Statistické metody a zpracování dat

III. Pravděpodobnost, teoretická rozdělení

Petr Dobrovolný

Pravděpodobnost, náhodná proměnná

- Popisné a průzkumové metody umožňují přehledné shrnutí informací, které se týkají jen objektů měřených či pozorovaných.
- Činíme závěry pouze z určitého zpracovávaného souboru – **výběrového**, popisujeme jen to, co bylo zjištěno, naměřeno.
- S popisnou statistikou často nevystačíme, potřebujeme činit **zobecňující úsudky**.

Příklady:
- Jak často se takováto povodeň může opakovat?
- Jakou hodnotu měřené veličiny nejpravděpodobněji získáme opakovaným měřením?
- Je vysoký počet dvojčat narozených v určitém okrese „normální"?
- Je rozdíl mezi dvěma jevy významný?

Náhodný jev, náhodná proměnná

Přírodní či společenské jevy mohou mít povahu jevů
- **deterministických**
- **náhodných**

Náhodný jev - za určitého souboru podmínek může nastat jeden z množiny výsledků, který závisí nejen na vstupních podmínkách, ale obsahuje i **prvek náhody** (tahání karet, měření teploty vzduchu, …).

Náhodná veličina – proměnná, u které nelze na základě určité zákonitosti předem stanovit její konkrétní hodnotu.

Náhodná proměnná může být:
- **Spojitá** – může nabývat jakékoliv hodnoty z určitého intervalu (teplota vzduchu)
- **Diskrétní** – může nabývat pouze konkrétních hodnot (házení kostkou, pohlaví narozeného dítěte)

Pravděpodobnost, náhodná proměnná

Pravděpodobnost

- Vyjadřuje **míru nejistoty**, s jakou určitý náhodný jev může nastat, s jakou může náhodná veličina nabývat určité hodnoty
- **Tuto míru nejistoty (pravděpodobnost) můžeme kvantifikovat.**

Jak často nastane určitý jev, pokud experiment (měření, výběr) provedu mnohokrát?

- Řada jevů a procesů studovaných v geografických disciplínách má charakter **náhodné proměnné.**
- Řada geografických jevů má **pravděpodobnostní charakter** (mohu nastat s určitou pravděpodobností) – např. výsledky prognóz (demografie, meteorologie apod.)
- Použití teorie pravděpodobnosti (např. pro testování, odhady) vyžaduje data získaná náhodným výběrem (náhodné proměnné).

Pravděpodobnost

- Pravděpodobnost jako vyjádření míry nejistoty o výskytu náhodného jevu, o výsledku náhodného jevu.
- Pravděpodobnost, že nastane určitý náhodný jev se pohybuje v intervalu <0,1> resp. <0,100> %.

- **Jev možný** – množina všech možných výsledků - náhodný jev
- **Jev jistý** – padne něco mezi 1 až 6
- **Jev nemožný** – padne 7
- **Jev elementární** – padne 6
- **Jev složený** – více možných výsledků (padne sudé číslo)

Pravděpodobnost

P(A) - Určení pravděpodobnosti *P*, s jakou náhodný jev *A* nastane, můžeme povést dvěma způsoby:

1. Určení pravděpodobnosti „a priori":

Podíl počtu požadovaných výsledků a počtu všech možných výsledků:

Př.: S jako pravděpodobností padne při házení kostkou šestka:

$$P(A) = \frac{m}{n}$$

A – jev který sledujeme
P(A) – pravděpodobnost jevu A
m - Počet požadovaných výsledků
n – počet všech možných výsledků

V našem případě: P(A=6): m=1 (šestku můžeme hodit jen jedním způsobem), n=6 (může padnout 6 různých čísel, tedy:

$$P(A=6) = \frac{m}{n} = \frac{1}{6} = 0{,}1667$$

Pravděpodobnost, že při hodu kostkou padne šestka je 16,67%

Pravděpodobnost

2. Určení pravděpodobnosti „a posteriori“:
Pomocí relativní četnosti výskytu studovaného jevu:

$$P(A) = \frac{n_i}{n}$$

ni – počet požadovaných výsledků, které nastaly při realizaci jevu (absolutní četnost)
n – celkový počet pokusů (rozsah souboru)

Příklad: Z deseti hodů kostkou (n=10) jsme získali následující výsledky: 2,4,6,1,6,3,5,6,2,1. Spočteme frekvenci výskytu jednotlivých výsledků a následně relativní četnost výsledku, při kterém padla šestka, tedy počet případů příznivých jevu A k počtu případů možných.

$$P(A=6) = \frac{n_i}{n} = \frac{3}{10} = 0{,}3$$

Empiricky zjištěná pravděpodobnost, že padne 6 je 30%.

Obě pravděpodobnosti se liší. Čím více **realizací náhodného pokusu provedeme**, tím si budou výsledky bližší, pro $n \to \infty$ budou shodné.

Některá pravidla pro počítání s pravděpodobnostmi

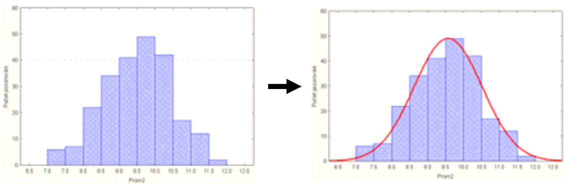
- Pravděpodobnost jistého jevu je jedna.
- Pravděpodobnost nemožného jevu je nula.
- Pravděpodobnost sjednocení dvou vzájemně neslučitelných (disjunktivních) jevů A a B je rovna součtu jejich pravděpodobností:
  platí-li $A \cap B = 0$, potom $P(A \cup B) = P(A) + P(B)$.
- Pro pravděpodobnost sjednocení jevů A a B, které se vzájemně nevylučují, platí, že se rovná součtu pravděpodobností obou jevů zmenšenému o pravděpodobnost jejich průniku.
  platí-li $P(A \cap B) \neq 0$, potom $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Pravděpodobnost současného objevení dvou nezávislých jevů (jejich průniku) je rovna součinu pravděpodobností těchto jevů:
  $P(A \cap B) = P(A) \cdot P(B)$
- Pro pravděpodobnosti dvou opačných jevů platí, že jejich součet je 1:
  $P(\overline{A}) = 1 - P(A)$

Rozdělení náhodné proměnné

Každému výsledku náhodného jevu či procesu přísluší určitá pravděpodobnost.

