

Benfordův zákon aneb o jedné překvapivé vlastnosti

OSMA, VŠB TU, Ostrava!!!

Mirko Rokyta
(...podesáté, **25. 3. 2025**)



matfyz

KMA MFF UK Praha

1. O co jde?

1. O co jde?

Problém: Jaké je rozložení prvních cifer u velkých souborů číselných dat?

1. O co jde?

Problém: Jaké je rozložení prvních cifer u velkých souborů číselných dat?

- **Intuitivní odpověď:** vždy rovnoměrné.

Problém: Jaké je rozložení prvních cifer u velkých souborů číselných dat?

- **Intuitivní odpověď**: vždy rovnoměrné.
- **Neintuitivní skutečnost**: existují soubory číselných dat, které často přicházejí tzv. "z praxe",

Problém: Jaké je rozložení prvních cifer u velkých souborů číselných dat?

- **Intuitivní odpověď:** vždy rovnoměrné.
- **Neintuitivní skutečnost:** existují soubory číselných dat, které často přicházejí tzv. "z praxe", u kterých je výskyt čísla 1 na první pozici mnohem pravděpodobnější než výskyt čísla 2,

Problém: Jaké je rozložení prvních cifer u velkých souborů číselných dat?

- **Intuitivní odpověď:** vždy rovnoměrné.
- **Neintuitivní skutečnost:** existují soubory číselných dat, které často přicházejí tzv. "z praxe", u kterých je výskyt čísla 1 na první pozici mnohem pravděpodobnější než výskyt čísla 2, a tak dál s klesající pravděpodobností - nejméně pravděpodobný je výskyt čísla 9 na první pozici.

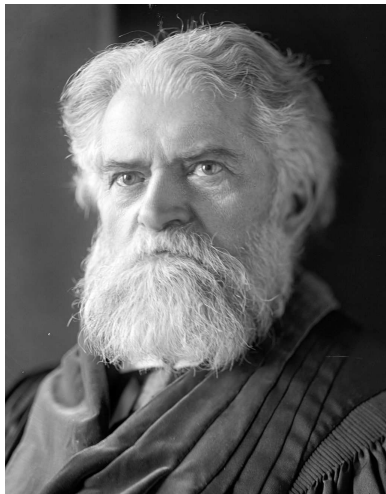
1. O co jde?

Problém: Jaké je rozložení prvních cifer u velkých souborů číselných dat?

- **Intuitivní odpověď:** vždy rovnoměrné.
- **Neintuitivní skutečnost:** existují soubory číselných dat, které často přicházejí tzv. "z praxe", u kterých je výskyt čísla 1 na první pozici mnohem pravděpodobnější než výskyt čísla 2, a tak dál s klesající pravděpodobností - nejméně pravděpodobný je výskyt čísla 9 na první pozici.

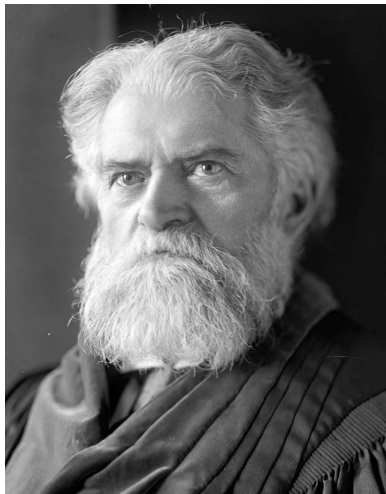
⇒ **Benfordův (či spíše Newcomb-Benfordův) zákon.**

2. Dva rodičové



Simon Newcomb
(1835 – 1909)

2. Dva rodičové

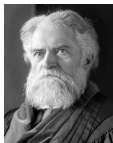


Simon Newcomb
(1835 – 1909)

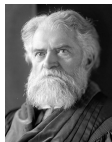


Frank Benford

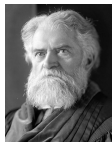
(1883 – 1948)



Simon Newcomb (fyzik-astronom, nikoli matematik):

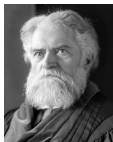


Simon Newcomb (fyzik-astronom, nikoli matematik): logaritmické tabulky jsou mnohem "ošahanější" u mantis začínajících číslem 1 než u čísel vyšších.



Simon Newcomb (fyzik-astronom, nikoli matematik): logaritmické tabulky jsou mnohem "ošahanější" u mantis začínajících číslem 1 než u čísel vyšších.

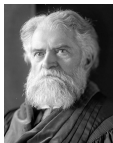
Newcomb, S.: Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics* 4, 39-40, **1881**.



Simon Newcomb (fyzik-astronom, nikoli matematik): logaritmické tabulky jsou mnohem "ošahanější" u mantis začínajících číslem 1 než u čísel vyšších.

Newcomb, S.: Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics* 4, 39-40, **1881**.

Matematické vysvětlení (odvození) nepodal,



Simon Newcomb (fyzik-astronom, nikoli matematik): logaritmické tabulky jsou mnohem "ošahanější" u mantis začínajících číslem 1 než u čísel vyšších.

Newcomb, S.: Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics* 4, 39-40, **1881**.

Matematické vysvětlení (odvození) nepodal, ale heuristicky pracoval s tím, že "*data, poskytovaná přírodou*, mají tu vlastnost, že jejich první cifry jsou rovnoměrně rozloženy na logaritmické škále."

Co tím tedy myslel?

Co tím tedy myslel?

Benfordův zákon

Cifra	Relativní četnost
1	30,10%
2	17,61%
3	12,49%
4	9,69%
5	7,92%
6	6,69%
7	5,80%
8	5,12%
9	4,58%

Newcomb vypožoroval, že jde o logaritmické rozložení (možná i pod vlivem tabulek logaritmů...):

Newcomb vypočítával, že jde o logaritmické rozložení (možná i pod vlivem tabulek logaritmů...):

Počáteční cifra	výpočet	v procentech
1	$\log 2 - \log 1 = 0,30103$	30,10%
2	$\log 3 - \log 2 = 0,17609$	17,61%
3	$\log 4 - \log 3 = 0,12494$	12,49%
4	$\log 5 - \log 4 = 0,09691$	9,69%
5	$\log 6 - \log 5 = 0,07918$	7,92%
6	$\log 7 - \log 6 = 0,06695$	6,69%
7	$\log 8 - \log 7 = 0,05799$	5,80%
8	$\log 9 - \log 8 = 0,05115$	5,12%
9	$\log 10 - \log 9 = 0,04576$	4,58%

Newcombův článek upadl asi na 50 let v zapomnění

...

Newcomb vypočítával, že jde o logaritmické rozložení (možná i pod vlivem tabulek logaritmů...):

Počáteční cifra	výpočet	v procentech
1	$\log 2 - \log 1 = 0,30103$	30,10%
2	$\log 3 - \log 2 = 0,17609$	17,61%
3	$\log 4 - \log 3 = 0,12494$	12,49%
4	$\log 5 - \log 4 = 0,09691$	9,69%
5	$\log 6 - \log 5 = 0,07918$	7,92%
6	$\log 7 - \log 6 = 0,06695$	6,69%
7	$\log 8 - \log 7 = 0,05799$	5,80%
8	$\log 9 - \log 8 = 0,05115$	5,12%
9	$\log 10 - \log 9 = 0,04576$	4,58%

Newcombův článek upadl asi na 50 let v zapomnění ... až do cca roku **1938**.



Frank Benford (také fyzik): studoval "benfordovské" rozložení na velké množině různých dat



Frank Benford (také fyzik): studoval "benfordovské" rozložení na velké množině různých dat (délky řek,



Frank Benford (také fyzik): studoval "benfordovské" rozložení na velké množině různých dat (délky řek, finanční transakce,



Frank Benford (také fyzik): studoval "benfordovské" rozložení na velké množině různých dat (délky řek, finanční transakce, počty obyvatel měst...).



Frank Benford (také fyzik): studoval "benfordovské" rozložení na velké množině různých dat (délky řek, finanční transakce, počty obyvatel měst...). Znovuoživení Newcombových myšlenek podal v článku



Frank Benford (také fyzik): studoval "benfordovské" rozložení na velké množině různých dat (délky řek, finanční transakce, počty obyvatel měst...). Znovuoživení Newcombových myšlenek podal v článku

Benford, F.: The Law of Anomalous Numbers. *Proceedings Of The American Philosophical Society*, vol. 78, no. 4, pp. 551–572, **1938**.



Frank Benford (také fyzik): studoval "benfordovské" rozložení na velké množině různých dat (délky řek, finanční transakce, počty obyvatel měst...). Znovuoživení Newcombových myšlenek podal v článku

Benford, F.: The Law of Anomalous Numbers. *Proceedings Of The American Philosophical Society*, vol. 78, no. 4, pp. 551–572, **1938**.

Benford neuměl vysvětlit důvod chování příslušných sad dat,



Frank Benford (také fyzik): studoval "benfordovské" rozložení na velké množině různých dat (délky řek, finanční transakce, počty obyvatel měst...). Znovuoživení Newcombových myšlenek podal v článku

Benford, F.: The Law of Anomalous Numbers. *Proceedings Of The American Philosophical Society*, vol. 78, no. 4, pp. 551–572, **1938**.

Benford neuměl vysvětlit důvod chování příslušných sad dat, dokonce se jednu dobu domníval, že to plyne z přirozené vlastnosti posloupnosti přirozených čísel.



Frank Benford (také fyzik): studoval "benfordovské" rozložení na velké množině různých dat (délky řek, finanční transakce, počty obyvatel měst...). Znovuoživení Newcombových myšlenek podal v článku

Benford, F.: The Law of Anomalous Numbers. *Proceedings Of The American Philosophical Society*, vol. 78, no. 4, pp. 551–572, **1938**.

Benford neuměl vysvětlit důvod chování příslušných sad dat, dokonce se jednu dobu domníval, že to plyne z přirozené vlastnosti posloupnosti přirozených čísel.

Tato chybná úvaha se občas objeví i dnes, proto se nejprve podívejme na ni.

3. Chybná úvaha pana Benforda

3. Chybná úvaha pana Benforda

Benford se snažil dokázat, že množina čísel

$$\{1, 10, 11, \dots, 19, 100, 101, 102, \dots, 199, \dots\}$$

3. Chybná úvaha pana Benforda

Benford se snažil dokázat, že množina čísel

$$\{1, 10, 11, \dots, 19, 100, 101, 102, \dots, 199, \dots\}$$

má v limitě "pravděpodobnost výskytu" $0,301 \approx \log_{10} 2$

...

3. Chybná úvaha pana Benforda

Benford se snažil dokázat, že množina čísel

$$\{1, 10, 11, \dots, 19, 100, 101, 102, \dots, 199, \dots\}$$

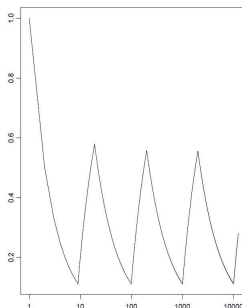
má v limitě "pravděpodobnost výskytu" $0,301 \approx \log_{10} 2$
... (aneb "v koši u šatnářky převažují takováto čísla").

3. Chybná úvaha pana Benforda

Benford se snažil dokázat, že množina čísel

$$\{1, 10, 11, \dots, 19, 100, 101, 102, \dots, 199, \dots\}$$

má v limitě "pravděpodobnost výskytu" $0,301 \approx \log_{10} 2$
... (aneb "v koši u šatnářky převažují takováto čísla").



