Kritická práce s daty 2

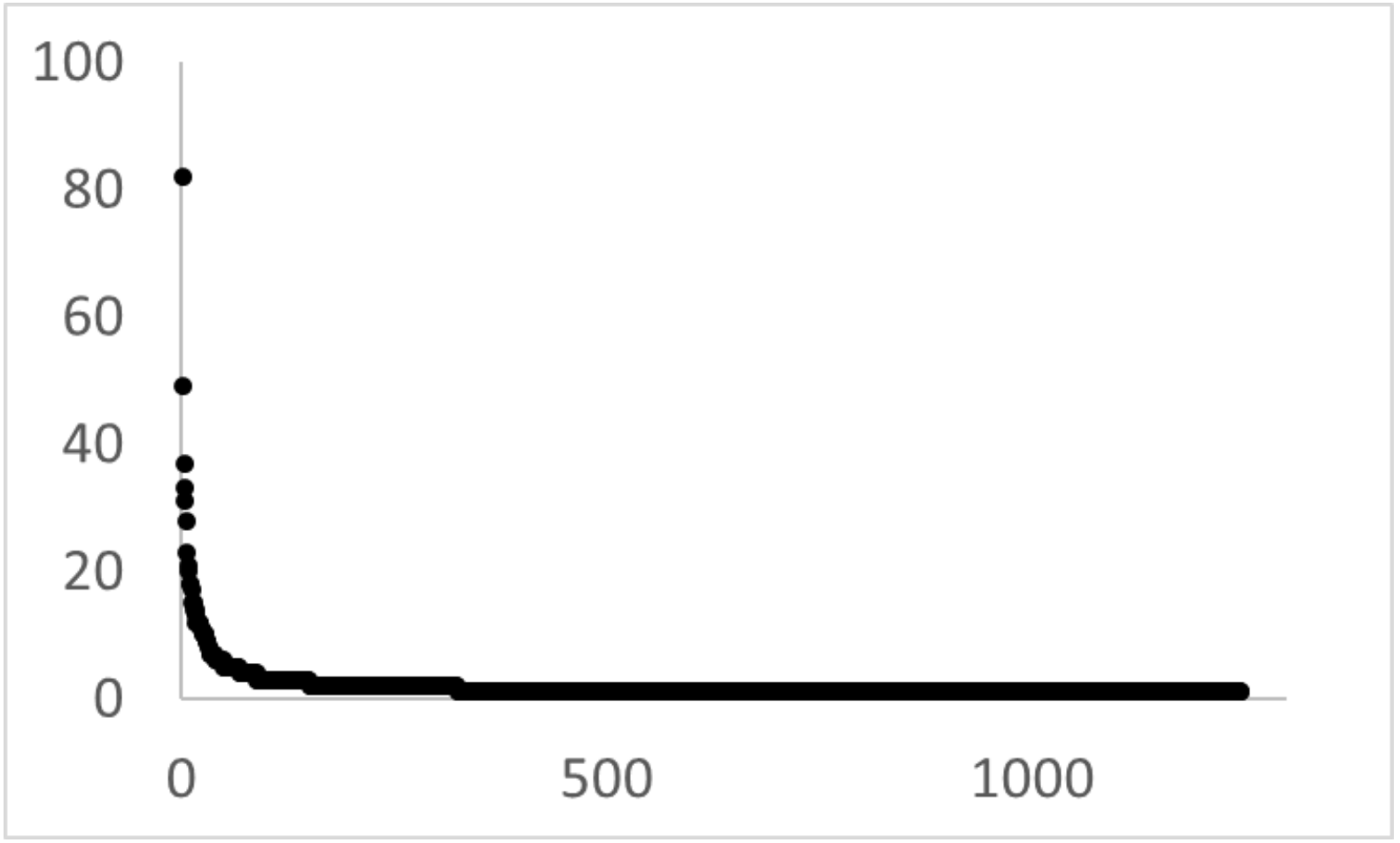
Radek Čech

Frekvence

- smysluplná pouze jako „vztahová“ veličina
  - **distribuce** jednotek určitého typu
    - ranková frekvenční distribuce
    - frekvence délek slov/vět…
    - …
  - **vztah** frekvence a jiných vlastností
    - frekvence slovních druhů vs. typ textu
    - frekvence vs. délka slova
    - frekvence vs. polysémie
    - …