K čemu je to tedy dobré vědět?

Můžeme určit s jakou pravděpodobností náhodný jev nabývá určité výsledné hodnoty či hodnoty z určitého intervalu.



Konstruujeme model umožňující zobecnění našich poznatků o chování hromadných náhodných jevů – **teoretické rozdělení pravděpodobnosti**

Teoretická rozdělení – základní pojmy

Teoretická rozdělení ve statistice charakterizujeme:

1. **průběhem frekvenční a distribuční funkce**
2. **parametry rozdělení**

1. Pro diskrétní náhodnou proměnnou konstruujeme:
• rozdělení pravděpodobností diskrétní náhodné proměnné (**frekvenční funkci**)
• distribuční funkci diskrétní náhodné proměnné

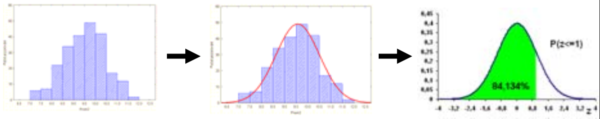
Pro spojitou náhodnou proměnnou konstruujeme:
• hustotu pravděpodobností spojité náhodné proměnné (**frekvenční funkci**)
• distribuční funkci spojité náhodné proměnné

2. Parametry rozdělení – čísla (statistiky)

Neznámé hodnoty základních statistických charakteristik základního souboru, které můžeme jen **odhadnout** z charakteristik výběrových

Výběrový soubor		Základní soubor
$\bar{x}$	$\rightarrow$	μ
s (s^2)		σ (σ^2)

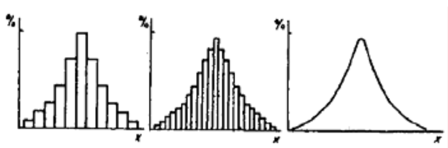
Teoretická rozdělení pravděpodobnosti



Rozdělení pravděpodobnosti *f(X)* náhodné proměnné X je funkce *f(X),* která každé hodnotě X přiřazuje určitou pravděpodobnost *p(x),* se kterou tato nabývá konkrétní velikosti.

Teoretická rozdělení spojité náhodné veličiny

K tzv. **frekvenční funkci** *f(x)*. Můžeme dospět jednoduše z histogramu relativních četností



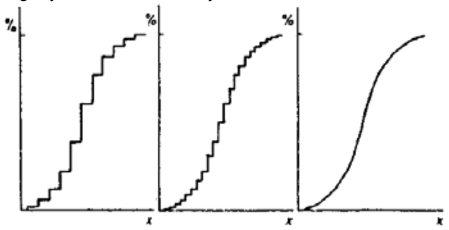
Frekvenční funkce f(x) představuje teoretické rozdělení četností základního souboru o parametrech μ, σ.

Cíl – nahradit výběrové soubory základními a pro ně odvozovat potřebné charakteristiky

(Příklad - hodnocení stupně normality výskytu určitých hodnot – povodně)

Teoretická rozdělení spojité náhodné veličiny

Analogicky lze ze součtové čáry definovat tzv. **distribuční funkci** *F(x)*.



Distribuční funkce udává pravděpodobnost, se kterou náhodná proměnná nabývá hodnoty menší nebo rovné určité konkrétní velikosti x.

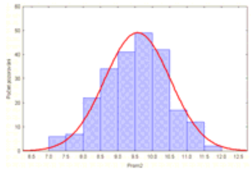
$$F(x) = P(X < x) = \sum_{x_i < X} P(X = x_i)$$

Normální rozdělení

- Nejčastěji používané rozdělení spojité náhodné veličiny.
- Opakované měření stejné veličiny za stejných podmínek.
- Naměřené veličiny více méně kolísají kolem skutečné hodnoty
- Má dva parametry μ, σ.

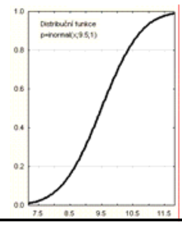
Frekvenční funkce:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

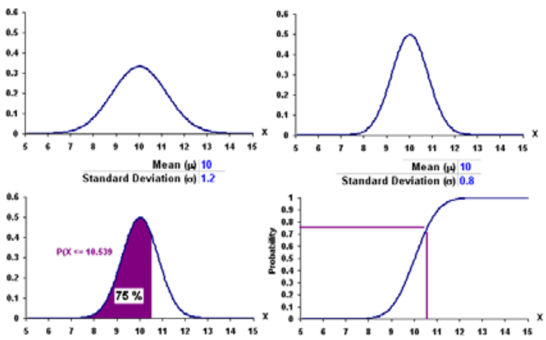


Distribuční funkce:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$



Normální rozdělení



Normované normální rozdělení

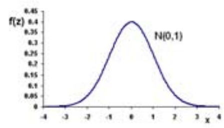
Protože se *μ* i *σ* výběr od výběru liší, má i frekvenční funkce (normální křivka) různý tvar. Proto se zavádí tzv. **normovaná náhodná proměnná**

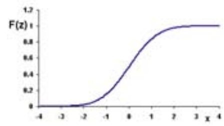
$$z = \frac{x - \mu}{\sigma}$$

Je to rozdělení hodnot, které mají střední hodnotu nula a jednotkový rozptyl ➡ N [0, 1]

Normované normální rozdělení již potom nezáleží na parametrech *μ* i *σ* a jeho frekvenční a distribuční funkce mají následující tvar:

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}} \qquad F(z) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{z} e^{-\frac{z^2}{2}} dz$$





Normované normální rozdělení

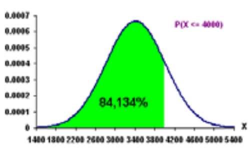
Pro hodnoty distribuční funkce *F(x)* normálního rozdělení N[μ,σ] a hodnoty distribuční funkce *F(z)* normovaného normálního rozdělení N[0,1] platí:

Jestliže $z_0 = \frac{x_0 - \mu}{\sigma}$ potom *F(x0) = F(z0)*

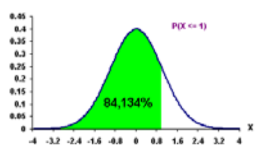
Obsahy ploch pod křivkami hustoty navzájem si odpovídajících si hodnot *F(x0) = F(z0)* jsou stejné (a tedy stejné jsou i pravděpodobnosti výskytu těchto hodnot). Například:

Protože $z_0 = \frac{x_0 - \mu}{\sigma} = \frac{4000 - 3400}{600} = 1$ potom $P(x \le 4000) = P(z \le 1) = 0{,}84134$

N[3400,600], x_0 = 4000

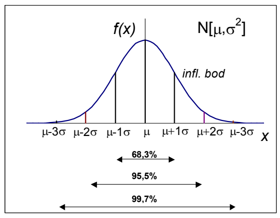


N[0,1], x_0 = 1



Hlavní vlastnosti normální křivky:

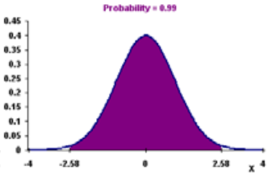
- zvonovitý tvar, asymptoticky se blíží k ose x, může nabývat hodnot ()
- souměrná podle osy procházející jejím vrcholem. x-ová souřadnice vrcholu je aritmetickým průměrem
- aritmetický průměr, medián a modus se rovnají
- s osou x omezuje normální křivka plochu o velikosti 100 % (1)



Pomocí násobků směrodatné odchylky lze stanovit pravděpodobnosti, s nimiž leží hodnoty v určitém intervalu:

Vlastnosti normální křivky:

a naopak
- 95% pravděpodobnosti odpovídá interval $\mu \pm 1{,}96\sigma$
- 99% pravděpodobnosti odpovídá interval $\mu \pm 2{,}58\sigma$



Stanovení mezí extremity

Slovní označení extremity	Symbol	Meze	Pravděpodobnost výskytu jevu (%)	Meze	Pravděpodobnost výskytu jevu (%)
extrémně podnormální	EP	$< \bar{x}-3s$	0,135	$< \bar{x}-3c$	2,15
silně podnormální	SP	$\bar{x}-3s$ až $\bar{x}-2s$	2,190	$\bar{x}-3c$ až $\bar{x}-2c$	8,87
podnormální	P	$\bar{x}-2s$ až $\bar{x}-s$	13,590	$\bar{x}-2c$ až $\bar{x}-c$	13,98
normální	O	$\bar{x}-s$ až $\bar{x}+s$	68,270	$\bar{x}-c$ až $\bar{x}+c$	50,00
nadnormální	N	$\bar{x}+s$ až $\bar{x}+2s$	13,590	$\bar{x}+c$ až $\bar{x}+2c$	13,98
silně nadnormální	SN	$\bar{x}+2s$ až $\bar{x}+3s$	2,190	$\bar{x}+2c$ až $\bar{x}+3c$	8,87
extrémně nadnormální	EN	$> \bar{x}+3s$	0,135	$> \bar{x}+3c$	2,15

c – tzv. pravděpodobná chyba: *c = 0,6745s*

Příklad použití

- Pro danou hodnotu jevu hledáme pravděpodobnost jejího výskytu
- Pro zadanou pravděpodobnost hledáme hodnotu studovaného jevu.

Dvě možnosti výpočtu
- převod na normované normální rozdělení a využití statistických tabulek
- pomocí sw modelujícího hodnoty *f(x)* a *F(x)* příslušného rozdělení

Příklad: Plochu obhospodařované zemědělské půdy u sledovaného souboru farmářů modelujeme normálním rozdělením. Zjistili jsme, že parametry rozdělení N [3400 m^2, 600 m^2]. Vypočtěte pravděpodobnost, že náhodně vybraný zemědělec bude mít:

a) méně než 4000 m^2 půdy
b) více než 4200 m^2 půdy
c) méně než 3000 m^2 půdy
d) mezi 2800 a 4000 m^2 půdy

Příklad řešení při použití tabulek distribuční funkce normovaného normálního rozdělení

Tables of the Normal Distribution

http://www.math.unb.ca/~knight/utility/NormTble.htm

Probability Content from -oo to Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

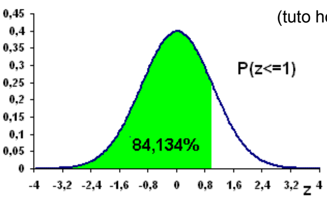
Ad a) Farmář má méně než 4000 m^2 půdy

Transformace hodnoty *x* na normovanou veličinu *z* (z-skóre):

$$z = \frac{x-\mu}{\sigma} = \frac{4000-3400}{600} = 1$$

Můžeme psát: $P(x \le 4000) = P(z \le 1) = 0{,}84134$

(tuto hodnotu nalezneme v tabulkách)



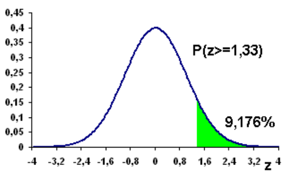
Ad b) Farmář má více než 4200 m^2 půdy

Transformace hodnoty *x* na normovanou veličinu *z* (z-skóre):

$$z = \frac{x-\mu}{\sigma} = \frac{4200-3400}{600} = 1{,}33$$

Pravděpodobnost, že normovaná proměnná překročí hodnotu 1,33 v tabulkách není. Určujeme obsah plochy pod křivkou hustoty rozdělení za hodnotou 1,33:

$$P(x \ge 4200) = P(z \ge 1{,}33) = 1 - P(z \le 1{,}33) = 1 - 0{,}90824 = 0{,}09176$$