Frekvence čísel, začínajících cifrou 1, mezi prvními n
přirozenými čísly.

4. Experimenty autora prezentace

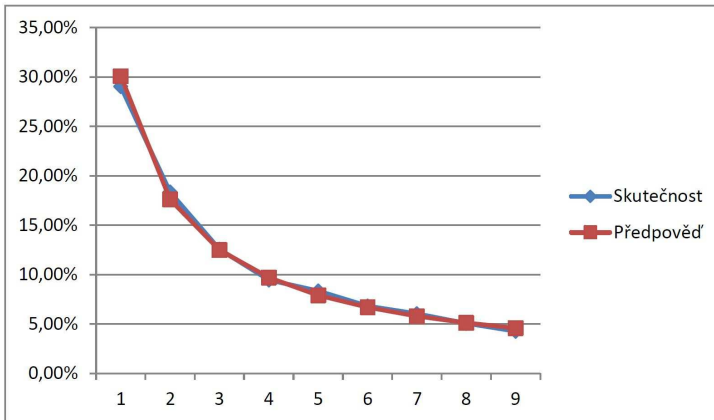
4. Experimenty autora prezentace

Počet obyvatel obcí ČR

6253

Cifra	Četnost výskytu první cifry	Relativní četnost výskytu první cifry	Co předpovídá Benford
1	1816	29,04%	30,10%
2	1148	18,36%	17,61%
3	786	12,57%	12,49%
4	592	9,47%	9,69%
5	522	8,35%	7,92%
6	427	6,83%	6,69%
7	378	6,05%	5,80%
8	317	5,07%	5,12%
9	267	4,27%	4,58%

Počet obyvatel obcí ČR - teorie a skutečnost



4. Experimenty autora prezentace

Benford a Rokytovy bankovní transakce

Cifra	Četnost výskytu první cifry	Relativní četnost výskytu. první cifry	Co předpovídá Benford
1	223	31,19%	30,10%
2	105	14,69%	17,61%
3	99	13,85%	12,49%
4	68	9,51%	9,69%
5	63	8,81%	7,92%
6	44	6,15%	6,69%
7	52	7,27%	5,80%
8	30	4,20%	5,12%
9	31	4,34%	4,58%

5. Tušení příčiny

5. Tušení příčiny

Jaké jsou (empiricky vypořizované) důvody, které způsobí, že daná množina číselných dat **nesplňuje** benfordovské rozdělení?

5. Tušení příčiny

Jaké jsou (empiricky vypořizované) důvody, které způsobí, že daná množina číselných dat **nesplňuje** benfordovské rozdělení?

- příliš málo dat

5. Tušení příčiny

Jaké jsou (empiricky vypočítávané) důvody, které způsobí, že daná množina číselných dat **nesplňuje** benfordovské rozdělení?

- příliš málo dat (obecný statistický požadavek)

5. Tušení příčiny

Jaké jsou (empiricky vypořizované) důvody, které způsobí, že daná množina číselných dat **nesplňuje** benfordovské rozdělení?

- příliš málo dat (obecný statistický požadavek)
- data, která jsou nějak a-priori ovlivněna

5. Tušení příčiny

Jaké jsou (empiricky vypořizované) důvody, které způsobí, že daná množina číselných dat **nesplňuje** benfordovské rozdělení?

- příliš málo dat (obecný statistický požadavek)
- data, která jsou nějak a-priori ovlivněna (telefonní čísla,

5. Tušení příčiny

Jaké jsou (empiricky vypořizované) důvody, které způsobí, že daná množina číselných dat **nesplňuje** benfordovské rozdělení?

- příliš málo dat (obecný statistický požadavek)
- data, která jsou nějak a-priori ovlivněna (telefonní čísla, "baťovské" ceny)

5. Tušení příčiny

Jaké jsou (empiricky vypořizované) důvody, které způsobí, že daná množina číselných dat **nesplňuje** benfordovské rozdělení?

- příliš málo dat (obecný statistický požadavek)
- data, která jsou nějak a-priori ovlivněna (telefonní čísla, "baťovské" ceny)
- příliš úzce vymezená soustava dat

5. Tušení příčiny

Jaké jsou (empiricky vypořizované) důvody, které způsobí, že daná množina číselných dat **nesplňuje** benfordovské rozdělení?

- příliš málo dat (obecný statistický požadavek)
- data, která jsou nějak a-priori ovlivněna (telefonní čísla, "baťovské" ceny)
- příliš úzce vymezená soustava dat (výška lidí v cm,

5. Tušení příčiny

Jaké jsou (empiricky vypořizované) důvody, které způsobí, že daná množina číselných dat **nesplňuje** benfordovské rozdělení?

- příliš málo dat (obecný statistický požadavek)
- data, která jsou nějak a-priori ovlivněna (telefonní čísla, "baťovské" ceny)
- příliš úzce vymezená soustava dat (výška lidí v cm, ale někdy také také data pouze jednoho či málo řádů)

5. Tušení příčiny

Jaké jsou (empiricky vypožorované) důvody, které způsobí, že daná množina číselných dat **nesplňuje** benfordovské rozdělení?

- příliš málo dat (obecný statistický požadavek)
- data, která jsou nějak a-priori ovlivněna (telefonní čísla, "baťovské" ceny)
- příliš úzce vymezená soustava dat (výška lidí v cm, ale někdy také také data pouze jednoho či málo řádů)

Podstata benfordovského rozdělení (stále ještě empiricky/heuristicky)

5. Tušení příčiny

Jaké jsou (empiricky vypořizované) důvody, které způsobí, že daná množina číselných dat **nesplňuje** benfordovské rozdělení?

- příliš málo dat (obecný statistický požadavek)
- data, která jsou nějak a-priori ovlivněna (telefonní čísla, "baťovské" ceny)
- příliš úzce vymezená soustava dat (výška lidí v cm, ale někdy také také data pouze jednoho či málo řádů)

Podstata benfordovského rozdělení (stále ještě empiricky/heuristicky)

- **Data, která nejsou deformována výše uvedenými podmínkami, musí navíc podléhat nějakému "univerzálnímu přírodnímu zákonu".**

■ Klíčový moment:

■ Klíčový moment:

Pinkham, R.S.: On the distribution of first significant digits.
Annals of Mathematical Statistics, 32, pp 1223-1230,
1961.

■ Klíčový moment:

Pinkham, R.S.: On the distribution of first significant digits.
Annals of Mathematical Statistics, 32, pp 1223-1230,
1961.

"Pokud existuje něco jako univerzální zákon přírody pro distribuci přirozených dat, tak by se měl projevit nezávisle na tom jaké jednotky měření používáme."

■ Klíčový moment:

Pinkham, R.S.: On the distribution of first significant digits.
Annals of Mathematical Statistics, 32, pp 1223-1230,
1961.

"Pokud existuje něco jako univerzální zákon přírody pro distribuci přirozených dat, tak by se měl projevit nezávisle na tom jaké jednotky měření používáme."

(Přírodě je jedno, jestli měříme v cm nebo palcích nebo yardech.)

■ Klíčový moment:

Pinkham, R.S.: On the distribution of first significant digits.
Annals of Mathematical Statistics, 32, pp 1223-1230,
1961.

"Pokud existuje něco jako univerzální zákon přírody pro distribuci přirozených dat, tak by se měl projevit nezávisle na tom jaké jednotky měření používáme."

(Přírodě je jedno, jestli měříme v cm nebo palcích nebo yardech.)

Jak to popsat matematicky?

Přejít od centimetrů k (např.) palcům znamená celou sadu dat přenásobit jistým (stejným) koeficientem.

Přejít od centimetrů k (např.) palcům znamená celou sadu dat přenásobit jistým (stejným) koeficientem.

Pokud existuje nějaké přirozené chování sady "přírodních" dat, nemělo by přenásobení celé sady dat jakýmkoli, ale stejným číslem mít vliv na statistické rozložení prvních cifer nově vzniklé sady dat.

Přejít od centimetrů k (např.) palcům znamená celou sadu dat přenásobit jistým (stejným) koeficientem.

Pokud existuje nějaké přirozené chování sady "přírodních" dat, nemělo by přenásobení celé sady dat jakýmkoli, ale stejným číslem mít vliv na statistické rozložení prvních cifer nově vzniklé sady dat.

OK, délky řek. . .

Přejít od centimetrů k (např.) palcům znamená celou sadu dat přenásobit jistým (stejným) koeficientem.

Pokud existuje nějaké přirozené chování sady "přírodních" dat, nemělo by přenásobení celé sady dat jakýmkoli, ale stejným číslem mít vliv na statistické rozložení prvních cifer nově vzniklé sady dat.

OK, délky řek... výška hor...

Přejít od centimetrů k (např.) palcům znamená celou sadu dat přenásobit jistým (stejným) koeficientem.

Pokud existuje nějaké přirozené chování sady "přírodních" dat, nemělo by přenásobení celé sady dat jakýmkoli, ale stejným číslem mít vliv na statistické rozložení prvních cifer nově vzniklé sady dat.

OK, délky řek... výška hor... bankovní transakce (v korunách nebo eurech?) ...

Přejít od centimetrů k (např.) palcům znamená celou sadu dat přenásobit jistým (stejným) koeficientem.

Pokud existuje nějaké přirozené chování sady "přírodních" dat, nemělo by přenásobení celé sady dat jakýmkoli, ale stejným číslem mít vliv na statistické rozložení prvních cifer nově vzniklé sady dat.

OK, délky řek... výška hor... bankovní transakce (v korunách nebo eurech?)... ale jak to, že například i počty obyvatel?

Přejít od centimetrů k (např.) palcům znamená celou sadu dat přenásobit jistým (stejným) koeficientem.

Pokud existuje nějaké přirozené chování sady "přírodních" dat, nemělo by přenásobení celé sady dat jakýmkoli, ale stejným číslem mít vliv na statistické rozložení prvních cifer nově vzniklé sady dat.

OK, délky řek... výška hor... bankovní transakce (v korunách nebo eurech?)... ale jak to, že například i počty obyvatel?

"Přenásobení" dat o počtech obyvatel může třeba hovořit o počtech nohou všech obyvatel,

Přejít od centimetrů k (např.) palcům znamená celou sadu dat přenásobit jistým (stejným) koeficientem.

Pokud existuje nějaké přirozené chování sady "přírodních" dat, nemělo by přenásobení celé sady dat jakýmkoli, ale stejným číslem mít vliv na statistické rozložení prvních cifer nově vzniklé sady dat.

OK, délky řek... výška hor... bankovní transakce (v korunách nebo eurech?)... ale jak to, že například i počty obyvatel?

"Přenásobení" dat o počtech obyvatel může třeba hovořit o počtech nohou všech obyvatel, počtech průměrně velkých rodin...