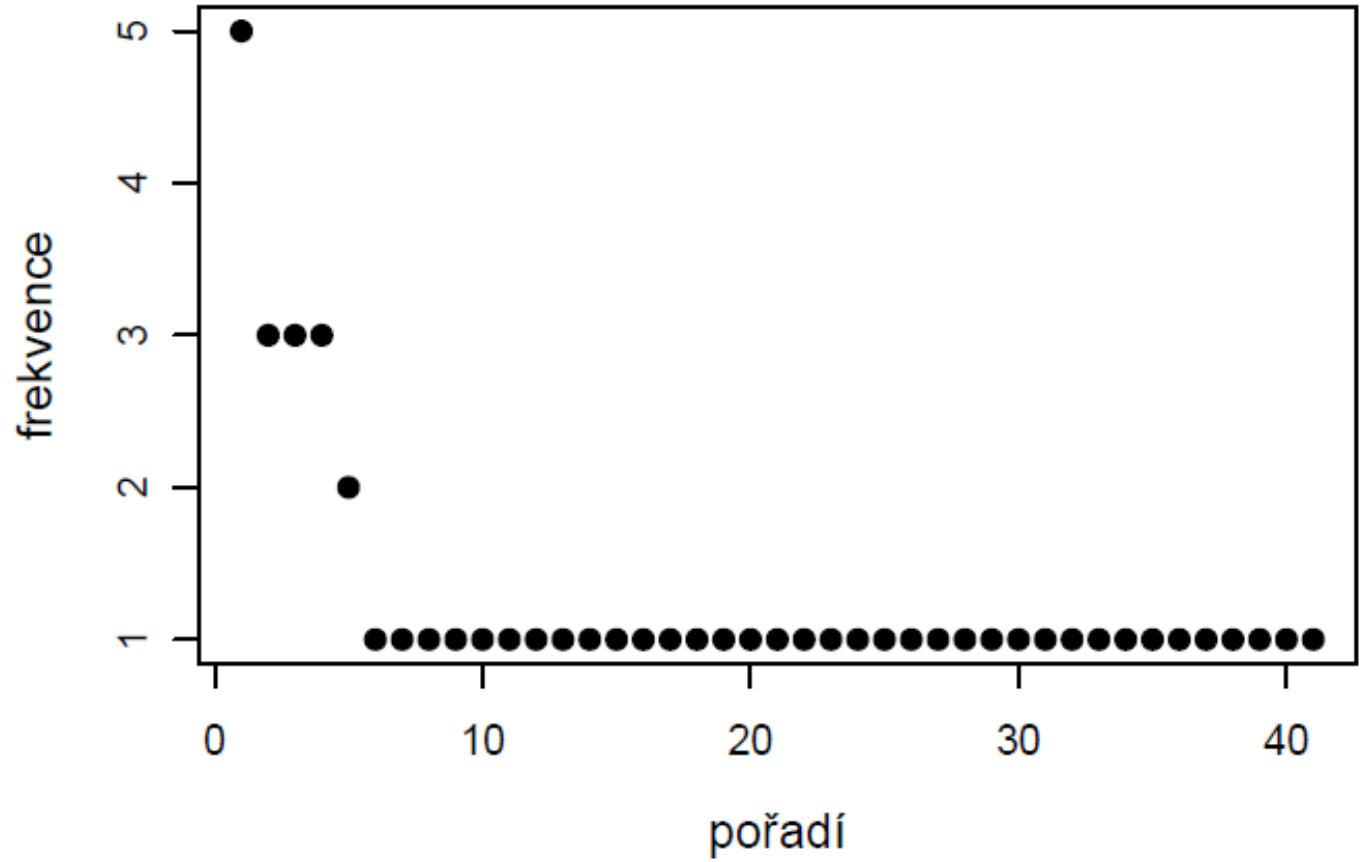
Distribuce jednotek

- Havel 1990: ranková frekvenční distribuce slov



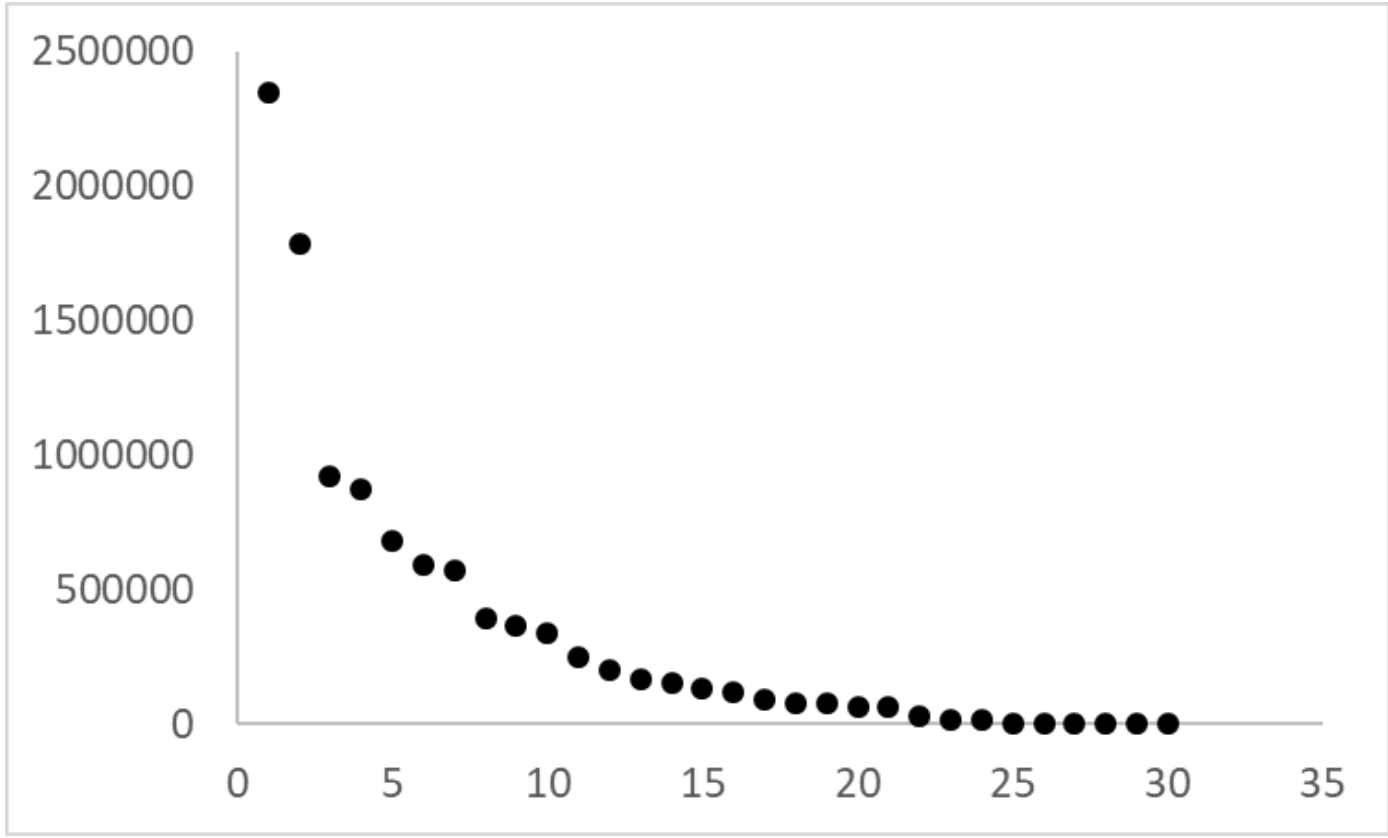
Distribuce jednotek

- Skácel: *Odvaha k tomu*: ranková frekvenční distribuce slov



Distribuce jednotek

- SYN2005: : ranková frekvenční distribuce primárních předložek



Distribuce – modely a interpretace

- diverzifikovanost sytému:

- type-token ratio
- repeat rate
- entropie

- výsledkem jedna hodnota

Příklad

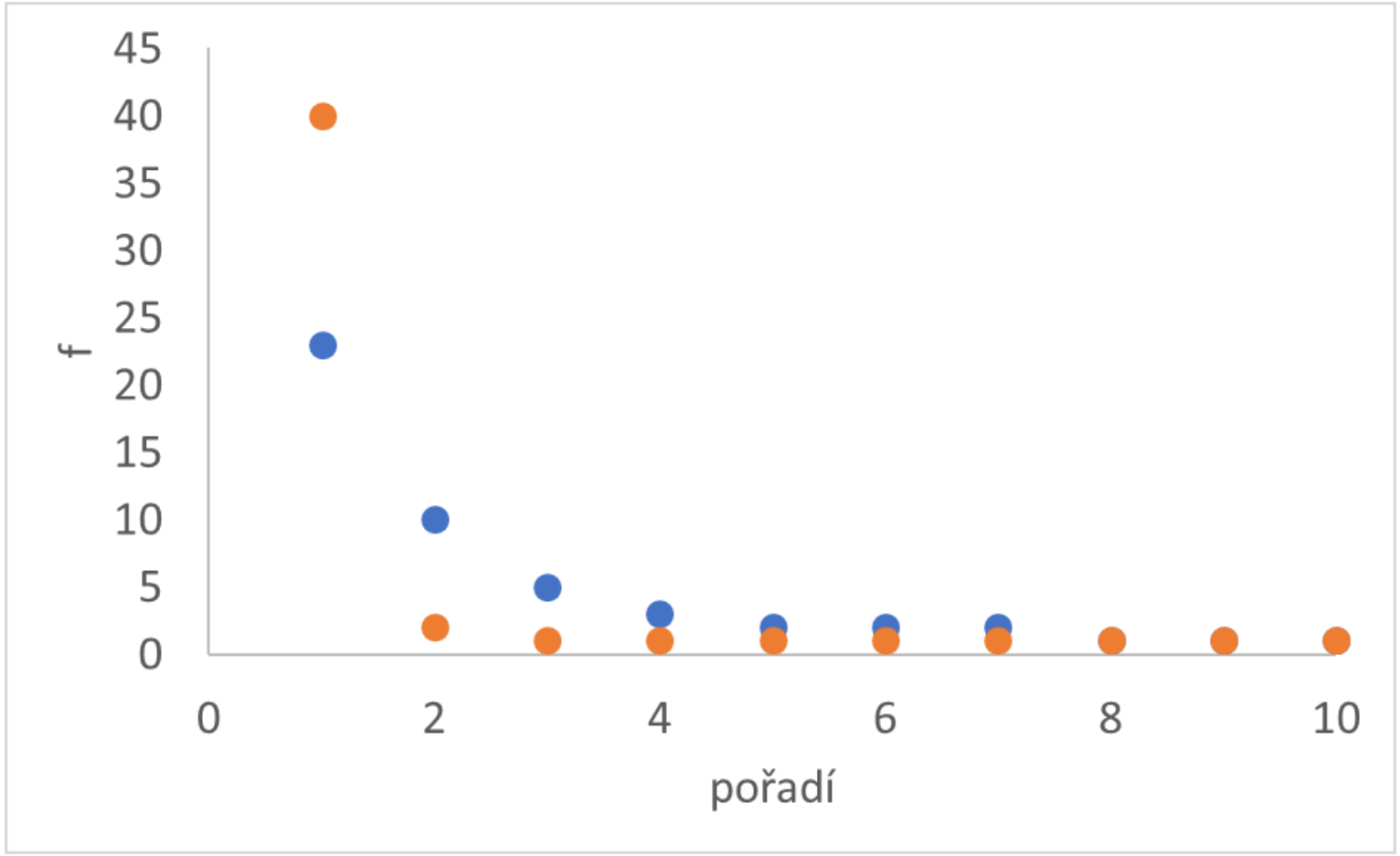
- V = 10 jednotek
- N = 50 výskytů

- dvě různé distribuce

Příklad

pořadí	f		pořadí	f
1	23		1	40
2	10		2	2
3	5		3	1
4	3		4	1
5	2		5	1
6	2		6	1
7	2		7	1
8	1		8	1
9	1		9	1
10	1		10	1

Příklad



Příklad – type-token poměr

- diverzifikovanost/slovní bohatství

$$TTR = \frac{V}{N}$$

- jaká bude teoreticky nejvyšší a nejnižší hodnota TTR u textu (souboru), který bude mít délku N = 50 slov?

Příklad – type-token poměr

- nejvyšší hodnota = každé slovo pouze jednou

$$TTR_{max} = \frac{V}{N}$$

Příklad – type-token poměr

- nejvyšší hodnota = každé slovo pouze jednou

$$TTR_{max} = \frac{V}{N} = \frac{50}{50} = 1$$

Příklad – type-token poměr

- nejvyšší hodnota = každé slovo pouze jednou

$$TTR_{max} = \frac{V}{N} = \frac{50}{50} = 1$$

- nejnižší hodnota = jedno slovo v celém texu

Příklad – type-token poměr

- nejvyšší hodnota = každé slovo pouze jednou

$$TTR_{max} = \frac{V}{N} = \frac{50}{50} = 1$$

- nejnižší hodnota = jedno slovo v celém texu

$$TTR_{min} = \frac{V}{N} = \frac{1}{50} = 0.02$$

Příklad – type-token poměr

- a co naše hypotetická data?
- liší se jejich TTR?

pořadí	f		pořadí	f
1	23		1	40
2	10		2	2
3	5		3	1
4	3		4	1
5	2		5	1
6	2		6	1
7	2		7	1
8	1		8	1
9	1		9	1
10	1		10	1

Příklad – type-token poměr

- a co naše hypotetická data?
- liší se jejich TTR?

$$TTR_{příklad} = \frac{V}{N} = \frac{10}{50} = 0.2$$

pořadí	f		pořadí	f
1	23		1	40
2	10		2	2
3	5		3	1
4	3		4	1
5	2		5	1
6	2		6	1
7	2		7	1
8	1		8	1
9	1		9	1
10	1		10	1

Příklad – index opakování (repeat rate)

- míra koncentrace jednotek (např. slov) v souboru

$$RR = \sum_{r=1}^{V} p_r^2$$

$$p_r = \frac{f_r}{N}$$

$$RR = \frac{1}{N^2} \sum_{r=1}^{V} f_r^2$$

Příklad – index opakování (repeat rate)

- nejvyšší koncentrace = jedno slovo v celém texu

$$RR_{max} = \frac{f_r^2}{N^2} = \frac{50^2}{50^2} = \frac{2500}{2500} = 1$$

- nejnižší koncentrace = každé slovo pouze jednou

$$RR_{min} = \frac{f_r^2}{N^2} = \frac{1^2 + 1^2 + 1^2 \ldots + 1^2}{50^2} = \frac{50}{2500} = 0.02$$

Příklad – index opakování (repeat rate)

- nejnižší koncentrace = zobecnění

$$RR_{min} = \frac{1}{N^2}\sum_{r=1}^{V}\left(\frac{N}{V}\right)^2 = \frac{1}{V}$$

- pro naše hypotetická data, V = 10 platí

$$RR_{min} = \frac{1}{V} = \frac{1}{10} = 0.1$$

Příklad – repeat rate

- a co naše hypotetická data?
- liší se jejich RR?

- Excel

pořadí	f		pořadí	f
1	23		1	40
2	10		2	2
3	5		3	1
4	3		4	1
5	2		5	1
6	2		6	1
7	2		7	1
8	1		8	1
9	1		9	1
10	1		10	1

Příklad – repeat rate

- Excel

pořadí	f	f^2
1	23	529
2	10	100
3	5	25
4	3	9
5	2	4
6	2	4
7	2	4
8	1	1
9	1	1
10	1	1
	Σf^2	678
	N^2	2500
	RR	0.2712

Příklad – repeat rate

- vypočítejte RR pro oranžová data

Příklad – repeat rate

pořadí	f	f^2
1	23	529
2	10	100
3	5	25
4	3	9
5	2	4
6	2	4
7	2	4
8	1	1
9	1	1
10	1	1
	Σf^2	678
	N^2	2500
	RR	0.2712

pořadí	f	f^2
1	40	1600
2	2	4
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
8	1	1
9	1	1
10	1	1
	Σf^2	1612
	N^2	2500
	RR	0.6448

Entropie

- míra neurčitosti systému
- míra diverzity
  - čím je hodnota entropie větší, tím systém diverzifikovanější (tj. méně koncentrovaný)
  - např. vysoká hodnota entropie je např. znakem velkého slovního bohatství

Entropie

$$H = -\sum_{r=1}^{V} p_r \, log_2 p_r$$

$$p_r = \frac{f_r}{N}$$

$$H = log_2 N - \frac{1}{N} \sum_{r=1}^{V} f_r \, log_2 f_r$$

Příklad – entropie

- nejnižší entropie = největší koncentrace slovníku (celý text z 1 slova)

$$H_{min} = log_2 50 - \frac{50(log_2 50)}{50} = log_2 50 - log_2 50 = 0$$

Příklad – entropie

- nejvyšší entropie = nejnižší koncentrace slovníku (každé slovo 1x)

$$H_{max} = log_2 50 - \frac{50 \sum_{r=1}^{V} log_2 1}{50} = log_2 50 - \frac{0}{50} = log_2 50 = 5.64$$

$$H_{max} = log_2 V$$

- nejvyšší entropie pro naše hypotetická data V = 10

$$H_{max} = log_2 10 = 3.32$$

Příklad – entropie

- pro porovnání dat různého rozsahu → relativní entropie
  - <0; 1>

$$H_{rel} = \frac{H}{H_{max}} = \frac{H}{log_2 V}$$

Příklad – entropie

- a co naše hypotetická data?
- liší se jejich H?