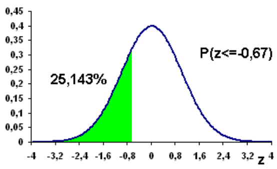
Ad c) Farmář má méně než 3000 m² půdy

Transformace hodnoty x na normovanou veličinu z (z-skóre):

$$z = \frac{x-\mu}{\sigma} = \frac{3000-3400}{600} = -0{,}67$$

Pravděpodobnost $P(z \le -0{,}67)$ určíme na základě symetrie normovaného normálního rozdělení, platí tedy:

$$P(z \le -0{,}67) = P(z \ge 0{,}67) = 1 - P(z \le 0{,}67) = 1 - 0{,}74857 = 0{,}25143$$



Ad d) Farmář má mezi 2800 a 4000 m² půdy

Transformace hodnoty x na normovanou veličinu z (z-skóre):

$$z_1 = \frac{x_1-\mu}{\sigma} = \frac{2800-3400}{600} = -1 \qquad z_2 = \frac{x_2-\mu}{\sigma} = \frac{4000-3400}{600} = 1$$

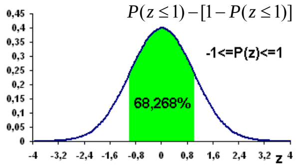
Plocha mezi hodnotami 2800 a 4000 m² u rozdělení N [3400, 600] je stejná jako plocha mezi hodnotami -1 a 1 u rozdělení N [0, 1].

$$P(2800 \le x \le 4000) = P(-1 \le z \le 1)$$

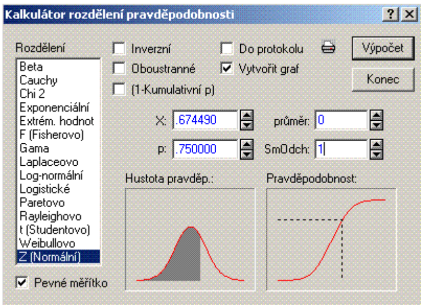
Od plochy před hodnotou 1 odečteme plochy před hodnotou -1, tedy:

$$P(-1 \le z \le 1) = P(z \le 1) - P(z \le -1) =$$

$$P(z \le 1) - [1 - P(z \le 1)] = 0{,}84134 - [1 - 0{,}84134] = 0{,}68268$$

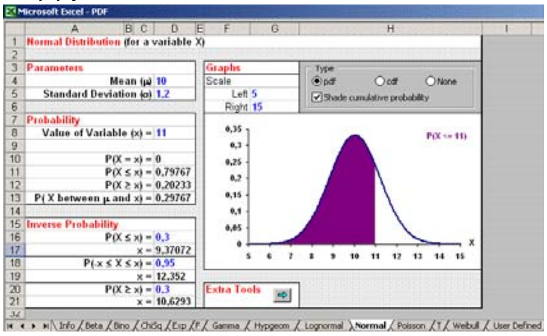


Řešení při použití pomocí sw modelujícího hodnoty f(x) a F(x) příslušného rozdělení



STATISTICA – Pravděpodobnostní kalkulátor

Řešení při použití sw modelujícího hodnoty f(x) a F(x) příslušného rozdělení



EXCEL – soubor PDF – viz. studijní materiály v IS

Pearsonova křivka III. typu

- Na řadu souborů geografických údajů nelze aplikovat normální rozdělení.
- Je to v případech, že studovaná veličina nemá teoreticky zdůvodněnou možnost nabývat nekonečných hodnot či je omezena z obou stran konečnými čísly.
- V těchto případech lze často využít některé z 12 křivek Pearsonovy systému.
- Především v meteorologii a klimatologii se ke konstrukci tzv. čar překročení využívá Pearsonovy křivky III. typu.

Průběh křivky je určen třemi parametry:

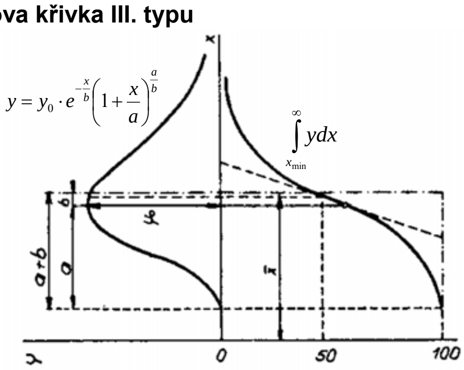
- aritmetickým průměrem
- variačním koeficientem
- koeficientem asymetrie

Pearsonova křivka III. typu

Tvar rovnice:

$$y = y_0 \cdot e^{-\frac{x}{b}} \left(1 + \frac{x}{a}\right)^{\frac{a}{b}}$$

$$\int_{x_{\min}}^{\infty} y\,dx$$



Hodnota y_0 značí největší pořadnici křivky a odpovídá modu rozdělení. b – vzdálenost pořadnice procházející aritmetickým průměrem od y_0, a – vzdálenost y_0 od počátku křivky.