Počet obyvatel obcí ČR / 2.5

Cifra	Četnost výskytu první cifry	Relativní četnost výskytu první cifry	Co předpovídá Benford
1	1875	29,99%	30,10%
2	1125	17,99%	17,61%
3	786	12,57%	12,49%
4	549	8,78%	9,69%
5	508	8,12%	7,92%
6	419	6,70%	6,69%
7	340	5,44%	5,80%
8	348	5,57%	5,12%
9	303	4,85%	4,58%

Počet obyvatel obcí ČR / 31 415

Cifra	Četnost výskytu první cifry	Relativní četnost výskytu první cifry	Co předpovídá Benford
1	1900	30,39%	30,10%
2	1108	17,72%	17,61%
3	711	11,37%	12,49%
4	621	9,93%	9,69%
5	466	7,45%	7,92%
6	432	6,91%	6,69%
7	403	6,44%	5,80%
8	322	5,15%	5,12%
9	290	4,64%	4,58%

6. Smělá hypotéza

6. Smělá hypotéza

Podmínka škálovatelnosti: řekneme, že sada kladných číselných dat splňuje **podmínku škálovatelnosti**, pokud přenásobením celé sady dat jakýmkoli, ale stejným kladným číslem nemá vliv na statistické rozložení prvních cifer nově vzniklé sady dat.

6. Smělá hypotéza

Podmínka škálovatelnosti: řekneme, že sada kladných číselných dat splňuje **podmínku škálovatelnosti**, pokud přenásobením celé sady dat jakýmkoli, ale stejným kladným číslem nemá vliv na statistické rozložení prvních cifer nově vzniklé sady dat.

Tvrzení: pokud nějaká dostatečně velká sada číselných dat splňuje **podmínku škálovatelnosti**, pak je rozložení prvních cifer v této sadě dat benfordovské.

7. Hledání hustoty pravděpodobnosti

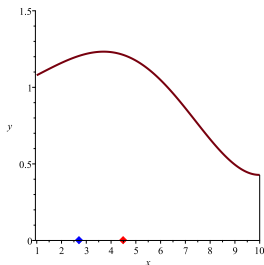
Hledáme funkci f , představující hustotu pravděpodobnosti rozložení prvních cifer dané škálovatelné sady dat.

7. Hledání hustoty pravděpodobnosti

Hledáme funkci f , představující hustotu pravděpodobnosti rozložení prvních cifer dané škálovatelné sady dat. Kde má být definována?

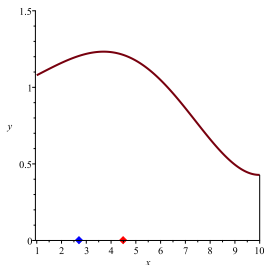
7. Hledání hustoty pravděpodobnosti

Hledáme funkci f , představující hustotu pravděpodobnosti rozložení prvních cifer dané škálovatelné sady dat. Kde má být definována?



7. Hledání hustoty pravděpodobnosti

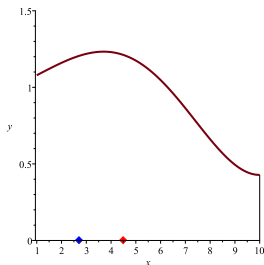
Hledáme funkci f , představující hustotu pravděpodobnosti rozložení prvních cifer dané škálovatelné sady dat. Kde má být definována?



■ 2718,

7. Hledání hustoty pravděpodobnosti

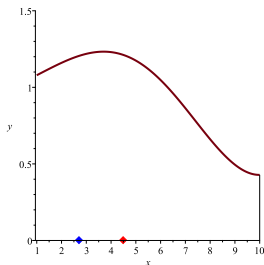
Hledáme funkci f , představující hustotu pravděpodobnosti rozložení prvních cifer dané škálovatelné sady dat. Kde má být definována?



■ 2718, 271.8,

7. Hledání hustoty pravděpodobnosti

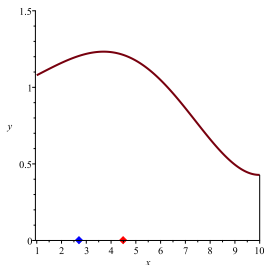
Hledáme funkci f , představující hustotu pravděpodobnosti rozložení prvních cifer dané škálovatelné sady dat. Kde má být definována?



■ 2718, 271.8, 27.18,

7. Hledání hustoty pravděpodobnosti

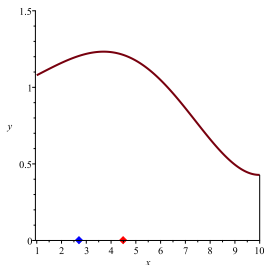
Hledáme funkci f , představující hustotu pravděpodobnosti rozložení prvních cifer dané škálovatelné sady dat. Kde má být definována?



■ 2718, 271.8, 27.18, 2.718

7. Hledání hustoty pravděpodobnosti

Hledáme funkci f , představující hustotu pravděpodobnosti rozložení prvních cifer dané škálovatelné sady dat. Kde má být definována?

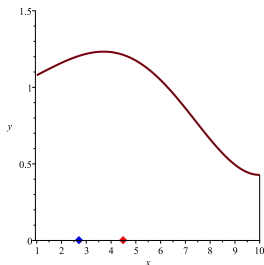


■ 2718, 271.8, 27.18, 2.718

■ 4500,

7. Hledání hustoty pravděpodobnosti

Hledáme funkci f , představující hustotu pravděpodobnosti rozložení prvních cifer dané škálovatelné sady dat. Kde má být definována?

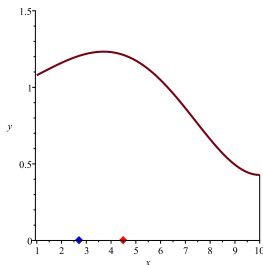


■ 2718, 271.8, 27.18, 2.718

■ 4500, 450.0,

7. Hledání hustoty pravděpodobnosti

Hledáme funkci f , představující hustotu pravděpodobnosti rozložení prvních cifer dané škálovatelné sady dat. Kde má být definována?

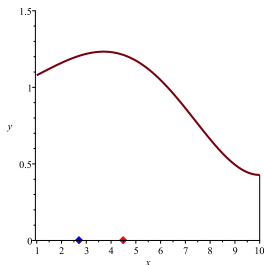


■ 2718, 271.8, 27.18, 2.718

■ 4500, 450.0, 45.00,

7. Hledání hustoty pravděpodobnosti

Hledáme funkci f , představující hustotu pravděpodobnosti rozložení prvních cifer dané škálovatelné sady dat. Kde má být definována?

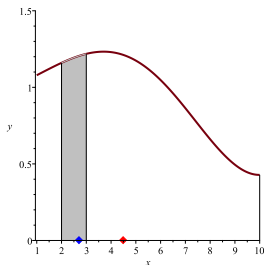


■ 2718, 271.8, 27.18, 2.718

■ 4500, 450.0, 45.00, 4.500

7. Hledání hustoty pravděpodobnosti

Hledáme funkci f , představující hustotu pravděpodobnosti rozložení prvních cifer dané škálovatelné sady dat. Kde má být definována?

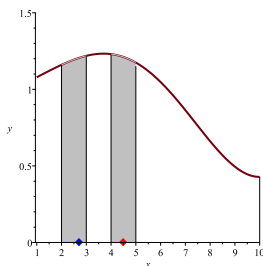


■ 2718, 271.8, 27.18, 2.718 (cifra 2: $\int_2^3 f(x) dx$)

■ 4500, 450.0, 45.00, 4.500

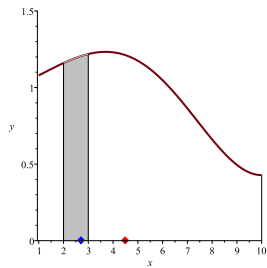
7. Hledání hustoty pravděpodobnosti

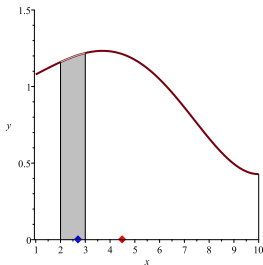
Hledáme funkci f , představující hustotu pravděpodobnosti rozložení prvních cifer dané škálovatelné sady dat. Kde má být definována?



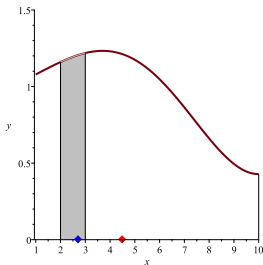
■ 2718, 271.8, 27.18, 2.718 (cifra 2: $\int_2^3 f(x) dx$)

■ 4500, 450.0, 45.00, 4.500 (cifra 4: $\int_4^5 f(x) dx$)



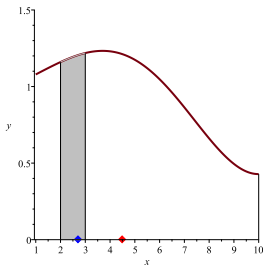


Pravděpodobnost, že číslo začíná cifrou 2, resp.
 $n \in \{1, 2, \dots, 9\}$:



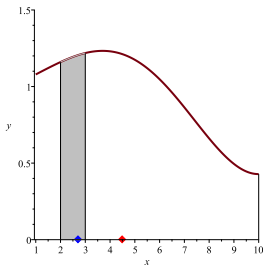
Pravděpodobnost, že číslo začíná cifrou 2, resp.
 $n \in \{1, 2, \dots, 9\}$:

$$p(2) = \int_2^3 f(x) dx$$



Pravděpodobnost, že číslo začíná cifrou 2, resp.
 $n \in \{1, 2, \dots, 9\}$:

$$p(2) = \int_2^3 f(x) dx \quad \text{resp.} \quad p(n) = \int_n^{n+1} f(x) dx,$$

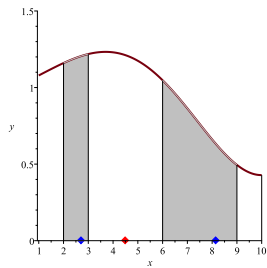


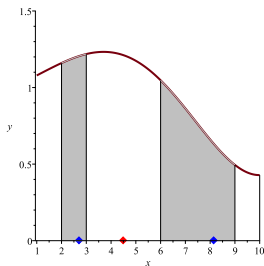
Pravděpodobnost, že číslo začíná cifrou 2, resp.
 $n \in \{1, 2, \dots, 9\}$:

$$p(2) = \int_2^3 f(x) dx \quad \text{resp.} \quad p(n) = \int_n^{n+1} f(x) dx ,$$

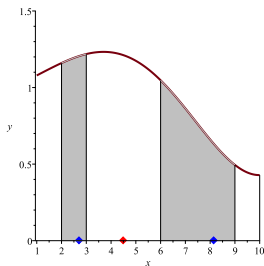
samozřejmě za podmínky

$$\int_1^{10} f(x) dx = 1 .$$



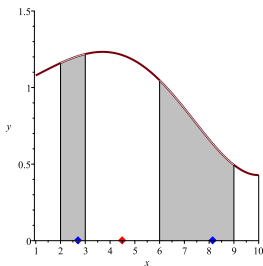


Podmínka škálovatelnosti znamená, že pro všechna $y > 0$ chceme



Podmínka škálovatelnosti znamená, že pro všechna $y > 0$ chceme

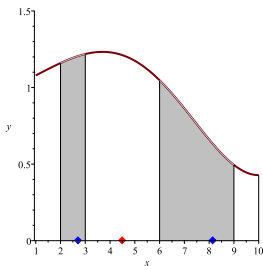
$$\int_n^{n+1} f(x) dx = \int_{y_n}^{y^{(n+1)}} f(x) dx ,$$



Podmínka škálovatelnosti znamená, že pro všechna $y > 0$ chceme

$$\int_n^{n+1} f(x) dx = \int_{y_n}^{y^{(n+1)}} f(x) dx ,$$

ovšem také za podmínky, že

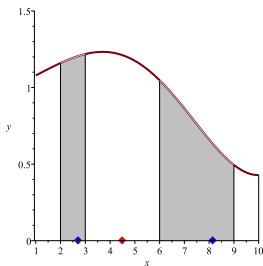


Podmínka škálovatelnosti znamená, že pro všechna $y > 0$ chceme

$$\int_n^{n+1} f(x) dx = \int_{yn}^{y(n+1)} f(x) dx ,$$

ovšem také za podmínky, že

$$\int_y^{10y} f(x) dx = 1$$



Podmínka škálovatelnosti znamená, že pro všechna $y > 0$ chceme

$$\int_n^{n+1} f(x) \, dx = \int_{yn}^{y(n+1)} f(x) \, dx ,$$

ovšem také za podmínky, že

$$\int_y^{10y} f(x) \, dx = 1 \quad \left(= \int_1^{10} f(x) \, dx \right) .$$

Ukážeme dokonce o něco víc:

Ukážeme dokonce o něco víc:

Najdeme (spojitou) funkci $f : (0, +\infty) \rightarrow (0, +\infty)$ takovou, že pro všechna $a > 0$ existuje $C(a) > 0$ taková, že

$$\int_{ya}^{y(a+1)} f(x) dx = C(a) \quad \text{a přitom} \quad \int_y^{10y} f(x) dx = 1,$$

pro všechna $y > 0$.