- Excel

pořadí	f		pořadí	f
1	23		1	40
2	10		2	2
3	5		3	1
4	3		4	1
5	2		5	1
6	2		6	1
7	2		7	1
8	1		8	1
9	1		9	1
10	1		10	1

Příklad – entropie

pořadí	f	f*log_2*f
1	23	104.042
2	10	33.219
3	5	11.610
4	3	4.755
5	2	2
6	2	2
7	2	2
8	1	0
9	1	0
10	1	0
	Σf*log_2*f	159.626
	Σf*log_2*f/N	3.193
	log_2N	5.644
	H	2.451
	H_max	3.322
	H_rel	0.738

Příklad – repeat rate

- vypočítejte H pro oranžová data

Příklad – entropie

pořadí	f	f*log_2*f
1	23	104.042
2	10	33.219
3	5	11.610
4	3	4.755
5	2	2
6	2	2
7	2	2
8	1	0
9	1	0
10	1	0
	Σf*log_2*f	159.626
	Σf*log_2*f/N	3.193
	log_2N	5.644
	H	2.451
	H_max	3.322
	H_rel	0.738

pořadí	f	f*log_2*f
1	40	212.877
2	2	2
3	1	0
4	1	0
5	1	0
6	1	0
7	1	0
8	1	0
9	1	0
10	1	0
	Σf*log_2*	214.877
	Σf*log_2*	4.298
	log_2N	5.644
	H	1.346
	H_max	3.322
	H_rel	0.405

TTR, RR, H

- ! nevhodné pro porovnávání souborů nestejné délky (např. textů)

- viz QuitaUp (https://korpus.cz/quitaup/)

- příklad vhodného použití → distribuce pádů

Příklad - distribuce pádů vs. sémantika

- Proč (a jaký) by měla mít sémantika vliv na distribuci pádů substantiv?
- východiska (předběžná)
  - substantiva denotující osoby mají tendenci se vyskytovat nejčastěji v nominativu (vlivem tendence vyskytovat se v sémantické roli agentu)
  - u substantiv denotujících např. neživé předměty nebo abstraktní entity není jejich morfosyntaktický status jednoznačný

Subst. maskulina anim. vs. inanim

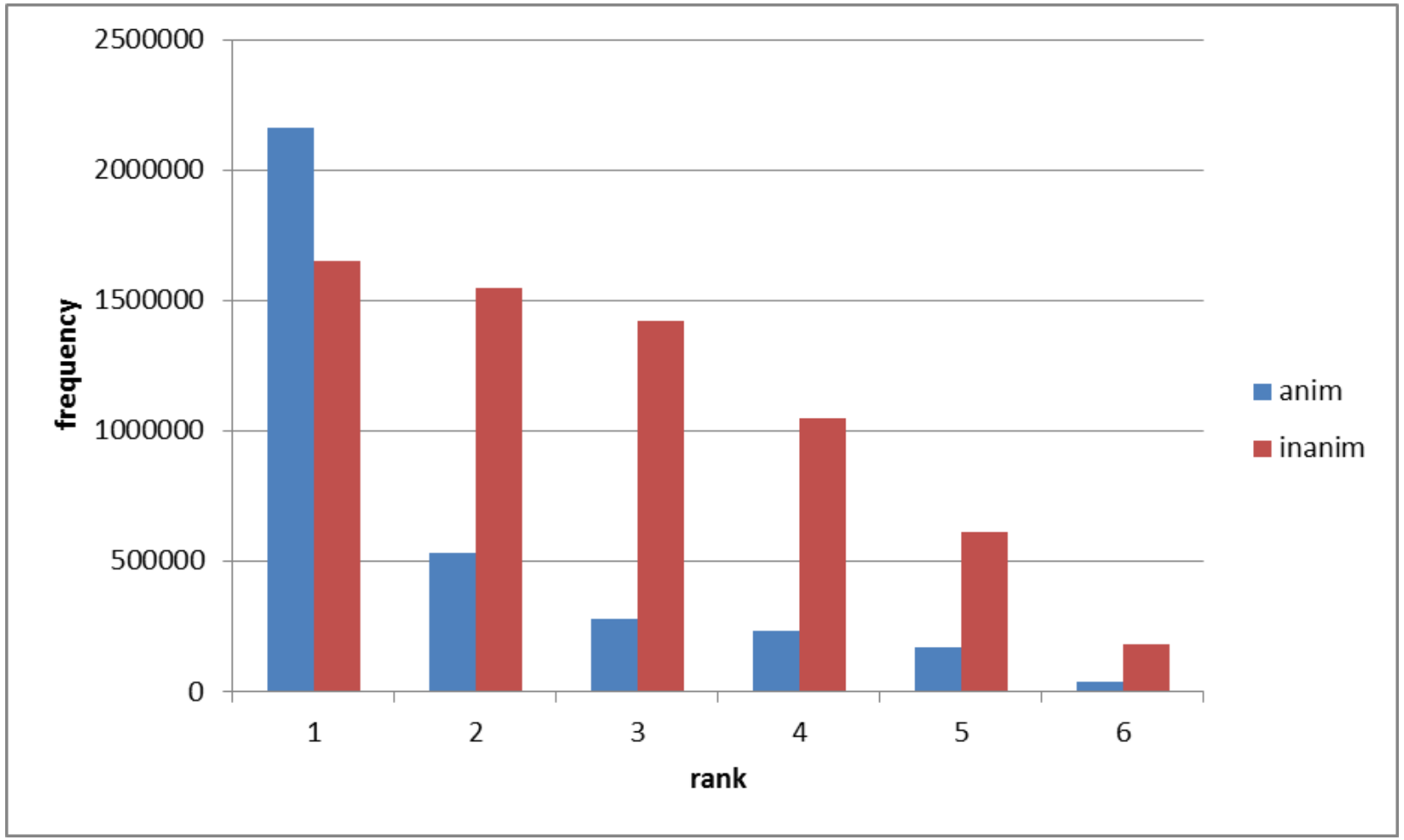
Anim. sg. (SYN2010)

pád	frekvence
nom.	2161013
gen.	532579
acc.	278806
instr.	233327
dat.	170042
loc.	39956

Inanim. sg. (SYN2010)

pád	frekvence
gen.	1649641
acc.	1546412
nom.	1422769
loc.	1045981
instr.	613918
dat.	184674

Subst. maskulina anim. vs. inanim



Subst. maskulina anim. vs. inanim

- $RR_{anim} = 0.439$ $RR_{inanim} = 0.207$

Distribuce pádů vs. sémantika

- Jaké obecné principy řídící jazykové chování by mohly mít vliv na předpokládaný vztah mezi sémantikou a distribucí pádů?
- Jaké jsou tzv. **hraniční podmínky**?
  - rod
  - číslo
  - polysémie atd.