Čára překročení

Čára překročení je součtová čára četností a lze z ní stanovit pravděpodobnost, se kterou bude znak určité hodnoty dosažený a překročený (či nebude dosažený).
Ke konstrukci čáry překročení musíme znát tři parametry:

Spočteme aritmetický průměr

Označíme-li $\frac{x_i}{\bar{x}} = k_i$ potom variační koeficient bude: $v = \sqrt{\frac{\sum (k_i - 1)^2}{n-1}}$

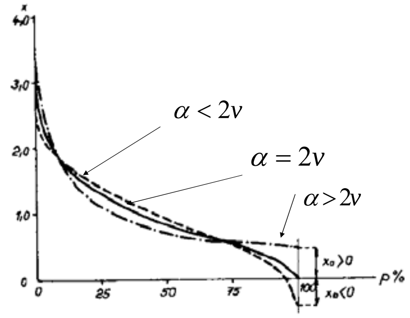
Koeficient asymetrie je roven $\alpha = \frac{2b}{v}$ a nebo $\alpha = \frac{\sum (k_i - 1)^3}{(n-1) \cdot v^3}$

a nebo $\alpha = 2v$

Uvedená podmínka je často nutná pro fyzikálně zdůvodnitelné výsledky

Čára překročení

Tvar čáry překročení pro různá α



Konstrukce čáry překročení I

1) Seřadíme hodnoty R_i v klesajícím pořadí
2) Vypočteme aritmetický průměr $\overline{R}$
3) Stanovíme hodnoty $k_i = \frac{R_i}{\overline{R}}$
4) Vypočteme výrazy $(k_i - 1), (k_i - 1)^2, (k_i - 1)^3$ a jejich sumy
5) Vypočteme hodnoty variačního koeficientu *v* a koeficientu asymetrie α

	A	B	C	D	E	F	G	H	I	J	K
1	Pořadí	Rok	Ri	ki	ki-1	(ki-1)²	(ki-1)³	m	P%	Pc	T
2	1	1981	90.0	2.662	1.662	2.762	4.590	1	1.78	0.02	56
3	2	1977	66.5	1.967	0.967	0.935	0.904	2	4.31	0.04	23
4	3	1979	62.5	1.849	0.849	0.720	0.611	3	6.85	0.07	15
5	4	1996	51.5	1.523	0.523	0.274	0.143	4	9.39	0.09	11
6	...	...	...	...	...	...	...	...	...	...	...
7	...	...	...	...	...	...	...	...	...	...	...
8	36	1998	19.6	0.580	-0.420	0.177	-0.074	36	90.61	0.91	1
9	37	1976	19.2	0.568	-0.432	0.187	-0.081	37	93.15	0.93	1
10	38	1990	18.4	0.544	-0.456	0.208	-0.095	38	95.69	0.96	1
11	39	1984	15.9	0.470	-0.530	0.281	-0.149	39	98.22	0.98	1

pravděpodobnost překročení

$P_c = \frac{m - 0{,}3}{n + 0{,}4}$

$T = \frac{1}{P_c}$

doba opakování

$v = \sqrt{\frac{\sum (k_i - 1)^2}{n-1}} = 0{,}447$ $\alpha \approx 2v = 0{,}90$

Konstrukce čáry překročení II

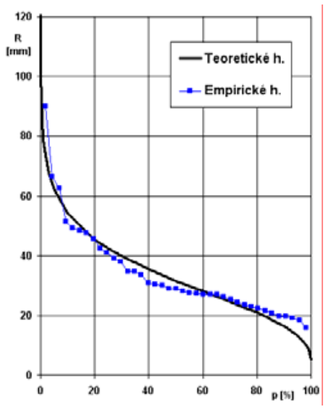
6) Pro jednotlivé hodnoty pravděpodobnosti překročení *p* a pro vypočtenou hodnotu koeficientu asymetrie α vypočteme pomocí tabelovaných hodnot *E* pro případ *v = 1* pořadnice čáry překročení – teoretické hodnoty *R*.

7) Každé hodnotě R odpovídá určité p a určitá doba opakování (T)

Pravděpodobnosti překročení pro koeficient asymetrie 0,9

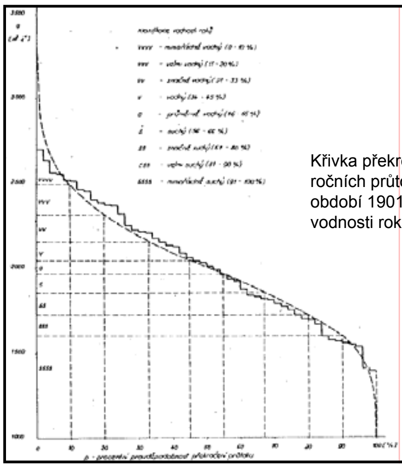
P	E	(E*v) + 1	R	T
0.01	5.73	3.56	120.52	100.00
0.1	4.38	2.96	100.09	10.00
1	2.96	2.32	78.60	1.00
3	2.22	1.99	67.41	0.33
5	1.86	1.83	61.96	0.20
10	1.34	1.60	54.09	0.10
20	0.77	1.34	45.46	0.05
25	0.57	1.26	42.44	0.04
30	0.40	1.18	39.86	0.03
40	0.11	1.05	35.47	0.03
50	-0.15	0.93	31.54	0.02
60	-0.38	0.83	28.06	0.02
70	-0.61	0.73	24.58	0.01
75	-0.73	0.67	22.76	0.01
80	-0.85	0.62	20.95	0.01
90	-1.15	0.49	16.41	0.01
95	-1.35	0.40	13.38	0.01
97	-1.47	0.34	11.56	0.01
99	-1.66	0.26	8.69	0.01
99.9	-1.90	0.15	5.06	0.01

Konstrukce čáry překročení III



Hodnocení extremity jevu na základě procenta pravděpodobnosti překročení určeného z čáry překročení:

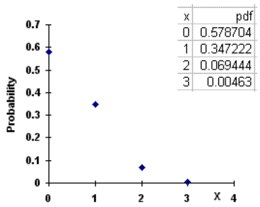
Pravděpodobnost	Jev je	Symbol
0 - 10	extrémně nadnormální	EN
11 - 20	velmi silně nadnormální	VN
21 - 33	silně nadnormální	SN
34 - 45	nadnormální	N
46 - 55	normální	O
56 - 67	podnormální	P
68 - 80	silně podnormální	SP
81 - 90	velmi silně podnormální	VP
91 - 100	extrémně podnormální	EP



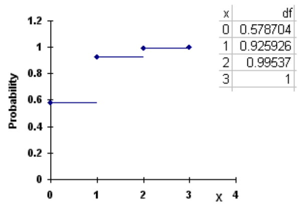
Křivka překročení průměrných ročních průtoků Dunaje v Bratislavě v období 1901-1950 a hodnocení vodnosti roků

Rozdělení diskrétní náhodné proměnné

Příklad: Třikrát hodíme kostkou. Jaká je pravděpodobnost, že šestka nepadne, že padne jednou, dvakrát, třikrát?