Máme tedy pro každé (pevné) $a > 0$ a pro všechna $y > 0$:

Máme tedy pro každé (pevné) $a > 0$ a pro všechna $y > 0$:

$$C(a) = \int_{ya}^{y(a+1)} f(x) dx$$

Máme tedy pro každé (pevné) $a > 0$ a pro všechna $y > 0$:

$$C(a) = \int_{ya}^{y(a+1)} f(x) dx = F(y(a+1)) - F(ya),$$

kde F je primitivní k f .

Máme tedy pro každé (pevné) $a > 0$ a pro všechna $y > 0$:

$$C(a) = \int_{ya}^{y(a+1)} f(x) \, dx = F(y(a+1)) - F(ya),$$

kde F je primitivní k f . Derivujme tuto rovnost podle y :

Máme tedy pro každé (pevné) $a > 0$ a pro všechna $y > 0$:

$$C(a) = \int_{ya}^{y(a+1)} f(x) dx = F(y(a+1)) - F(ya),$$

kde F je primitivní k f . Derivujme tuto rovnost podle y :

$$0 = F'(y(a+1))(a+1) - F'(ya)a$$

Máme tedy pro každé (pevné) $a > 0$ a pro všechna $y > 0$:

$$C(a) = \int_{ya}^{y(a+1)} f(x) \, dx = F(y(a+1)) - F(ya),$$

kde F je primitivní k f . Derivujme tuto rovnost podle y :

$$\begin{aligned} 0 &= F'(y(a+1))(a+1) - F'(ya)a \\ &= f(y(a+1))(a+1) - f(ya)a. \end{aligned}$$

Máme tedy pro každé (pevné) $a > 0$ a pro všechna $y > 0$:

$$C(a) = \int_{ya}^{y(a+1)} f(x) dx = F(y(a+1)) - F(ya),$$

kde F je primitivní k f . Derivujme tuto rovnost podle y :

$$\begin{aligned} 0 &= F'(y(a+1))(a+1) - F'(ya)a \\ &= f(y(a+1))(a+1) - f(ya)a. \end{aligned}$$

Položme $y = 1/a$:

Máme tedy pro každé (pevné) $a > 0$ a pro všechna $y > 0$:

$$C(a) = \int_{ya}^{y(a+1)} f(x) dx = F(y(a+1)) - F(ya),$$

kde F je primitivní k f . Derivujme tuto rovnost podle y :

$$\begin{aligned} 0 &= F'(y(a+1))(a+1) - F'(ya)a \\ &= f(y(a+1))(a+1) - f(ya)a. \end{aligned}$$

Položme $y = 1/a$:

$$0 = f\left(\frac{a+1}{a}\right)(a+1) - f(1)a$$

Máme tedy pro každé (pevné) $a > 0$ a pro všechna $y > 0$:

$$C(a) = \int_{ya}^{y(a+1)} f(x) dx = F(y(a+1)) - F(ya),$$

kde F je primitivní k f . Derivujme tuto rovnost podle y :

$$\begin{aligned} 0 &= F'(y(a+1))(a+1) - F'(ya)a \\ &= f(y(a+1))(a+1) - f(ya)a. \end{aligned}$$

Položme $y = 1/a$:

$$0 = f\left(\frac{a+1}{a}\right)(a+1) - f(1)a$$

a vydělme tuto rovnost číslem $a > 0$:

$$f\left(\frac{a+1}{a}\right) \frac{a+1}{a} = f(1)$$

$$f\left(\frac{a+1}{a}\right) \frac{a+1}{a} = f(1) := c.$$

$$f\left(\frac{a+1}{a}\right) \frac{a+1}{a} = f(1) := c.$$

Položme nyní $t = \frac{a+1}{a}$

$$f\left(\frac{a+1}{a}\right) \frac{a+1}{a} = f(1) := c.$$

Položme nyní $t = \frac{a+1}{a}$ (pokud $a > 0$, je $t > 1$):

$$f\left(\frac{a+1}{a}\right) \frac{a+1}{a} = f(1) := c.$$

Položme nyní $t = \frac{a+1}{a}$ (pokud $a > 0$, je $t > 1$):

$$f(t)t = c$$

$$f\left(\frac{a+1}{a}\right) \frac{a+1}{a} = f(1) := c.$$

Položme nyní $t = \frac{a+1}{a}$ (pokud $a > 0$, je $t > 1$):

$$f(t)t = c \quad \implies \quad f(t) = \frac{c}{t} \quad t > 1.$$

$$f\left(\frac{a+1}{a}\right) \frac{a+1}{a} = f(1) := c.$$

Položme nyní $t = \frac{a+1}{a}$ (pokud $a > 0$, je $t > 1$):

$$f(t)t = c \quad \implies \quad f(t) = \frac{c}{t} \quad t > 1.$$

(f lze pochopitelně dodefinovat na celý interval $(0, +\infty)$).

$$f\left(\frac{a+1}{a}\right) \frac{a+1}{a} = f(1) := c.$$

Položme nyní $t = \frac{a+1}{a}$ (pokud $a > 0$, je $t > 1$):

$$f(t)t = c \quad \implies \quad f(t) = \frac{c}{t} \quad t > 1.$$

(f lze pochopitelně dodefinovat na celý interval $(0, +\infty)$).
Konstantu c najdeme z podmínky

$$f\left(\frac{a+1}{a}\right) \frac{a+1}{a} = f(1) := c.$$

Položme nyní $t = \frac{a+1}{a}$ (pokud $a > 0$, je $t > 1$):

$$f(t)t = c \quad \implies \quad f(t) = \frac{c}{t} \quad t > 1.$$

(f lze pochopitelně dodefinovat na celý interval $(0, +\infty)$).
Konstantu c najdeme z podmínky

$$1 = \int_1^{10} f(x) dx$$

$$f\left(\frac{a+1}{a}\right) \frac{a+1}{a} = f(1) := c.$$

Položme nyní $t = \frac{a+1}{a}$ (pokud $a > 0$, je $t > 1$):

$$f(t)t = c \quad \implies \quad f(t) = \frac{c}{t} \quad t > 1.$$

(f lze pochopitelně dodefinovat na celý interval $(0, +\infty)$).
Konstantu c najdeme z podmínky

$$1 = \int_1^{10} f(x) dx = \int_1^{10} \frac{c}{x} dx$$

$$f\left(\frac{a+1}{a}\right) \frac{a+1}{a} = f(1) := c.$$

Položme nyní $t = \frac{a+1}{a}$ (pokud $a > 0$, je $t > 1$):

$$f(t)t = c \quad \implies \quad f(t) = \frac{c}{t} \quad t > 1.$$

(f lze pochopitelně dodefinovat na celý interval $(0, +\infty)$).
Konstantu c najdeme z podmínky

$$1 = \int_1^{10} f(x) dx = \int_1^{10} \frac{c}{x} dx = c(\ln 10 - \ln 1)$$

$$f\left(\frac{a+1}{a}\right) \frac{a+1}{a} = f(1) := c.$$

Položme nyní $t = \frac{a+1}{a}$ (pokud $a > 0$, je $t > 1$):

$$f(t)t = c \quad \implies \quad f(t) = \frac{c}{t} \quad t > 1.$$

(f lze pochopitelně dodefinovat na celý interval $(0, +\infty)$).
Konstantu c najdeme z podmínky

$$1 = \int_1^{10} f(x) dx = \int_1^{10} \frac{c}{x} dx = c(\ln 10 - \ln 1) = c \ln 10.$$

$$f\left(\frac{a+1}{a}\right) \frac{a+1}{a} = f(1) := c.$$

Položme nyní $t = \frac{a+1}{a}$ (pokud $a > 0$, je $t > 1$):

$$f(t)t = c \quad \implies \quad f(t) = \frac{c}{t} \quad t > 1.$$

(f lze pochopitelně dodefinovat na celý interval $(0, +\infty)$).
Konstantu c najdeme z podmínky

$$1 = \int_1^{10} f(x) dx = \int_1^{10} \frac{c}{x} dx = c(\ln 10 - \ln 1) = c \ln 10.$$

Tedy

$$c = \frac{1}{\ln 10}$$

$$f\left(\frac{a+1}{a}\right) \frac{a+1}{a} = f(1) := c.$$

Položme nyní $t = \frac{a+1}{a}$ (pokud $a > 0$, je $t > 1$):

$$f(t)t = c \quad \implies \quad f(t) = \frac{c}{t} \quad t > 1.$$

(f lze pochopitelně dodefinovat na celý interval $(0, +\infty)$).
Konstantu c najdeme z podmínky

$$1 = \int_1^{10} f(x) dx = \int_1^{10} \frac{c}{x} dx = c(\ln 10 - \ln 1) = c \ln 10.$$

Tedy

$$c = \frac{1}{\ln 10}$$

a hledaná funkce má tvar

$$f(x) = \frac{1}{x \ln 10}, \quad x > 0.$$

Ověření podmínek (pro radost):

Ověření podmínek (pro radost):

$$\int_{y^a}^{y^{(a+1)}} \frac{dx}{x \ln 10} dx$$

Ověření podmínek (pro radost):

$$\int_{ya}^{y(a+1)} \frac{dx}{x \ln 10} = \frac{1}{\ln 10} (\ln y(a+1) - \ln(ya))$$

Ověření podmínek (pro radost):

$$\begin{aligned}\int_{ya}^{y(a+1)} \frac{dx}{x \ln 10} &= \frac{1}{\ln 10} (\ln y(a+1) - \ln(ya)) \\ &= \frac{1}{\ln 10} (\ln y + \ln(a+1) - \ln y - \ln a)\end{aligned}$$

Ověření podmíněk (pro radost):

$$\begin{aligned}\int_{ya}^{y(a+1)} \frac{dx}{x \ln 10} &= \frac{1}{\ln 10} (\ln y(a+1) - \ln(ya)) \\ &= \frac{1}{\ln 10} (\ln y + \ln(a+1) - \ln y - \ln a) \\ &= \frac{1}{\ln 10} (\ln(a+1) - \ln a) = C(a) .\end{aligned}$$