  - pozn. více o hraničních podmínkách u tématu *Hypotéza a její vlastnosti*

Teoretické předpoklady

- distribuce pádů je výsledkem tzv. diverzifikačního procesu
- diverzifikace (obecně)
  - jednotka (např. slovo) – kategorie (pád, rod, číslo atd.) – jednotlivé instance (nom., gen…; mask., fem., neut.…)
  - pokud jednotka v rámci kategorie podléhá diverzifikaci, frekvence **nejsou** distribuovány rovnoměrně
  - jedná se o obecný jev, který je charakteristický pro jazykový systém

Hypotéza

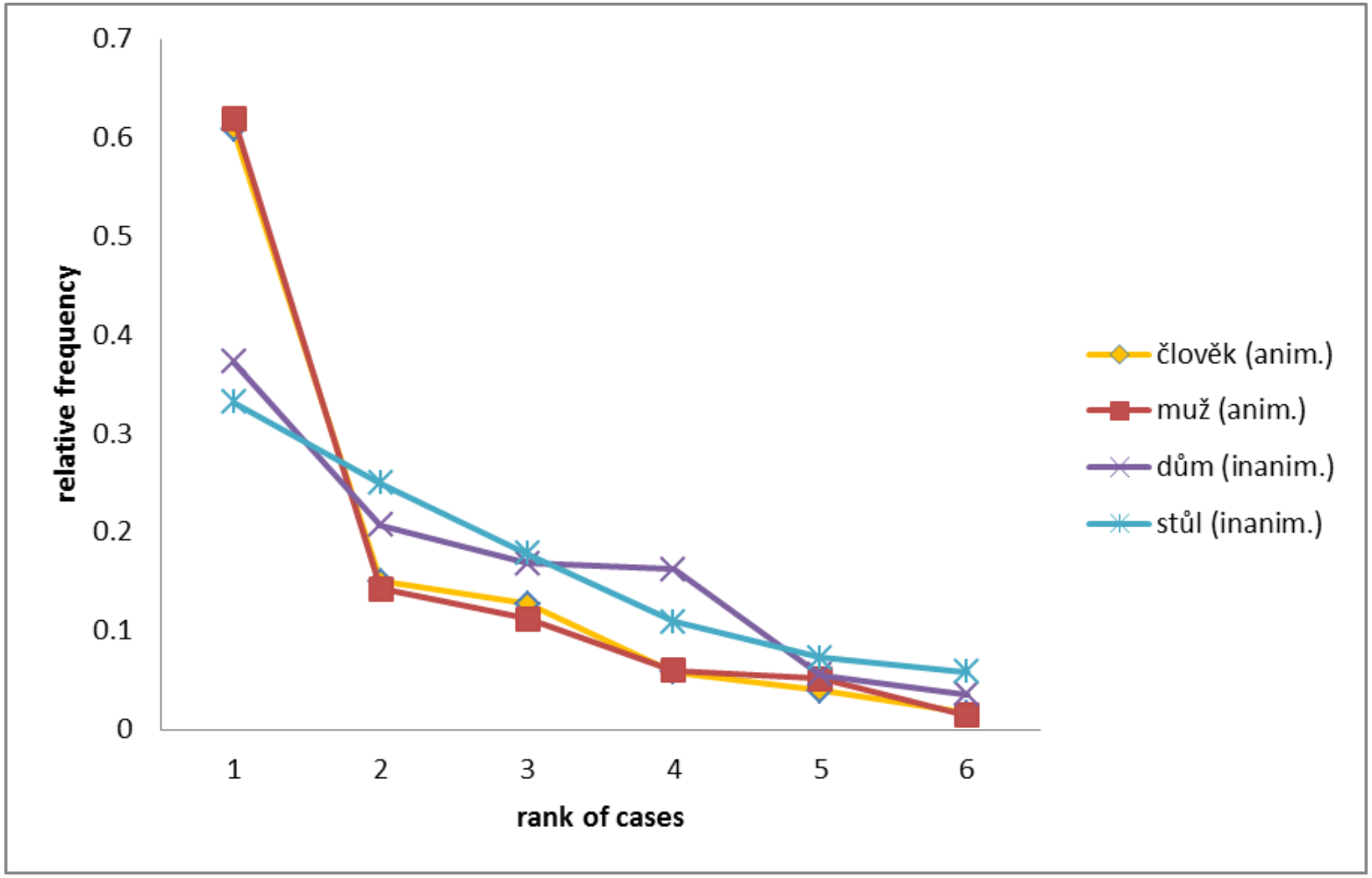
- pádové distribuce jednotlivých substantiv (anim. a inanim u všech tří rodů) se významně liší vzhledem k životnosti

- H0: mezi distribucí pádů životných a neživotných substantiv není rozdíl
- H1: mezi distribucí pádů životných a neživotných substantiv je rozdíl

Data

- SYN 2010
- 5 nejfrekventovanějších anim. a inanim. substantiv
- 10 v rámci každého rodu (mask., fem., neut.)
- celkem analyzováno 30 substantiv
  - konkrétní substantiva
  - bez vlastních jmen
  - pouze singulár

Distribuce pádů vs. sémantika



Distribuce pádů vs. sémantika

anim.

- člověk
  - RR = 0.414
  - H_{rel} = 0.676
- muž
  - RR = 0.423
  - H_{rel} = 0.67

inanim.

- dům
  - RR = 0.24
  - H_{rel} = 0.874
- stůl
  - RR = 0.225
  - H_{rel} = 0.903

Distribuce pádů vs. sémantika

anim.

- člověk
  - RR = 0.414
  - H_{rel} = 0.676
- muž
  - RR = 0.423
  - H_{rel} = 0.67

inanim.

- dům
  - RR = 0.24
  - H_{rel} = 0.874
- stůl
  - RR = 0.225
  - H_{rel} = 0.903

pozn. z daných hodnot je možné udělat aritmetický průměr a statisticky testovat rozdíly mezi skupinami slov → více později