Rozdělení pravděpodobnosti



Distribuční funkce

Binomické rozdělení

Rozdělení diskrétní náhodné proměnné. Udává rozdělení výsledků jednoho a téhož pokusu za stejných podmínek, kdy výsledkem pokusu mohou být pouze dvě alternativy *A* nebo *B*.

Pravděpodobnost, že nastane alternativa *A* značíme *p*, pravděpodobnost, že nastane alternativa *B* značíme *q*.

Přitom platí: $p + q = 1$

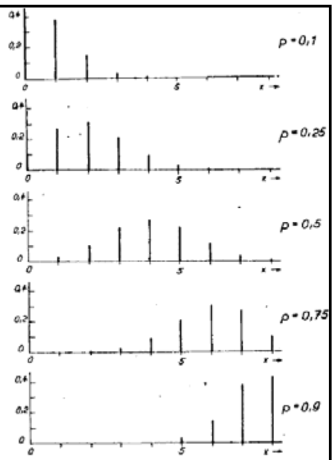
Pokus provedeme *n*-krát a hledáme pravděpodobnost, že alternativa *A* nastane právě *x*-krát. Výpočet této pravděpodobnosti určuje výraz:

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

Uvedený vztah vyjadřuje rozdělení pravděpodobností binomického rozdělení (analogie frekvenční funkce u spojitých veličin).

Binomické rozdělení

U binomického rozdělení nabývá náhodná veličina diskrétních hodnot od 0 do n.



Rozdělení pravděpodobnosti binomického rozdělení pro *n=8* a různé hodnoty pravděpodobnosti *p*

Je-li *p = q = 0,5*, potom je binomické rozdělení souměrné

Základní momenty binomického rozdělení

$$\mu = n \cdot p \qquad \sigma = \sqrt{n \cdot p \cdot q} \qquad \alpha = \frac{1-2p}{\sqrt{n \cdot p \cdot q}} \qquad \varepsilon = \frac{1-6p \cdot q}{n \cdot p \cdot q}$$

Příklady použití binomického rozdělení

- rozdělení počtu dní s určitým meteorologickým jevem za měsíc
- pravděpodobnost narození dvou chlapců v rodinách se třemi dětmi
- pravděpodobnost pozdního příchodu na jednu ze 12 přednášek ze statistiky

Binomické rozdělení

Příklad: Pravděpodobnost, že se v určitém roce vyskytne na studovaném toku povodeň je 0,25. Jaká je pravděpodobnost, že se během příštích čtyř let vyskytnou 3 povodně?

- každý rok je nezávislý „pokus“
- každý rok se povodeň může vyskytnout či nemusí
- pravděpodobnost výskytu povodně p=0,25
- pro 4 roky (n=4) hledáme pravděpodobnost výskytu tří povodní (x=3)

$$f(x) = \binom{n}{x} \cdot p^x \cdot q^{n-x} = \binom{4}{3} \cdot 0{,}25^3 \cdot 0{,}75^{4-3} = 0{,}0469$$

Pravděpodobnost výskytu tří povodní ve 4 rocích je necelých 5 procent.

Binomické rozdělení

Binomial Distribution (for a variable K)	
Parameters	
Number of Trials (n)	4
Probability of a Success (π)	0.25
Probability	
Number of "Successful" Outcomes (k) =	3
P(K = k) =	0.04688
P(K ≤ k) =	0.99609
P(K ≥ k) =	0.05078
P(K < k) =	0.94922
P(K > k) =	0.00391

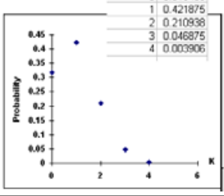
Příklad – pokračování

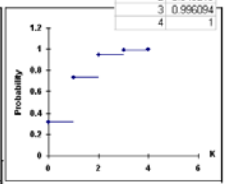
Pravděpodobnostní funkce

Distribuční funkce

k	pdf
0	0.316406
1	0.421875
2	0.210938
3	0.046875
4	0.003906

k	cdf
0	0.316406
1	0.738281
2	0.949219
3	0.996094
4	1





Poissonovo rozdělení

Popisuje pravděpodobnost výskytu **vzácných jevů**. Je vhodné v případech, kdy *p* v binomickém rozdělení je příliš malé nebo naopak příliš blízké 1.

Náhodná veličina s Poissonovým rozdělením může nabývat hodnot *x = 0,1,2,3,...* (kolikrát jev nastal v určitém časovém úseku) a to s rozdělením pravděpodobnosti:

$$f(x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$$

Pro aritmetický průměr a rozptyl platí: $\mu = \sigma^2 = \lambda$

kde lambda (λ) je očekávaná hodnota a jediný parametr Poissonova rozdělení: $\lambda = n \cdot p$

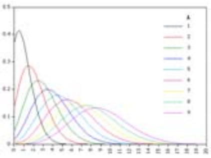
Označuje se jako rozdělení vzácných případů (bouřky v zimě, výskyt krupobití v roce, …). Jeho použití se doporučuje, pokud n > 30 (resp. 50) a $p \leq 0{,}1$ nebo $p \geq 0{,}9$.