Ověření podmínek (pro radost):

$$\begin{aligned}\int_{ya}^{y(a+1)} \frac{dx}{x \ln 10} dx &= \frac{1}{\ln 10} (\ln y(a+1) - \ln(ya)) \\ &= \frac{1}{\ln 10} (\ln y + \ln(a+1) - \ln y - \ln a) \\ &= \frac{1}{\ln 10} (\ln(a+1) - \ln a) = C(a) .\end{aligned}$$

$$\int_y^{10y} \frac{dx}{x \ln 10} dx$$

Ověření podmínek (pro radost):

$$\begin{aligned}\int_{ya}^{y(a+1)} \frac{dx}{x \ln 10} dx &= \frac{1}{\ln 10} (\ln y(a+1) - \ln(ya)) \\ &= \frac{1}{\ln 10} (\ln y + \ln(a+1) - \ln y - \ln a) \\ &= \frac{1}{\ln 10} (\ln(a+1) - \ln a) = C(a) .\end{aligned}$$

$$\int_y^{10y} \frac{dx}{x \ln 10} dx = \frac{1}{\ln 10} (\ln 10y - \ln y)$$

Ověření podmínek (pro radost):

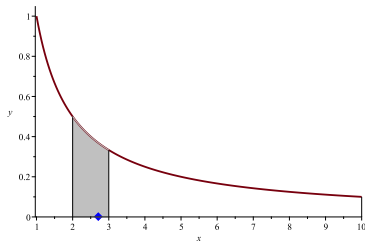
$$\begin{aligned}\int_{ya}^{y(a+1)} \frac{dx}{x \ln 10} dx &= \frac{1}{\ln 10} (\ln y(a+1) - \ln(ya)) \\ &= \frac{1}{\ln 10} (\ln y + \ln(a+1) - \ln y - \ln a) \\ &= \frac{1}{\ln 10} (\ln(a+1) - \ln a) = C(a) .\end{aligned}$$

$$\begin{aligned}\int_y^{10y} \frac{dx}{x \ln 10} dx &= \frac{1}{\ln 10} (\ln 10y - \ln y) \\ &= \frac{1}{\ln 10} (\ln 10 + \ln y - \ln y)\end{aligned}$$

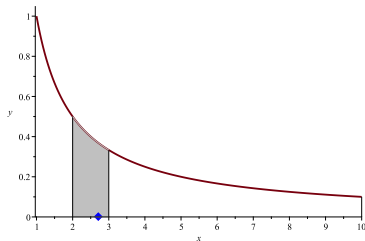
Ověření podmínek (pro radost):

$$\begin{aligned}\int_{ya}^{y(a+1)} \frac{dx}{x \ln 10} dx &= \frac{1}{\ln 10} (\ln y(a+1) - \ln(ya)) \\ &= \frac{1}{\ln 10} (\ln y + \ln(a+1) - \ln y - \ln a) \\ &= \frac{1}{\ln 10} (\ln(a+1) - \ln a) = C(a) .\end{aligned}$$

$$\begin{aligned}\int_y^{10y} \frac{dx}{x \ln 10} dx &= \frac{1}{\ln 10} (\ln 10y - \ln y) \\ &= \frac{1}{\ln 10} (\ln 10 + \ln y - \ln y) \\ &= \frac{\ln 10}{\ln 10} = 1 .\end{aligned}$$

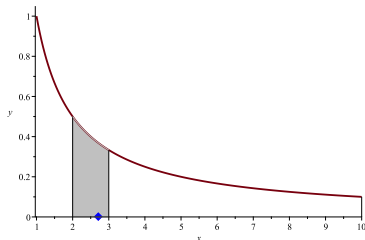


Pravděpodobnost, že
číslo začíná cifrou 2, resp.
 $n \in \{1, 2, \dots, 9\}$ je tedy



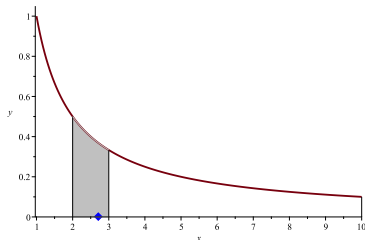
Pravděpodobnost, že
číslo začíná cifrou 2, resp.
 $n \in \{1, 2, \dots, 9\}$ je tedy

$$p(2) = \int_2^3 \frac{dx}{x \ln 10}$$



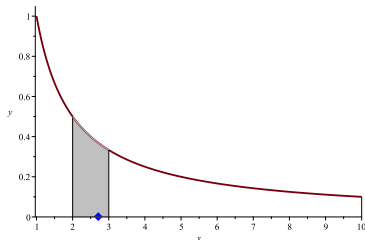
Pravděpodobnost, že
číslo začíná cifrou 2, resp.
 $n \in \{1, 2, \dots, 9\}$ je tedy

$$p(2) = \int_2^3 \frac{dx}{x \ln 10} = \frac{\ln 3}{\ln 10} - \frac{\ln 2}{\ln 10}$$



Pravděpodobnost, že
číslo začíná cifrou 2, resp.
 $n \in \{1, 2, \dots, 9\}$ je tedy

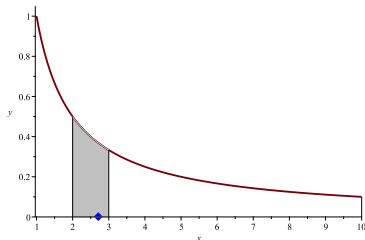
$$\begin{aligned} p(2) &= \int_2^3 \frac{dx}{x \ln 10} = \frac{\ln 3}{\ln 10} - \frac{\ln 2}{\ln 10} \\ &= \log 3 - \log 2, \end{aligned}$$



Pravděpodobnost, že
číslo začíná cifrou 2, resp.
 $n \in \{1, 2, \dots, 9\}$ je tedy

$$p(2) = \int_2^3 \frac{dx}{x \ln 10} = \frac{\ln 3}{\ln 10} - \frac{\ln 2}{\ln 10} \\ = \log 3 - \log 2,$$

resp.

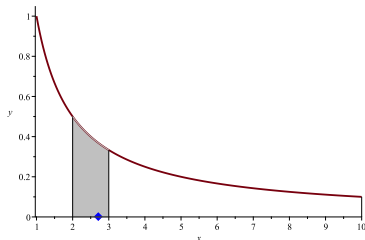


Pravděpodobnost, že
číslo začíná cifrou 2, resp.
 $n \in \{1, 2, \dots, 9\}$ je tedy

$$p(2) = \int_2^3 \frac{dx}{x \ln 10} = \frac{\ln 3}{\ln 10} - \frac{\ln 2}{\ln 10} \\ = \log 3 - \log 2,$$

resp.

$$p(n) = \int_n^{n+1} \frac{dx}{x \ln 10}$$

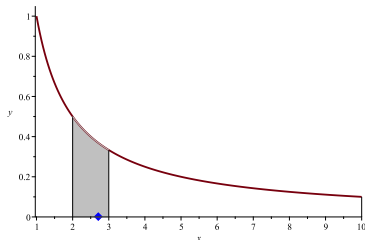


Pravděpodobnost, že
číslo začíná cifrou 2, resp.
 $n \in \{1, 2, \dots, 9\}$ je tedy

$$p(2) = \int_2^3 \frac{dx}{x \ln 10} = \frac{\ln 3}{\ln 10} - \frac{\ln 2}{\ln 10} \\ = \log 3 - \log 2,$$

resp.

$$p(n) = \int_n^{n+1} \frac{dx}{x \ln 10} = \frac{\ln(n+1)}{\ln 10} - \frac{\ln n}{\ln 10}$$



Pravděpodobnost, že
číslo začíná cifrou 2, resp.
 $n \in \{1, 2, \dots, 9\}$ je tedy

$$p(2) = \int_2^3 \frac{dx}{x \ln 10} = \frac{\ln 3}{\ln 10} - \frac{\ln 2}{\ln 10} \\ = \log 3 - \log 2,$$

resp.

$$p(n) = \int_n^{n+1} \frac{dx}{x \ln 10} = \frac{\ln(n+1)}{\ln 10} - \frac{\ln n}{\ln 10} \\ = \log(n+1) - \log n.$$

A tato tabulka už tu jednou byla:

A tato tabulka už tu jednou byla:

Počáteční cifra	výpočet	v procentech
1	$\log 2 - \log 1 = 0,30103$	30,10%
2	$\log 3 - \log 2 = 0,17609$	17,61%
3	$\log 4 - \log 3 = 0,12494$	12,49%
4	$\log 5 - \log 4 = 0,09691$	9,69%
5	$\log 6 - \log 5 = 0,07918$	7,92%
6	$\log 7 - \log 6 = 0,06695$	6,69%
7	$\log 8 - \log 7 = 0,05799$	5,80%
8	$\log 9 - \log 8 = 0,05115$	5,12%
9	$\log 10 - \log 9 = 0,04576$	4,58%

A tato tabulka už tu jednou byla:

Počáteční cifra	výpočet	v procentech
1	$\log 2 - \log 1 = 0,30103$	30,10%
2	$\log 3 - \log 2 = 0,17609$	17,61%
3	$\log 4 - \log 3 = 0,12494$	12,49%
4	$\log 5 - \log 4 = 0,09691$	9,69%
5	$\log 6 - \log 5 = 0,07918$	7,92%
6	$\log 7 - \log 6 = 0,06695$	6,69%
7	$\log 8 - \log 7 = 0,05799$	5,80%
8	$\log 9 - \log 8 = 0,05115$	5,12%
9	$\log 10 - \log 9 = 0,04576$	4,58%

Pokud nějaká dostatečně velká sada číselných dat splňuje **podmínku škálovatelnosti**, pak je rozložení prvních cifer v této sadě dat benfordovské.

8. A co další cifry?

8. A co další cifry?

"Dobrá teorie musí nejen vysvětlit již známé, ale také předpovědět něco neznámého."

8. A co další cifry?

"Dobrá teorie musí nejen vysvětlit již známé, ale také předpovědět něco neznámého."

Uměli bychom podle námi dosaženého výsledku správně predikovat i rozložení **druhých** (nebo i dalších) cifer?

8. A co další cifry?

"Dobrá teorie musí nejen vysvětlit již známé, ale také předpovědět něco neznámého."

Uměli bychom podle námi dosaženého výsledku správně predikovat i rozložení **druhých** (nebo i dalších) cifer?

Co to znamená, že číslo začíná dvojicí čísel 23?

8. A co další cifry?

"Dobrá teorie musí nejen vysvětlit již známé, ale také předpovědět něco neznámého."

Uměli bychom podle námi dosaženého výsledku správně predikovat i rozložení **druhých** (nebo i dalších) cifer?

Co to znamená, že číslo začíná dvojicí čísel 23?

$$\int_{2.3}^{2.4} \frac{dx}{x \ln 10}$$

8. A co další cifry?

"Dobrá teorie musí nejen vysvětlit již známé, ale také předpovědět něco neznámého."

Uměli bychom podle námi dosaženého výsledku správně predikovat i rozložení **druhých** (nebo i dalších) cifer?

Co to znamená, že číslo začíná dvojicí čísel 23?

$$\int_{2.3}^{2.4} \frac{dx}{x \ln 10} = \log 2.4 - \log 2.3$$

8. A co další cifry?

"Dobrá teorie musí nejen vysvětlit již známé, ale také předpovědět něco neznámého."