Distribuce – modely a interpretace

- matematická funkce jako model

- distribuční funkce
  - spojité veličiny
  - diskrétní veličiny

Model – funkce

- lineární

$$y = x$$

$$y = ax$$

$$y = b + ax$$

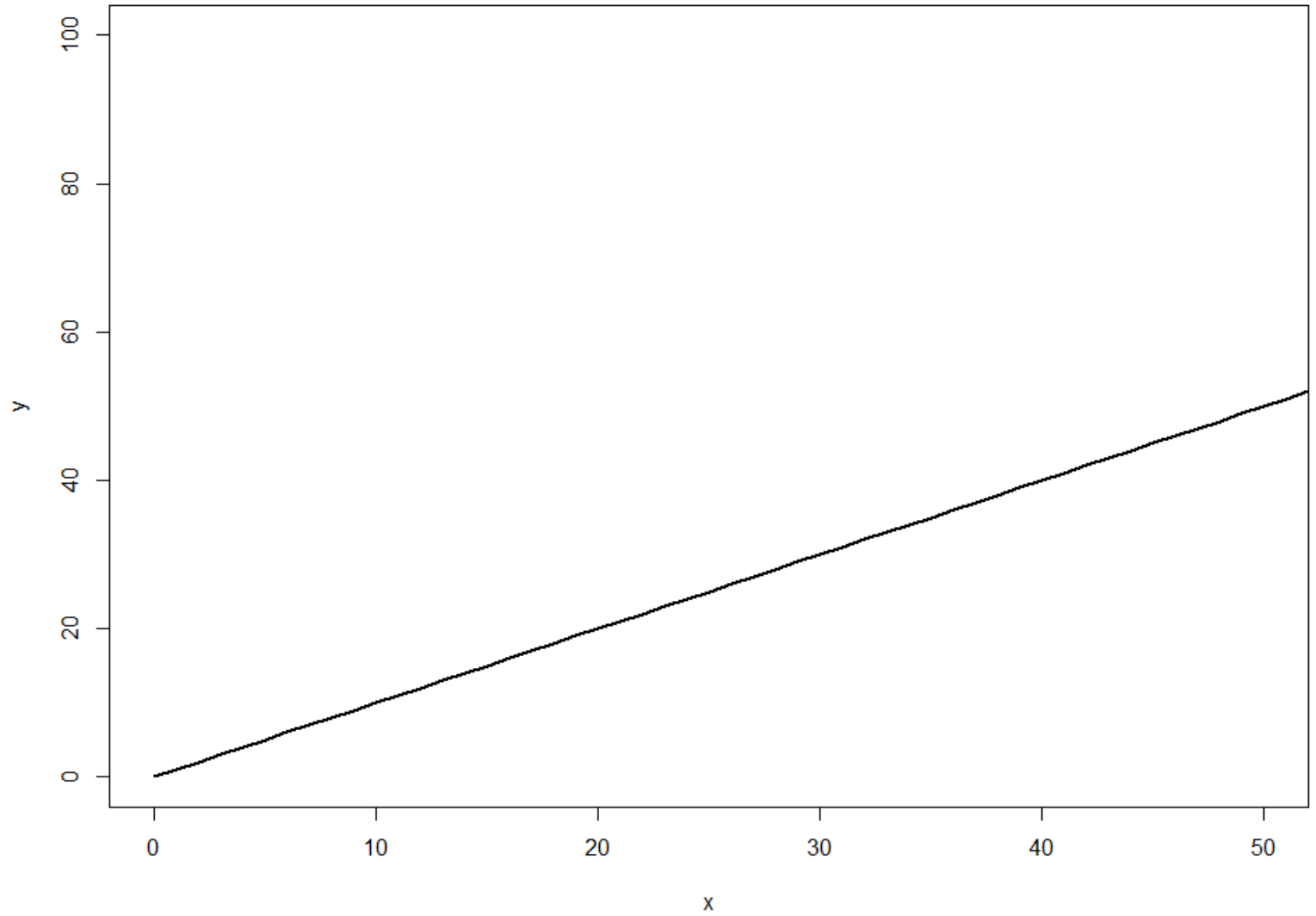