Poissonovo rozdělení

Použití tohoto rozdělení lze charakterizovat následujícími vlastnostmi:

1. Pravděpodobnost výskytu jedné události v daném intervalu (čase nebo prostoru) je úměrná délce tohoto intervalu.
2. Události se vyskytují nezávisle jak ve stejném intervalu, tak mezi po sobě jdoucími intervaly.
3. Událost může nastat v kterémkoliv okamžiku
4. Výskyt dvou či většího počtu událostí během krátkého časového okamžiku (ale i v malém prostoru) je prakticky nemožný

Použití:
- počet dětí ztracených v obchodním domě v určité časovém úseku
- počet telefonních hovorů v určitém časovém úseku
- počet borovic na jednotku plochy smíšeného lesa



S rostoucí hodnotou λ se tvar tohoto rozdělení blíží normálnímu

Poissonovo rozdělení - příklad

Průměrný počet těžkých dopravních nehod na určité křižovatce za měsíc je 5. Rozdělení četností nehod modelujeme Poissonovým rozdělením. Jaká je pravděpodobnost, že počet nehod bude více než 4.

- Rozdělení četností nehod obecně pro x = 0,1,2,3,4 má tvar:

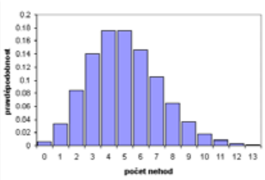
$$f(x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!} = \frac{5^x \cdot e^{-5}}{x!}$$

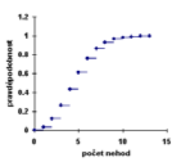
- Vypočteme pravděpodobnosti *f(x)* pro jednotlivá *x*: *f(0),…. f(4)*
- Součet těchto pravděpodobností je 0,44
- Hledaná pravděpodobnost – tedy že počet nehod přesáhne 4:

1 - 0,44 = 0,56

Poissonovo rozdělení – příklad (pokračování)

x	pdf	cdf
0	0.007	0.007
1	0.034	0.040
2	0.084	0.125
3	0.140	0.265
4	0.175	0.440
5	0.175	0.616
6	0.146	0.762
7	0.104	0.867
8	0.065	0.932
9	0.036	0.968
10	0.018	0.986
11	0.008	0.995
12	0.003	0.998
13	0.001	0.999





Rozdělení CHÍ - kvadrát χ^2

Ze základního souboru s normovaným normálním rozdělením provedeme náhodný výběr *n* prvků, které označíme $x_1, x_2, \ldots, x_n$.

Součet čtverců těchto hodnot se označuje jako χ^2 („chí – kvadrát"):

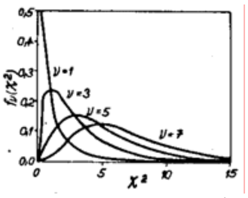
$$\sum_{i=1}^{n} x_i^2 = \chi^2$$

Hodnota χ^2 může nabývat v různých výběrech různých hodnot v intervalu $(0;\infty)$ a má své vlastní rozdělení - rozdělení s vlastní frekvenční ($f_\nu(\chi^2)$) a distribuční funkcí ($F_\nu(\chi^2)$)

Symbol ν značí počet stupňů volnosti a je jediným parametrem rozdělení. Je roven rozsahu náhodného výběru.

Každé hodnotě $\nu = n$ přísluší jiná křivka. S rostoucím ν se rozdělení blíží rozdělení normálnímu.

Rozdělení CHÍ - kvadrát χ^2



Frekvenční funkce chí-kvadrát rozdělení pro různý počet stupňů volnosti

Použití:
- v teorii odhadu a testování hypotéz
- při ověřování předpokladu zda empirické rozdělení četností má určité teoretické pravděpodobností rozdělení
- testování rozptylu dvou výběrových souborů při neznámé střední hodnotě
- při ověřování nezávislosti kvalitativních znaků
- pro testy nezávislosti v kontingenčních tabulkách.

Rozdělení t (Studentovo)

Využívá se především pro hodnocení odchylek hodnot aritmetického průměru základního souboru μ a aritmetického průměru výběrového souboru $\bar{x}$.

Pro hodnocení odchylek $(x-\mu)$ se definuje náhodná veličina

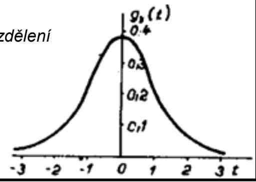
$$t = \frac{\bar{x}-\mu}{s}\sqrt{n-1}$$

Této přísluší tzv. *t-rozdělení* (Studentovo). Spojitá náhodná veličina *t* může nabývat hodnot $(-\infty;\infty)$. Frekvenční funkce $q_\nu(t)$ je souměrná podle osy procházející vrcholem a má jeden parametr $\nu = n-1$

S rostoucím počtem stupňů volnosti se *t-rozdělení* blíží rozdělení normálnímu.

Teoreticky se shodují při $\nu = \infty$.

V praxi však postačuje $\nu > 30$.

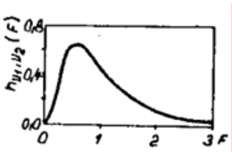


Rozdělení F (Fisherovo-Snedecorovo)

Uvažujeme dvě nezávislé náhodné veličiny, které mají χ^2 rozdělení s ν_1 a ν_2 stupni volnosti. Veličina *F*, určená jako jejich poměr

$$F = \frac{\chi_1^2}{\nu_1} : \frac{\chi_2^2}{\nu_2}$$

má tzv. *F – rozdělení.*

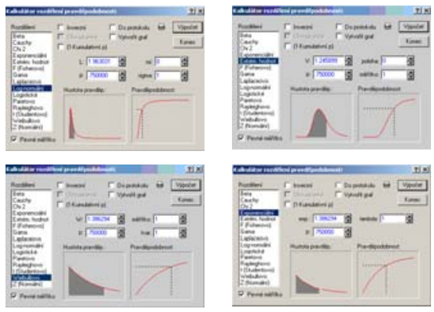


Jak je patrné z obrázku, náhodná veličina *F* nabývá pouze kladných hodnot.

