L_2 = \frac{(4+1) \cdot 2{,}08}{200 - 4 + (4+1) \cdot 2{,}08} = \underline{0{,}050}$

Obr. 4. Výpočet intervalu spolehlivosti pro parametr binomického rozdělení bez aproximace na normální rozdělení – příklad.

a interpretace i grafické vyjádření bude připomínat práci s rozdělením hodnot spojité veličiny.

Binomické rozdělení je užitečným nástrojem, popisujícím velké množství reálných situací. Známe-li hodnotu *p* a *n*, můžeme rovnici na obr. 1 pohodlně používat k velmi cenným výpočtům. Jako příklad uveďme tyto:

- výpočet pravděpodobnosti nastání libovolného počtu jevů, od 0 až do maxima *n*
- simulační výpočet, jak se změní pravděpodobnost při snížení nebo zvýšení *n*
- simulační výpočet, jak se mění tvar rozdělení při změně *p*.

V experimentální praxi je *p* často neznámé a sledování provádíme právě proto, abychom ho odhadli. I v případě binomického rozdělení odhadujeme hodnotu v cílové populaci π, a to jednak bodovým odhadem, anebo pomocí intervalu spolehlivosti. Příklad výpočtu 95% intervalu spolehlivosti uvádí obr. 2. Interval má stejnou interpretaci jako např. u odhadu aritmetického průměru normálního rozdělení. Udává pravděpodobnostní hranice výskytu opakovaně odhadovaných hodnot π. Pokud bychom odhad opakovali 100krát při stejném *n*, pak by se získané odhady p_1, p_2 až p_{100} celkem 5krát mohly dostat mimo hranice 95% intervalu spolehlivosti. Interval spolehlivosti informuje čtenáře o stupni nejistoty při odhadu pravděpodobnosti nastání jevu. Pro další výpočty pak můžeme podle okolností použít i hranice intervalu místo vlastní hodnoty *p*. Například při odhadu pravděpodobnosti škodlivého jevu lze při analýze rizik pracovat s horní hranicí intervalu spolehlivosti jako s bezpečnějším údajem. Snižujeme tak pravděpodobnost podhodnocení rizika.

Z výpočtu na obr. 2 vidíme všechny obecně platné vlastnosti intervalů spolehlivosti, z těch hlavních uveďme následující:

- je použit 97,5% kvantil standardizovaného normálního rozložení (Z), počítáme tedy oboustranný 95% interval. Pokud bychom použili kvantil s nižší pravděpodobnostní hodnotou, byl by číselně menší než $Z_{0{,}975}$ = 1,96 (např.

$Z_{0,95}$ = 1,645 by vedl k 90% intervalu) a interval by se numericky zužoval. A naopak dosazením $Z_{0,995}$ = 2,576 bychom hranice rozšířili a získali bychom 99% interval

- šířka intervalu závisí na hodnotě *p*, přičemž nejširší bude pro *p = 0,5*
- šířka intervalu logicky závisí na hodnotě *n*, čím větší *n*, tím přesnější je odhad π, a tím je interval užší.

Výpočet intervalu spolehlivosti na obr. 2 je založen na předpokladu, že odhad *p* se chová podle modelu normálního rozložení. Proto je ve vztahu používán kvantil standardizovaného normálního rozložení, a proto je interval symetrický. Avšak tento předpoklad samozřejmě nebude platit u velmi nízkých a vysokých hodnot *p*, neboť ty jsou ohraničeny hodnotou 0 nebo 1. Například, naměříme-li při malém *n = 10* hodnotu *p = 0,2* nebo nižší, budeme mít problémy se spodní hranicí intervalu spolehlivosti. Z tohoto důvodu uvádíme na obr. 3 a 4 vztahy pro asymetrický interval spolehlivosti, kde je používáno tzv. Fisher-Snedecorovo rozdělení. Toto rozdělení má kvantily mezinárodně označované jako F a má dva druhy stupňů volnosti, které jsou označovány jako $\nu 1$ a $\nu 2$. Hodnoty kvantilů F pro danou dvojici hodnot $\nu 1$ a $\nu 2$ lze nalézt v tabulkách, statistickém software nebo i v MS Excel. Výpočet na obr. 3 a 4 umožní ošetřit intervalem spolehlivosti odhady parametru π blízké limitním hodnotám 0 nebo 1.

Další možností, jak pracovat s binomickým rozdělením při velmi nízkých hodnotách p, je aproximace na tzv. Poissonovo rozdělení. Tato aproximace je efektivní při *n > 30* a *p < 0,1*. Poissonovo rozdělení umožňuje modelování výskytu vzácných a velmi vzácných jevů. Tomuto tématu bude věnován další díl našeho seriálu.