Uměli bychom podle námi dosaženého výsledku správně predikovat i rozložení **druhých** (nebo i dalších) cifer?

Co to znamená, že číslo začíná dvojicí čísel 23?

$$\int_{2.3}^{2.4} \frac{dx}{x \ln 10} = \log 2.4 - \log 2.3 = \log \frac{2.4}{2.3}$$

8. A co další cifry?

"Dobrá teorie musí nejen vysvětlit již známé, ale také předpovědět něco neznámého."

Uměli bychom podle námi dosaženého výsledku správně predikovat i rozložení **druhých** (nebo i dalších) cifer?

Co to znamená, že číslo začíná dvojicí čísel 23?

$$\int_{2.3}^{2.4} \frac{dx}{x \ln 10} = \log 2.4 - \log 2.3 = \log \frac{2.4}{2.3} = \log \frac{24}{23}$$

8. A co další cifry?

"Dobrá teorie musí nejen vysvětlit již známé, ale také předpovědět něco neznámého."

Uměli bychom podle námi dosaženého výsledku správně predikovat i rozložení **druhých** (nebo i dalších) cifer?

Co to znamená, že číslo začíná dvojicí čísel 23?

$$\int_{2.3}^{2.4} \frac{dx}{x \ln 10} = \log 2.4 - \log 2.3 = \log \frac{2.4}{2.3} = \log \frac{24}{23} = \int_{23}^{24} \frac{dx}{x \ln 10}$$

8. A co další cifry?

"Dobrá teorie musí nejen vysvětlit již známé, ale také předpovědět něco neznámého."

Uměli bychom podle námi dosaženého výsledku správně predikovat i rozložení **druhých** (nebo i dalších) cifer?

Co to znamená, že číslo začíná dvojicí čísel 23?

$$\begin{aligned}\int_{2.3}^{2.4} \frac{dx}{x \ln 10} &= \log 2.4 - \log 2.3 = \log \frac{2.4}{2.3} = \log \frac{24}{23} = \int_{23}^{24} \frac{dx}{x \ln 10} \\ &= \log \left(1 + \frac{1}{23} \right).\end{aligned}$$

Pro kontrolu: pravděpodobnost, že číslo začíná některou z dvojic $\{10, 11, \dots, 18, 19\}$

Pro kontrolu: pravděpodobnost, že číslo začíná některou z dvojic $\{10, 11, \dots, 18, 19\}$ (Všimněte si, že na druhém místě už může být nula.)

Pro kontrolu: pravděpodobnost, že číslo začíná některou z dvojic $\{10, 11, \dots, 18, 19\}$ (Všimněte si, že na druhém místě už může být nula.) musí být rovna pravděpodobosti, že číslo začíná jedničkou:

Pro kontrolu: pravděpodobnost, že číslo začíná některou z dvojic $\{10, 11, \dots, 18, 19\}$ (Všimněte si, že na druhém místě už může být nula.) musí být rovna pravděpodobosti, že číslo začíná jedničkou:

$$\log \frac{11}{10} + \log \frac{12}{11} + \dots + \log \frac{19}{18} + \log \frac{20}{19}$$

Pro kontrolu: pravděpodobnost, že číslo začíná některou z dvojic $\{10, 11, \dots, 18, 19\}$ (Všimněte si, že na druhém místě už může být nula.) musí být rovna pravděpodobosti, že číslo začíná jedničkou:

$$\begin{aligned} \log \frac{11}{10} + \log \frac{12}{11} + \dots + \log \frac{19}{18} + \log \frac{20}{19} \\ = \log \left(\frac{11}{10} \cdot \frac{12}{11} \cdot \frac{13}{12} \dots \frac{19}{18} \cdot \frac{20}{19} \right) \end{aligned}$$

Pro kontrolu: pravděpodobnost, že číslo začíná některou z dvojic $\{10, 11, \dots, 18, 19\}$ (Všimněte si, že na druhém místě už může být nula.) musí být rovna pravděpodobosti, že číslo začíná jedničkou:

$$\begin{aligned} \log \frac{11}{10} + \log \frac{12}{11} + \dots + \log \frac{19}{18} + \log \frac{20}{19} \\ = \log \left(\frac{11}{10} \cdot \frac{12}{11} \cdot \frac{13}{12} \cdot \dots \cdot \frac{19}{18} \cdot \frac{20}{19} \right) \\ = \log \frac{20}{10} \end{aligned}$$

Pro kontrolu: pravděpodobnost, že číslo začíná některou z dvojic $\{10, 11, \dots, 18, 19\}$ (Všimněte si, že na druhém místě už může být nula.) musí být rovna pravděpodobosti, že číslo začíná jedničkou:

$$\begin{aligned} & \log \frac{11}{10} + \log \frac{12}{11} + \dots + \log \frac{19}{18} + \log \frac{20}{19} \\ &= \log \left(\frac{11}{10} \cdot \frac{12}{11} \cdot \frac{13}{12} \dots \frac{19}{18} \cdot \frac{20}{19} \right) \\ &= \log \frac{20}{10} \\ &= \log 2. \end{aligned}$$

A teď: co to znamená, že číslo má na druhé pozici cifru 3?

A teď: co to znamená, že **číslo má na druhé pozici cifru 3**?
To znamená, že první dvojice cifer v jeho zápisu začíná některou z dvojic **13, 23, ... 83, 93**:

A teď: co to znamená, že **číslo má na druhé pozici cifru 3**?
To znamená, že první dvojice cifer v jeho zápisu začíná některou z dvojic **13, 23, ... 83, 93**:

$$p_2(3) = \log \frac{14}{13} + \log \frac{24}{23} + \cdots + \log \frac{84}{83} + \log \frac{94}{93}.$$

A teď: co to znamená, že **číslo má na druhé pozici cifru 3**?
To znamená, že první dvojice cifer v jeho zápisu začíná některou z dvojic **13, 23, ... 83, 93**:

$$\begin{aligned} p_2(3) &= \log \frac{14}{13} + \log \frac{24}{23} + \cdots + \log \frac{84}{83} + \log \frac{94}{93} . \\ &= \log \left(1 + \frac{1}{13} \right) + \log \left(1 + \frac{1}{23} \right) + \cdots + \log \left(1 + \frac{1}{93} \right) . \end{aligned}$$

A teď: co to znamená, že **číslo má na druhé pozici cifru 3**?
To znamená, že první dvojice cifer v jeho zápisu začíná některou z dvojic **13, 23, ... 83, 93**:

$$\begin{aligned} p_2(3) &= \log \frac{14}{13} + \log \frac{24}{23} + \cdots + \log \frac{84}{83} + \log \frac{94}{93} . \\ &= \log \left(1 + \frac{1}{13} \right) + \log \left(1 + \frac{1}{23} \right) + \cdots + \log \left(1 + \frac{1}{93} \right) . \end{aligned}$$

To už nejde jednoduše upravit, je potřeba se smířit s tím, že

A teď: co to znamená, že **číslo má na druhé pozici cifru 3**?
To znamená, že první dvojice cifer v jeho zápisu začíná některou z dvojic **13, 23, ... 83, 93**:

$$\begin{aligned} p_2(3) &= \log \frac{14}{13} + \log \frac{24}{23} + \cdots + \log \frac{84}{83} + \log \frac{94}{93} . \\ &= \log \left(1 + \frac{1}{13} \right) + \log \left(1 + \frac{1}{23} \right) + \cdots + \log \left(1 + \frac{1}{93} \right) . \end{aligned}$$

To už nejde jednoduše upravit, je potřeba se smířit s tím, že

$$p_2(3) = \sum_{k=1}^9 \log \left(1 + \frac{1}{10k + 3} \right)$$

A teď: co to znamená, že **číslo má na druhé pozici cifru 3**?
To znamená, že první dvojice cifer v jeho zápisu začíná některou z dvojic **13, 23, ... 83, 93**:

$$\begin{aligned} p_2(3) &= \log \frac{14}{13} + \log \frac{24}{23} + \cdots + \log \frac{84}{83} + \log \frac{94}{93} . \\ &= \log \left(1 + \frac{1}{13} \right) + \log \left(1 + \frac{1}{23} \right) + \cdots + \log \left(1 + \frac{1}{93} \right) . \end{aligned}$$

To už nejde jednoduše upravit, je potřeba se smířit s tím, že

$$p_2(3) = \sum_{k=1}^9 \log \left(1 + \frac{1}{10k+3} \right)$$

a obecně

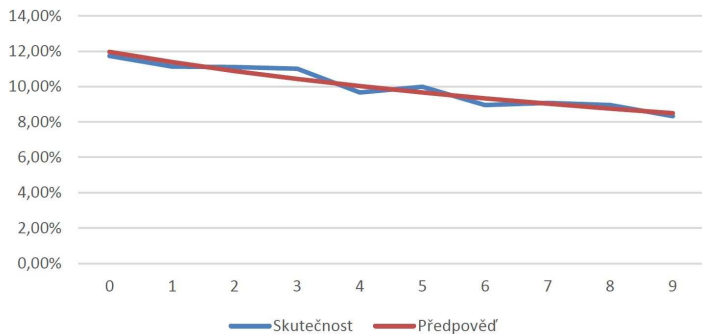
$$p_2(n) = \sum_{k=1}^9 \log \left(1 + \frac{1}{10k+n} \right), \quad n = 0, 1, \dots, 9.$$

Počet obyvatel obcí ČR

6253

Cifra	Četnost výskytu druhé cifry	Relativní četnost výskytu druhé cifry	Co předpovídá Benford
0	734	11,74%	11,97%
1	696	11,13%	11,39%
2	695	11,11%	10,88%
3	689	11,02%	10,43%
4	605	9,68%	10,03%
5	625	10,00%	9,67%
6	560	8,96%	9,34%
7	568	9,08%	9,04%
8	560	8,96%	8,76%
9	521	8,33%	8,50%

Druhé cifry



Podobně pro **třetí** cifry:

Podobně pro **třetí** cifry:

Číslo 4 na třetí pozici znamená uvažovat trojice $\{104, 114, 124, \dots, 984, 994\}$, s odpovídající pravděpodobností

$$p_3(4) = \log \left(1 + \frac{1}{104} \right) + \log \left(1 + \frac{1}{114} \right) + \dots + \log \left(1 + \frac{1}{994} \right).$$

Podobně pro **třetí** cifry:

Číslo 4 na třetí pozici znamená uvažovat trojice $\{104, 114, 124, \dots, 984, 994\}$, s odpovídající pravděpodobností

$$p_3(4) = \log \left(1 + \frac{1}{104} \right) + \log \left(1 + \frac{1}{114} \right) + \dots + \log \left(1 + \frac{1}{994} \right).$$

což dá v tomto případě

Podobně pro **třetí** cifry:

Číslo 4 na třetí pozici znamená uvažovat trojice $\{104, 114, 124, \dots, 984, 994\}$, s odpovídající pravděpodobností

$$p_3(4) = \log \left(1 + \frac{1}{104} \right) + \log \left(1 + \frac{1}{114} \right) + \dots + \log \left(1 + \frac{1}{994} \right).$$

což dá v tomto případě

$$p_3(4) = \sum_{k=10}^{99} \log \left(1 + \frac{1}{10k+4} \right)$$

Podobně pro **třetí** cifry:

Číslo 4 na třetí pozici znamená uvažovat trojice $\{104, 114, 124, \dots, 984, 994\}$, s odpovídající pravděpodobností

$$p_3(4) = \log \left(1 + \frac{1}{104} \right) + \log \left(1 + \frac{1}{114} \right) + \dots + \log \left(1 + \frac{1}{994} \right).$$

což dá v tomto případě

$$p_3(4) = \sum_{k=10}^{99} \log \left(1 + \frac{1}{10k+4} \right)$$

a obecně

$$p_3(n) = \sum_{k=10}^{99} \log \left(1 + \frac{1}{10k+n} \right), \quad n = 0, 1, \dots, 9.$$

Výskyt cifer na první, druhé a třetí pozici

cifra	1. pozice	2. pozice	3.pozice
0	0,00%	11,97%	10,18%
1	30,10%	11,39%	10,14%
2	17,61%	10,88%	10,10%
3	12,49%	10,43%	10,06%
4	9,69%	10,03%	10,02%
5	7,92%	9,67%	9,98%
6	6,69%	9,34%	9,94%
7	5,80%	9,04%	9,90%
8	5,12%	8,76%	9,86%
9	4,58%	8,50%	9,83%
	100,00%	100,00%	100,00%

9. Limita pro znalce

Podobně se dá odvodit rozložení cifer na N -tém místě studované sady dat.