Frekvenční funkce $h_{\nu_1,\nu_2}(F)$ je nesymetrická s dvěma parametry ν_1 a ν_2

Používá se u testů v regresní analýze, při analýze rozptylu a při testu shody rozptylů dvou výběrů z normálního rozdělení.

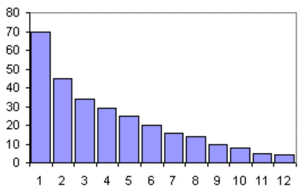
Další v geografii využívaná teoretická rozdělení



- Gama rozdělení
- Gumbelovo rozdělení
- GEV (General Extreme Value)

Asymetrická rozdělení

- vzdálenosti na které se lidé stěhují, cestují na dovolenou, do práce …
- osobní příjmy
- vzdálenost dojížďky



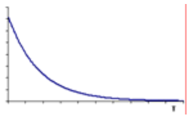
Většina těchto rozdělení má kladnou asymetrii

Nalezněte příklady pro rozdělení se zápornou asymetrií (?!)

Jak lze charakterizovat rozdělení takových veličin jako:
- počet členů domácnosti
- četnost zaměstnanců podle průměrné měsíční mzdy

Exponenciální rozdělení

Frekvenční funkce *f(x)* má následující tvar a její sklon klesá s rostoucí hodnotou x



$f(x) = \lambda e^{-\lambda x}$ pro $x \geq 0$

Rozdělení má jeden parametr (*λ*), jeho velké hodnoty indikují velký sklon frekvenční funkce a naopak.

Hodnoty pravděpodobnosti lze určovat přímo z distribuční funkce:

$$F(x) = p(X < x) = 1 - e^{-\lambda x}$$

Očekávaná hodnota průměru je 1/*λ*, a očekávaná hodnota rozptylu 1/*λ*^{2}.
Z očekávané hodnoty průměru potom lze určit hodnotu parametru (*λ*).

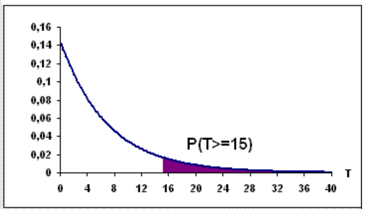
Hodláme-li výběrovým souborem, který má silně pozitivní asymetrii proložit exponenciální rozdělení, určíme průměrnou hodnotu výběrového souboru a parametr rozdělení *λ* bude jeho převrácenou hodnotou.

Exponenciální rozdělení - příklad

40 studentů dojíždí do školy z průměrné vzdálenosti 7 km. Histogram hodnot vzdálenosti vykazuje pozitivní asymetrii. Hodnota parametru exponenciálního rozdělení bude λ = 1/7 = 0,143. Jaká je pravděpodobnost, že student dojíždí ze vzdálenosti 15 km a delší?

Exponential Distribution (for a variable T)

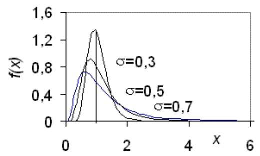
Parameters	
Failure Rate (λ)	0,143
Probability	
Value of Variable (t) =	15
P(T = t) =	0
P(0 ≤ T ≤ t) =	0,88293
P(T ≥ t) =	0,11707
Inverse Probability	
P(T ≤ t) =	0,95
t =	20,9492
P(T ≥ t) =	0,05
t =	20,9492



Lognormální rozdělení

Proměnná X má lognormální rozdělení pravděpodobností, když logaritmickou transformací ($Y=lnX$) získá právě rozdělení normální s parametry μ a σ^2.

Frekvenční funkce log-normálního rozdělení: $f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$



Příklady:

- rozdělení věku obyvatelstva v populaci
- koncentrace stopových prvků v horninách (stopová analýza),

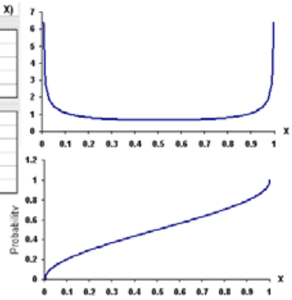
Beta rozdělení
U - rozdělení

α,β – parametry tvaru rozdělení

A,B – dolní a horní mez

Beta Distribution (for a variable X)

Parameters	
Alpha (α)	0.5
Beta (β)	0.5
A	0
B	1
Probability	
Value of Variable (x) =	0.5
P(X = x) =	0
P(X ≤ x) =	0.5
P(X ≥ x) =	0.5



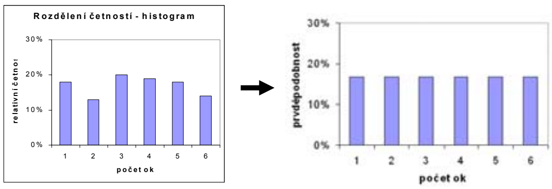
Limitní věty (dobrá zpráva)

Formulují obecná tvrzení o teoretických rozděleních v případě rostoucího rozsahu výběrového souboru (n)

Zákon velkých čísel

S rostoucím počtem pokusů se empirické rozdělení blíží teoretickému (relativní četnost se blíží pravděpodobnosti).

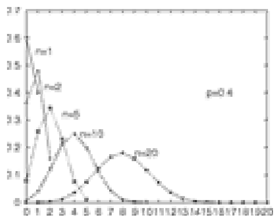
Střední hodnotu náhodné veličiny můžeme odhadovat průměrem z počtu dostatečně velkého pozorování



Centrální limitní věta

Pravděpodobnostní rozdělení náhodné veličiny, která vznikla jako součet velkého počtu vzájemně nezávislých náhodných veličin, je možné aproximovat rozdělením normálním, i když výchozí dílčí náhodné veličiny mají rozdělení různá.

Rozdělení takovéto náhodné veličiny se totiž blíží k normálnímu rozdělení jako limitnímu, pokud je počet dílčích veličin dostatečně velký.



Pro dostatečně velké n můžeme binomické rozdělení aproximovat normálním