x, y … proměnné

a, b … parametry

Model – funkce

$y = ax$

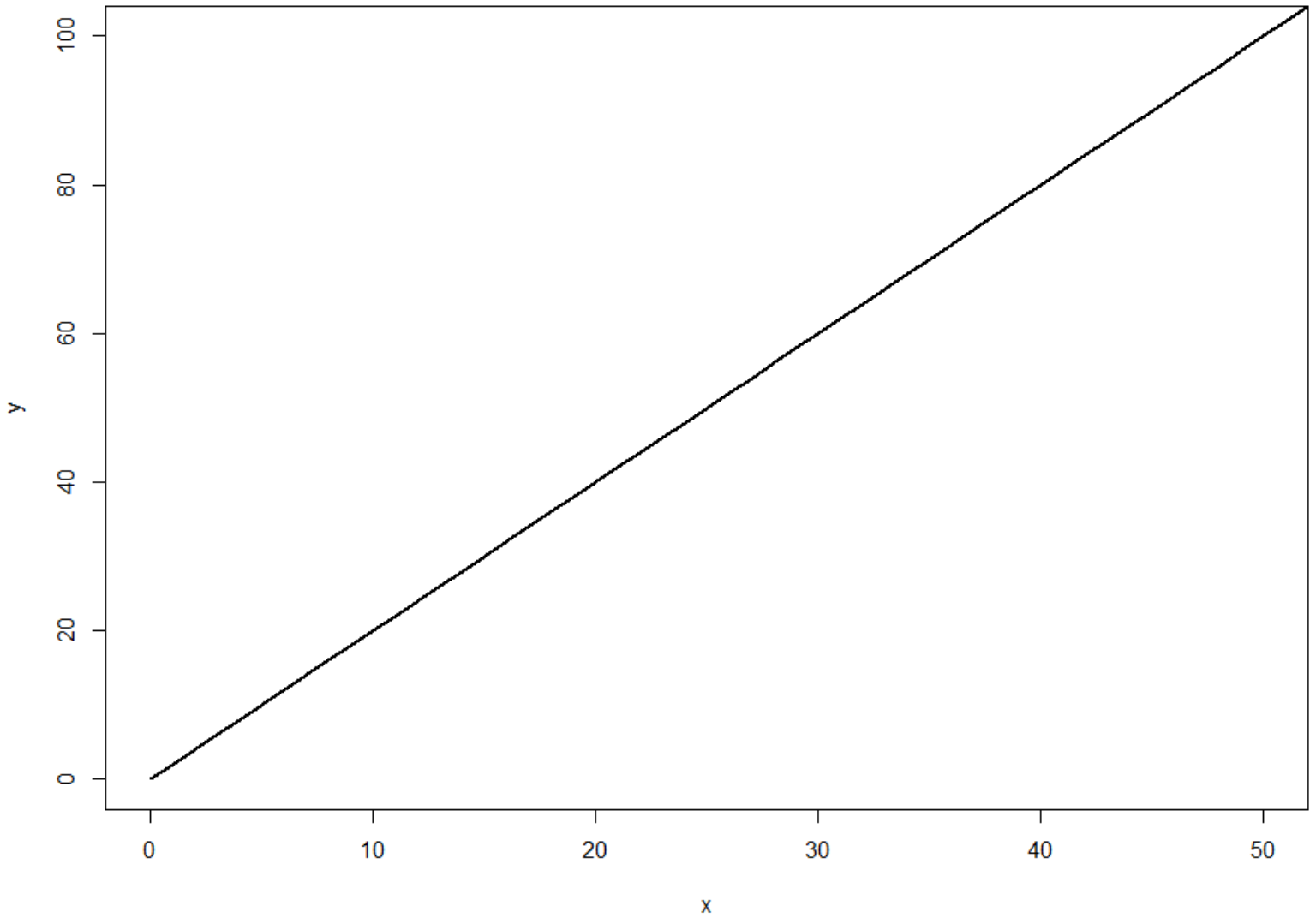
$a = 1$



Model – funkce

$y = ax$

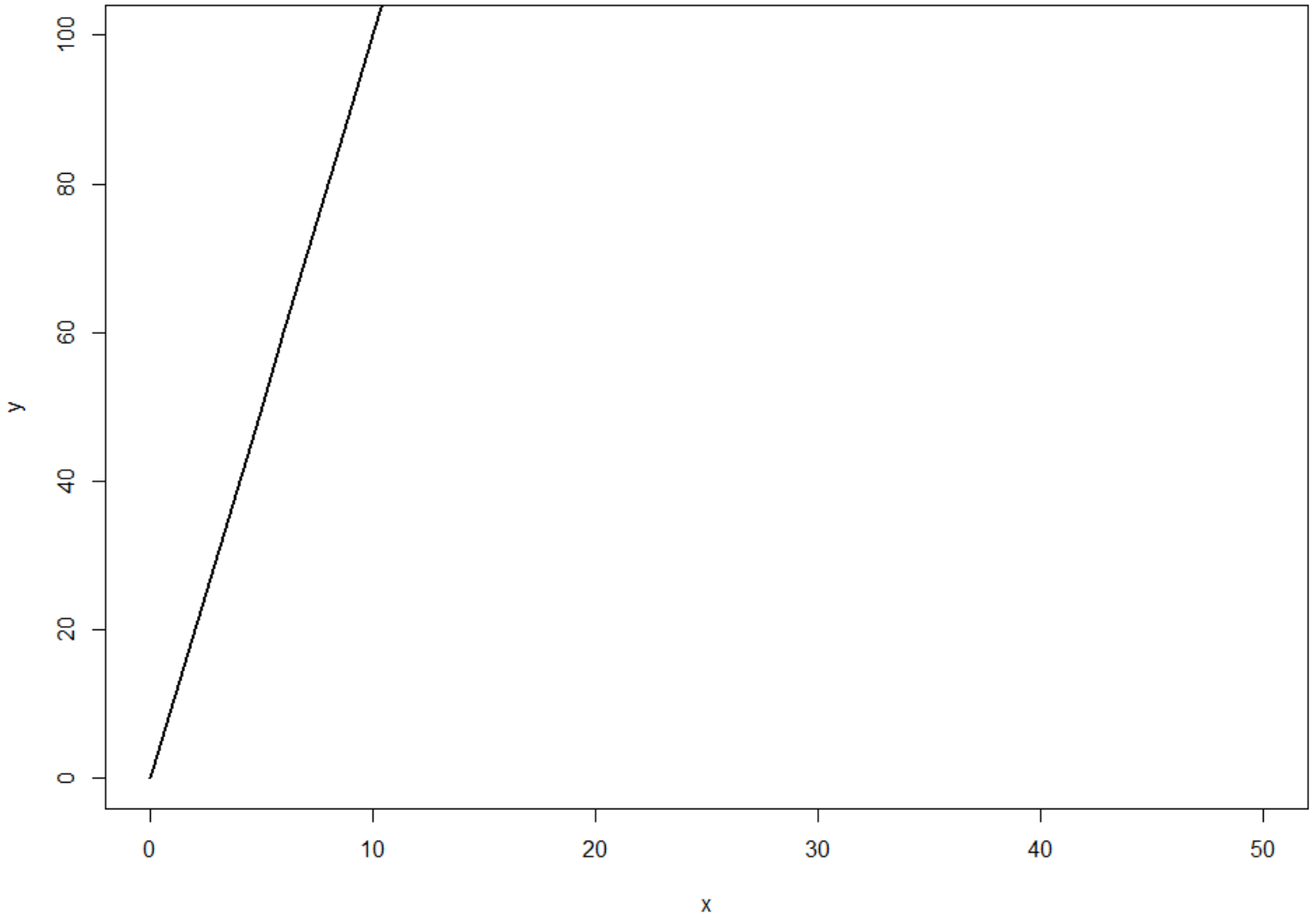
$a = 2$



Model – funkce

$y = ax$

a = 10