9. Limita pro znalce

Podobně se dá odvodit rozložení cifer na N -tém místě studované sady dat. Analogicky:

$$p_2(n) = \sum_{k=1}^9 \log \left(1 + \frac{1}{10k+n} \right),$$

9. Limita pro znalce

Podobně se dá odvodit rozložení cifer na N -tém místě studované sady dat. Analogicky:

$$p_2(n) = \sum_{k=1}^9 \log \left(1 + \frac{1}{10k+n} \right), \quad p_3(n) = \sum_{k=10}^{99} \log \left(1 + \frac{1}{10k+n} \right),$$

9. Limita pro znalce

Podobně se dá odvodit rozložení cifer na N -tém místě studované sady dat. Analogicky:

$$p_2(n) = \sum_{k=1}^9 \log \left(1 + \frac{1}{10k+n} \right), \quad p_3(n) = \sum_{k=10}^{99} \log \left(1 + \frac{1}{10k+n} \right),$$

$$n = 0, 1, \dots, 9.$$

9. Limita pro znalce

Podobně se dá odvodit rozložení cifer na N -tém místě studované sady dat. Analogicky:

$$p_2(n) = \sum_{k=1}^9 \log \left(1 + \frac{1}{10k+n} \right), \quad p_3(n) = \sum_{k=10}^{99} \log \left(1 + \frac{1}{10k+n} \right),$$

$n = 0, 1, \dots, 9$. Není těžké nahlédnout, že skutečně platí

9. Limita pro znalce

Podobně se dá odvodit rozložení cifer na N -tém místě studované sady dat. Analogicky:

$$p_2(n) = \sum_{k=1}^9 \log \left(1 + \frac{1}{10k+n} \right), \quad p_3(n) = \sum_{k=10}^{99} \log \left(1 + \frac{1}{10k+n} \right),$$

$n = 0, 1, \dots, 9$. Není těžké nahlédnout, že skutečně platí

$$p_N(n) = \sum_{k=10^{N-2}}^{10^{N-1}-1} \log \left(1 + \frac{1}{10k+n} \right), \quad n = 0, 1, \dots, 9.$$

9. Limita pro znalce

Podobně se dá odvodit rozložení cifer na N -tém místě studované sady dat. Analogicky:

$$p_2(n) = \sum_{k=1}^9 \log \left(1 + \frac{1}{10k+n} \right), \quad p_3(n) = \sum_{k=10}^{99} \log \left(1 + \frac{1}{10k+n} \right),$$

$n = 0, 1, \dots, 9$. Není těžké nahlédnout, že skutečně platí

$$p_N(n) = \sum_{k=10^{N-2}}^{10^{N-1}-1} \log \left(1 + \frac{1}{10k+n} \right), \quad n = 0, 1, \dots, 9.$$

Přirozeně vzniká otázka, co se s tímto výrazem bude dít, kdy budeme uvažovat $N \rightarrow \infty$.

9. Limita pro znalce

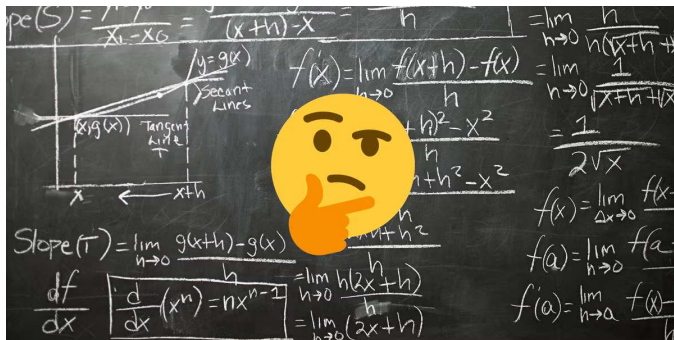
Podobně se dá odvodit rozložení cifer na N -tém místě studované sady dat. Analogicky:

$$p_2(n) = \sum_{k=1}^9 \log \left(1 + \frac{1}{10k+n} \right), \quad p_3(n) = \sum_{k=10}^{99} \log \left(1 + \frac{1}{10k+n} \right),$$

$n = 0, 1, \dots, 9$. Není těžké nahlédnout, že skutečně platí

$$p_N(n) = \sum_{k=10^{N-2}}^{10^{N-1}-1} \log \left(1 + \frac{1}{10k+n} \right), \quad n = 0, 1, \dots, 9.$$

Přirozeně vzniká otázka, co se s tímto výrazem bude dít, kdy budeme uvažovat $N \rightarrow \infty$. (Pochopitelně je dobré mít pro numerické ověření například data typu desetinná či dokonce "iracionální" čísla).



$$\lim_{N \rightarrow \infty} \sum_{k=10^{N-2}}^{10^N-1-1} \log \left(1 + \frac{1}{10k+n} \right) = ?$$

Odhad hodnoty limity:

Odhad hodnoty limity: $f(N) \approx g(N)$ znamená $\lim \frac{f(N)}{g(N)} = 1$.

Odhad hodnoty limity: $f(N) \approx g(N)$ znamená $\lim \frac{f(N)}{g(N)} = 1$.

$$\sum_{k=10^{N-2}}^{10^{N-1}-1} \log \left(1 + \frac{1}{10k+n} \right)$$

Odhad hodnoty limity: $f(N) \approx g(N)$ znamená $\lim \frac{f(N)}{g(N)} = 1$.

$$\sum_{k=10^{N-2}}^{10^{N-1}-1} \log \left(1 + \frac{1}{10k+n} \right)$$

$$\left| \log x = \frac{\ln x}{\ln 10} \right.$$

Odhad hodnoty limity: $f(N) \approx g(N)$ znamená $\lim \frac{f(N)}{g(N)} = 1$.

$$\sum_{k=10^{N-2}}^{10^{N-1}-1} \log \left(1 + \frac{1}{10k+n} \right)$$

$$\left| \log x = \frac{\ln x}{\ln 10} \right.$$

$$= \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \ln \left(1 + \frac{1}{10k+n} \right)$$

Odhad hodnoty limity: $f(N) \approx g(N)$ znamená $\lim \frac{f(N)}{g(N)} = 1$.

$$\begin{aligned} \sum_{k=10^{N-2}}^{10^{N-1}-1} \log \left(1 + \frac{1}{10k+n} \right) & \quad \left| \quad \log x = \frac{\ln x}{\ln 10} \right. \\ &= \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \ln \left(1 + \frac{1}{10k+n} \right) \quad \left| \quad \ln(1+1/N) \approx 1/N \right. \end{aligned}$$

Odhad hodnoty limity: $f(N) \approx g(N)$ znamená $\lim \frac{f(N)}{g(N)} = 1$.

$$\sum_{k=10^{N-2}}^{10^{N-1}-1} \log \left(1 + \frac{1}{10k+n} \right)$$

$$\left| \log x = \frac{\ln x}{\ln 10} \right.$$

$$= \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \ln \left(1 + \frac{1}{10k+n} \right)$$

$$\left| \ln(1+1/N) \approx 1/N \right.$$

$$\approx \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \frac{1}{10k+n}$$

Odhad hodnoty limity: $f(N) \approx g(N)$ znamená $\lim \frac{f(N)}{g(N)} = 1$.

$$\begin{aligned}
 & \sum_{k=10^{N-2}}^{10^{N-1}-1} \log \left(1 + \frac{1}{10k+n} \right) \quad \left| \quad \log x = \frac{\ln x}{\ln 10} \right. \\
 &= \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \ln \left(1 + \frac{1}{10k+n} \right) \quad \left| \quad \ln(1+1/N) \approx 1/N \right. \\
 &\approx \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \frac{1}{10k+n} \quad \left| \quad \frac{1}{10(k+1)} < \frac{1}{10k+n} < \frac{1}{10k} \right.
 \end{aligned}$$

Odhad hodnoty limity: $f(N) \approx g(N)$ znamená $\lim \frac{f(N)}{g(N)} = 1$.

$$\begin{aligned}
 & \sum_{k=10^{N-2}}^{10^{N-1}-1} \log \left(1 + \frac{1}{10k+n} \right) \quad \left| \quad \log x = \frac{\ln x}{\ln 10} \right. \\
 &= \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \ln \left(1 + \frac{1}{10k+n} \right) \quad \left| \quad \ln(1+1/N) \approx 1/N \right. \\
 &\approx \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \frac{1}{10k+n} \quad \left| \quad \frac{1}{10(k+1)} < \frac{1}{10k+n} < \frac{1}{10k} \right. \\
 &\approx \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \frac{1}{10k}
 \end{aligned}$$

Odhad hodnoty limity: $f(N) \approx g(N)$ znamená $\lim \frac{f(N)}{g(N)} = 1$.

$$\begin{aligned}
 & \sum_{k=10^{N-2}}^{10^{N-1}-1} \log \left(1 + \frac{1}{10k+n} \right) & \left| \log x = \frac{\ln x}{\ln 10} \right. \\
 &= \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \ln \left(1 + \frac{1}{10k+n} \right) & \left| \ln(1+1/N) \approx 1/N \right. \\
 &\approx \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \frac{1}{10k+n} & \left| \frac{1}{10(k+1)} < \frac{1}{10k+n} < \frac{1}{10k} \right. \\
 &\approx \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \frac{1}{10k} & \left| \sum_{k=1}^n \frac{1}{k} \approx \ln n \right.
 \end{aligned}$$

Odhad hodnoty limity: $f(N) \approx g(N)$ znamená $\lim \frac{f(N)}{g(N)} = 1$.

$$\begin{aligned}
 & \sum_{k=10^{N-2}}^{10^{N-1}-1} \log \left(1 + \frac{1}{10k+n} \right) & \left| \log x = \frac{\ln x}{\ln 10} \right. \\
 &= \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \ln \left(1 + \frac{1}{10k+n} \right) & \left| \ln(1+1/N) \approx 1/N \right. \\
 &\approx \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \frac{1}{10k+n} & \left| \frac{1}{10(k+1)} < \frac{1}{10k+n} < \frac{1}{10k} \right. \\
 &\approx \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \frac{1}{10k} & \left| \sum_{k=1}^n \frac{1}{k} \approx \ln n \right. \\
 &\approx \frac{1}{\ln 10} \cdot \frac{1}{10} \ln \left(\frac{10^{N-1} - 1}{10^{N-2} - 1} \right)
 \end{aligned}$$

Odhad hodnoty limity: $f(N) \approx g(N)$ znamená $\lim \frac{f(N)}{g(N)} = 1$.

$$\begin{aligned}
 & \sum_{k=10^{N-2}}^{10^{N-1}-1} \log \left(1 + \frac{1}{10k+n} \right) \quad \left| \quad \log x = \frac{\ln x}{\ln 10} \right. \\
 &= \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \ln \left(1 + \frac{1}{10k+n} \right) \quad \left| \quad \ln(1+1/N) \approx 1/N \right. \\
 &\approx \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \frac{1}{10k+n} \quad \left| \quad \frac{1}{10(k+1)} < \frac{1}{10k+n} < \frac{1}{10k} \right. \\
 &\approx \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \frac{1}{10k} \quad \left| \quad \sum_{k=1}^n \frac{1}{k} \approx \ln n \right. \\
 &\approx \frac{1}{\ln 10} \cdot \frac{1}{10} \ln \left(\frac{10^{N-1}-1}{10^{N-2}-1} \right) \approx \frac{1}{10 \ln 10} \ln 10
 \end{aligned}$$

Odhad hodnoty limity: $f(N) \approx g(N)$ znamená $\lim \frac{f(N)}{g(N)} = 1$.

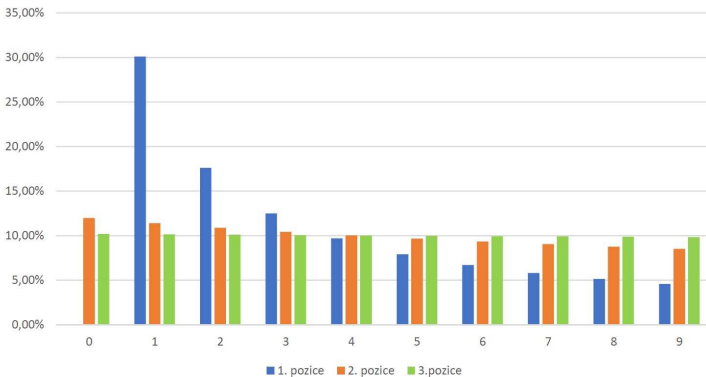
$$\begin{aligned}
 & \sum_{k=10^{N-2}}^{10^{N-1}-1} \log \left(1 + \frac{1}{10k+n} \right) \quad \left| \quad \log x = \frac{\ln x}{\ln 10} \right. \\
 &= \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \ln \left(1 + \frac{1}{10k+n} \right) \quad \left| \quad \ln(1+1/N) \approx 1/N \right. \\
 &\approx \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \frac{1}{10k+n} \quad \left| \quad \frac{1}{10(k+1)} < \frac{1}{10k+n} < \frac{1}{10k} \right. \\
 &\approx \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \frac{1}{10k} \quad \left| \quad \sum_{k=1}^n \frac{1}{k} \approx \ln n \right. \\
 &\approx \frac{1}{\ln 10} \cdot \frac{1}{10} \ln \left(\frac{10^{N-1}-1}{10^{N-2}-1} \right) \approx \frac{1}{10 \ln 10} \ln 10 = \frac{1}{10},
 \end{aligned}$$

Odhad hodnoty limity: $f(N) \approx g(N)$ znamená $\lim \frac{f(N)}{g(N)} = 1$.

$$\begin{aligned}
 & \sum_{k=10^{N-2}}^{10^{N-1}-1} \log \left(1 + \frac{1}{10k+n} \right) \quad \left| \quad \log x = \frac{\ln x}{\ln 10} \right. \\
 &= \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \ln \left(1 + \frac{1}{10k+n} \right) \quad \left| \quad \ln(1+1/N) \approx 1/N \right. \\
 &\approx \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \frac{1}{10k+n} \quad \left| \quad \frac{1}{10(k+1)} < \frac{1}{10k+n} < \frac{1}{10k} \right. \\
 &\approx \frac{1}{\ln 10} \sum_{k=10^{N-2}}^{10^{N-1}-1} \frac{1}{10k} \quad \left| \quad \sum_{k=1}^n \frac{1}{k} \approx \ln n \right. \\
 &\approx \frac{1}{\ln 10} \cdot \frac{1}{10} \ln \left(\frac{10^{N-1}-1}{10^{N-2}-1} \right) \approx \frac{1}{10 \ln 10} \ln 10 = \frac{1}{10},
 \end{aligned}$$

nezávisle na n .

Výskyt cifer na první, druhé a třetí pozici



10. Domácí cvičení pro zvědavé

Lze ukázat, že pro benfordovská data, zapsaná v číselné bázi $b \in \{2, 3, \dots\}$ je hustota pravděpodobnosti rozložení prvních cifer

10. Domácí cvičení pro zvědavé

Lze ukázat, že pro benfordovská data, zapsaná v číselné bázi $b \in \{2, 3, \dots\}$ je hustota pravděpodobnosti rozložení prvních cifer

$$f_b(x) = \frac{1}{x \ln b},$$

10. Domácí cvičení pro zvědavé

Lze ukázat, že pro benfordovská data, zapsaná v číselné bázi $b \in \{2, 3, \dots\}$ je hustota pravděpodobnosti rozložení prvních cifer

$$f_b(x) = \frac{1}{x \ln b},$$

a tedy pravděpodobnost výskytu cifry n na prvním místě takového souboru je

10. Domácí cvičení pro zvědavé

Lze ukázat, že pro benfordovská data, zapsaná v číselné bázi $b \in \{2, 3, \dots\}$ je hustota pravděpodobnosti rozložení prvních cifer

$$f_b(x) = \frac{1}{x \ln b},$$

a tedy pravděpodobnost výskytu cifry n na prvním místě takového souboru je

$$p_1^b(n) = \int_n^{n+1} \frac{dx}{x \ln b},$$

10. Domácí cvičení pro zvědavé

Lze ukázat, že pro benfordovská data, zapsaná v číselné bázi $b \in \{2, 3, \dots\}$ je hustota pravděpodobnosti rozložení prvních cifer

$$f_b(x) = \frac{1}{x \ln b},$$

a tedy pravděpodobnost výskytu cifry n na prvním místě takového souboru je

$$p_1^b(n) = \int_n^{n+1} \frac{dx}{x \ln b},$$

přičemž pro všechna $a, y > 0$ platí:

10. Domácí cvičení pro zvědavé

Lze ukázat, že pro benfordovská data, zapsaná v číselné bázi $b \in \{2, 3, \dots\}$ je hustota pravděpodobnosti rozložení prvních cifer

$$f_b(x) = \frac{1}{x \ln b},$$

a tedy pravděpodobnost výskytu cifry n na prvním místě takového souboru je

$$p_1^b(n) = \int_n^{n+1} \frac{dx}{x \ln b},$$

přičemž pro všechna $a, y > 0$ platí:

$$\int_a^{a+1} \frac{dx}{x \ln b} = \int_{ya}^{y(a+1)} \frac{dx}{x \ln b}, \quad \int_1^b \frac{dx}{x \ln b} = \int_y^{by} \frac{dx}{x \ln b} = 1$$

10. Domácí cvičení pro zvědavé

Lze ukázat, že pro benfordovská data, zapsaná v číselné bázi $b \in \{2, 3, \dots\}$ je hustota pravděpodobnosti rozložení prvních cifer

$$f_b(x) = \frac{1}{x \ln b},$$

a tedy pravděpodobnost výskytu cifry n na prvním místě takového souboru je

$$p_1^b(n) = \int_n^{n+1} \frac{dx}{x \ln b},$$

přičemž pro všechna $a, y > 0$ platí:

$$\int_a^{a+1} \frac{dx}{x \ln b} = \int_{ya}^{y(a+1)} \frac{dx}{x \ln b}, \quad \int_1^b \frac{dx}{x \ln b} = \int_y^{by} \frac{dx}{x \ln b} = 1$$

(škálovatelnost)

10. Domácí cvičení pro zvědavé

Lze ukázat, že pro benfordovská data, zapsaná v číselné bázi $b \in \{2, 3, \dots\}$ je hustota pravděpodobnosti rozložení prvních cifer

$$f_b(x) = \frac{1}{x \ln b},$$

a tedy pravděpodobnost výskytu cifry n na prvním místě takového souboru je

$$p_1^b(n) = \int_n^{n+1} \frac{dx}{x \ln b},$$

přičemž pro všechna $a, y > 0$ platí:

$$\int_a^{a+1} \frac{dx}{x \ln b} = \int_{ya}^{y(a+1)} \frac{dx}{x \ln b},$$

(škálovatelnost)

$$\int_1^b \frac{dx}{x \ln b} = \int_y^{by} \frac{dx}{x \ln b} = 1$$

(normování)

Příklad 1: pravděpodobnost, že ve dvojkové soustavě benfordovská data začínají jedničkou, je:

Příklad 1: pravděpodobnost, že ve dvojkové soustavě benfordovská data začínají jedničkou, je:

$$p_1^2(1) = \int_1^2 \frac{dx}{x \ln 2}$$

Příklad 1: pravděpodobnost, že ve dvojkové soustavě benfordovská data začínají jedničkou, je:

$$p_1^2(1) = \int_1^2 \frac{dx}{x \ln 2} = \frac{\ln 2 - \ln 1}{\ln 2}$$

Příklad 1: pravděpodobnost, že ve dvojkové soustavě benfordovská data začínají jedničkou, je:

$$p_1^2(1) = \int_1^2 \frac{dx}{x \ln 2} = \frac{\ln 2 - \ln 1}{\ln 2} = 1$$

Příklad 1: pravděpodobnost, že ve dvojkové soustavě benfordovská data začínají jedničkou, je:

$$p_1^2(1) = \int_1^2 \frac{dx}{x \ln 2} = \frac{\ln 2 - \ln 1}{\ln 2} = 1,$$

což není překvapivé.

Příklad 1: pravděpodobnost, že ve dvojkové soustavě benfordovská data začínají jedničkou, je:

$$p_1^2(1) = \int_1^2 \frac{dx}{x \ln 2} = \frac{\ln 2 - \ln 1}{\ln 2} = 1,$$

což není překvapivé.

Příklad 2: vyjádření počtu obyvatel všech obcí v (například) pětkové soustavě, vede k tomuto rozložení cifer:

Příklad 1: pravděpodobnost, že ve dvojkové soustavě benfordovská data začínají jedničkou, je:

$$p_1^2(1) = \int_1^2 \frac{dx}{x \ln 2} = \frac{\ln 2 - \ln 1}{\ln 2} = 1,$$

což není překvapivé.

Příklad 2: vyjádření počtu obyvatel všech obcí v (například) pětkové soustavě, vede k tomuto rozložení cifer:

Cifra	Četnost výskytu první cifry	Relativní četnost výskytu první cifry	Co předpovídá Benford
1	2718	43,47%	43,07%
2	1554	24,85%	25,19%
3	1104	17,66%	17,87%
4	877	14,03%	13,86%

Co říci na závěr...

11. Závěr

Co říci na závěr...

... například to, že důležité je také včas skončit.



11. Závěr

Co říci na závěr...

... například to, že důležité je také včas skončit.



Děkuji za pozornost.

Mirko Rokyta
KMA MFF Praha
Sokolovská 83
Praha 8 - Karlín

mirko.rokyta@matfyz.cuni.cz