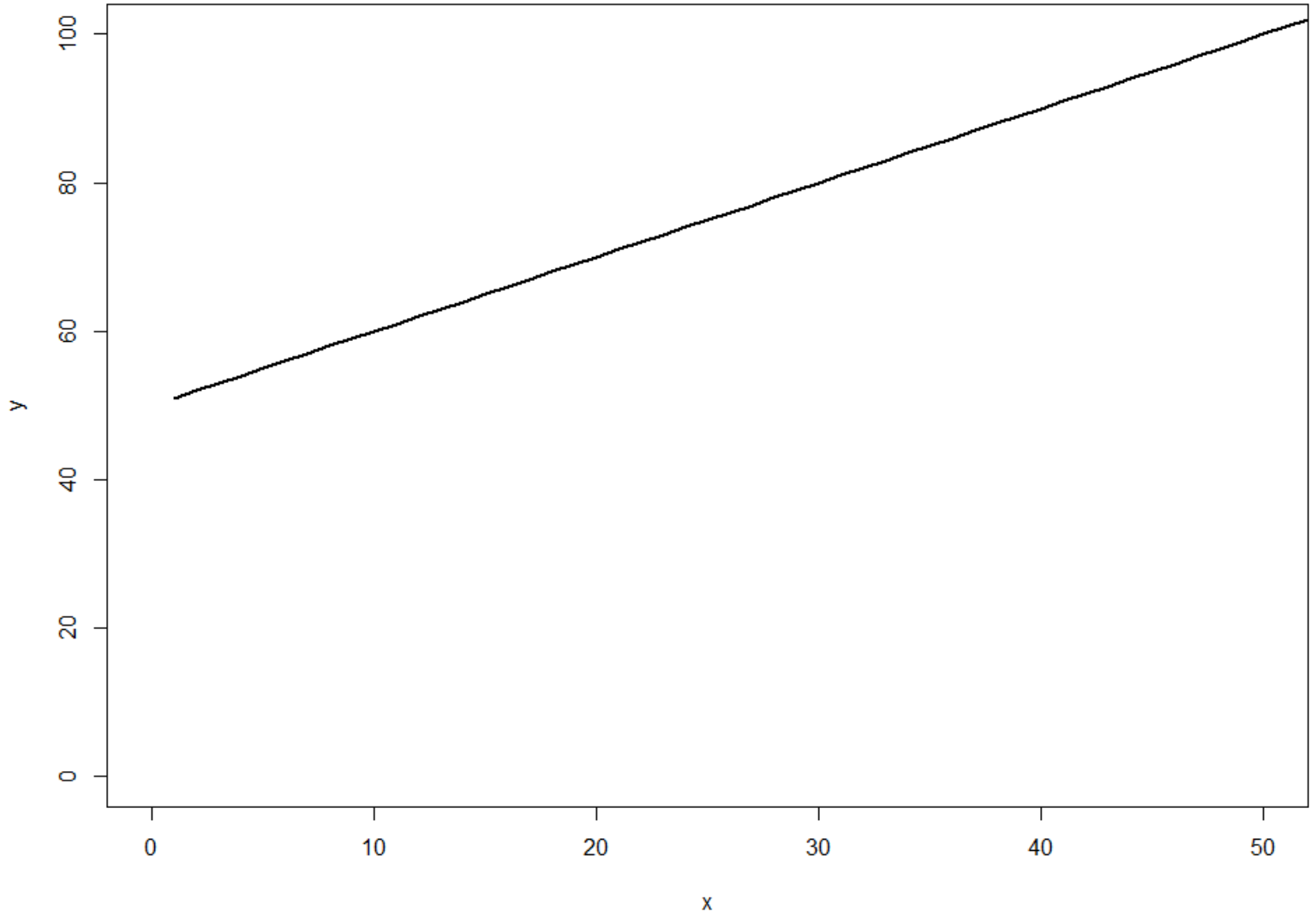


Model – funkce

$y = b + ax$

$a = 1$

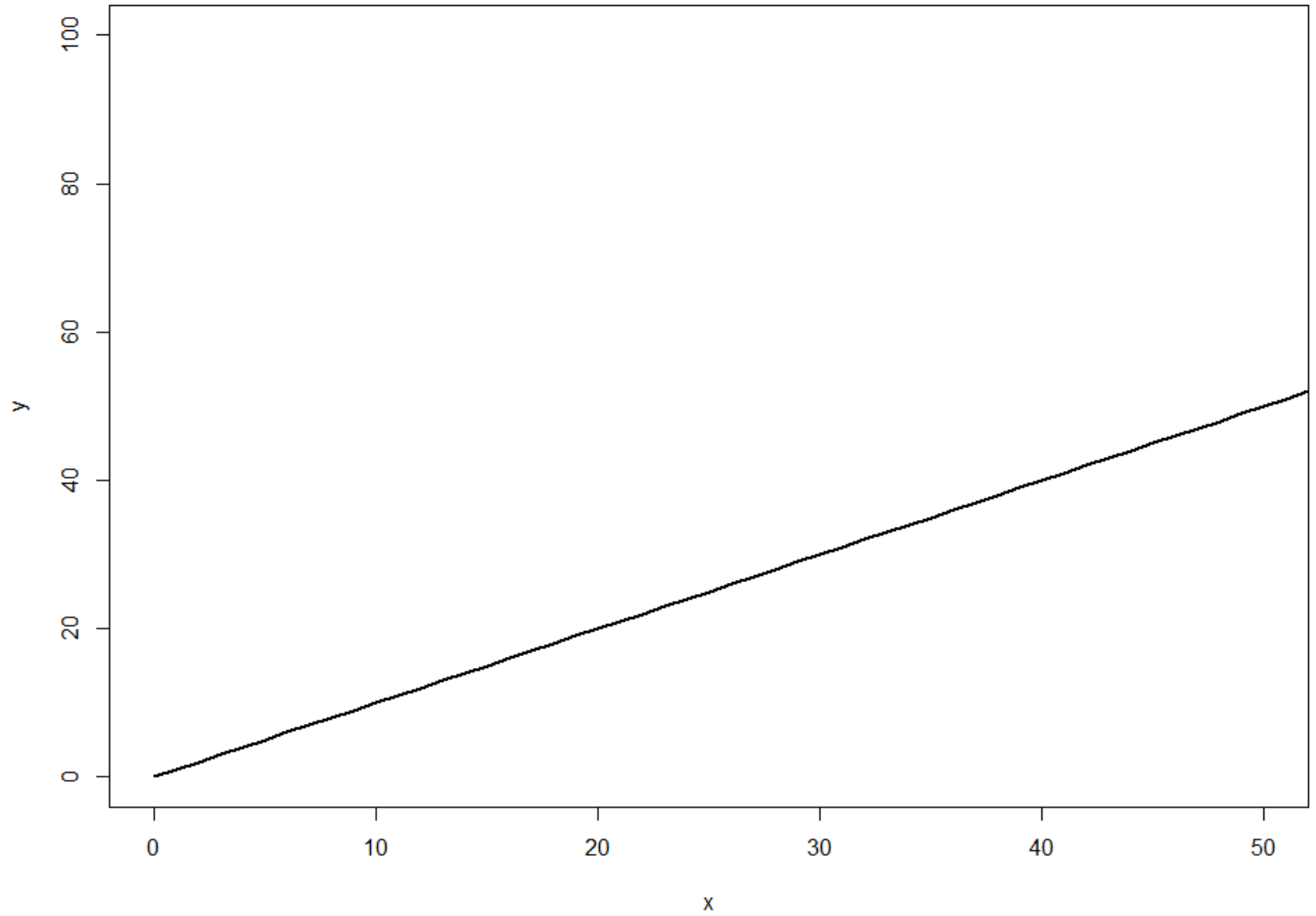
$b = 50$



Model – funkce

$y = ax$

$a = 1$

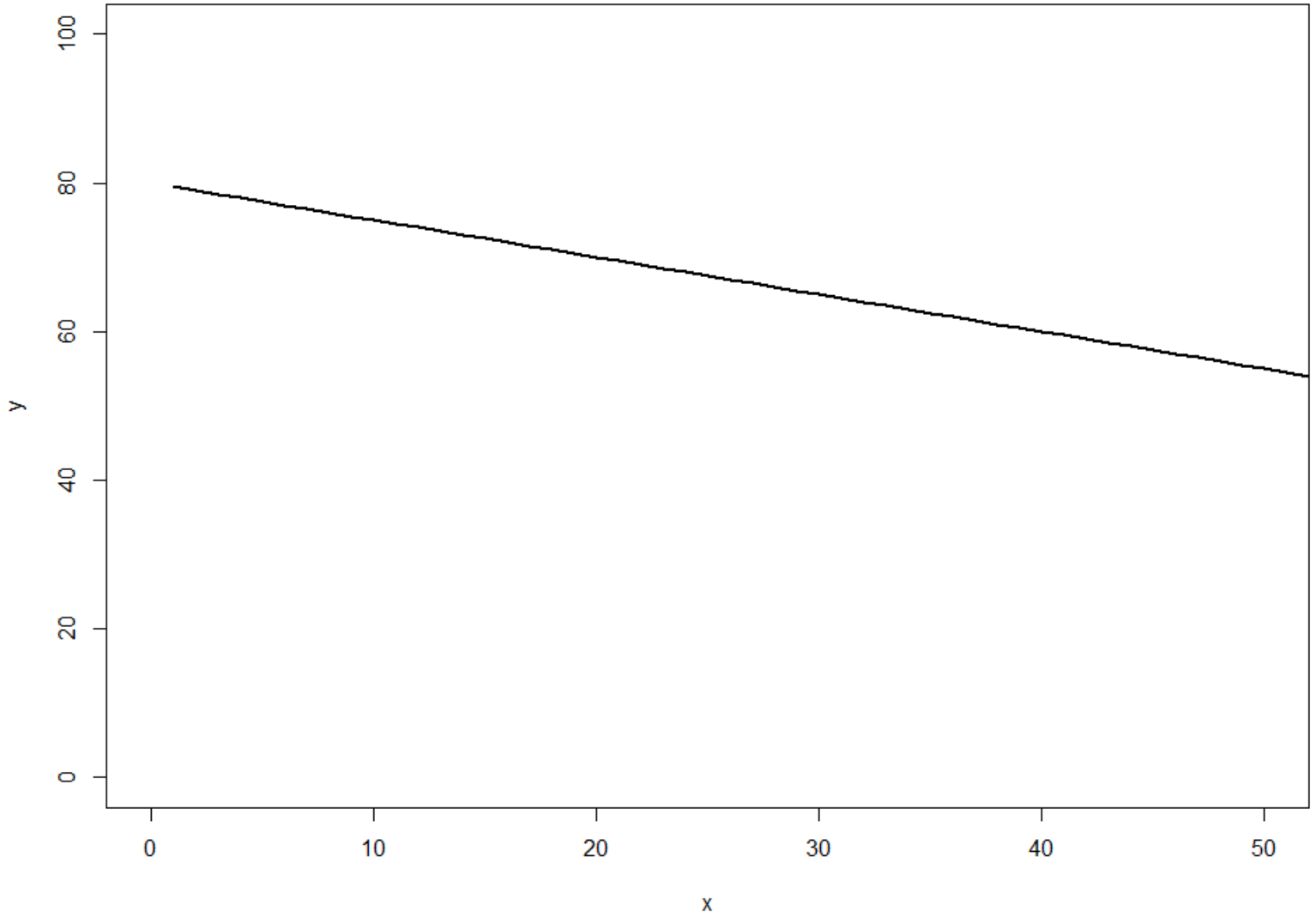


Model – funkce

$y = b + ax$

a = -0.5

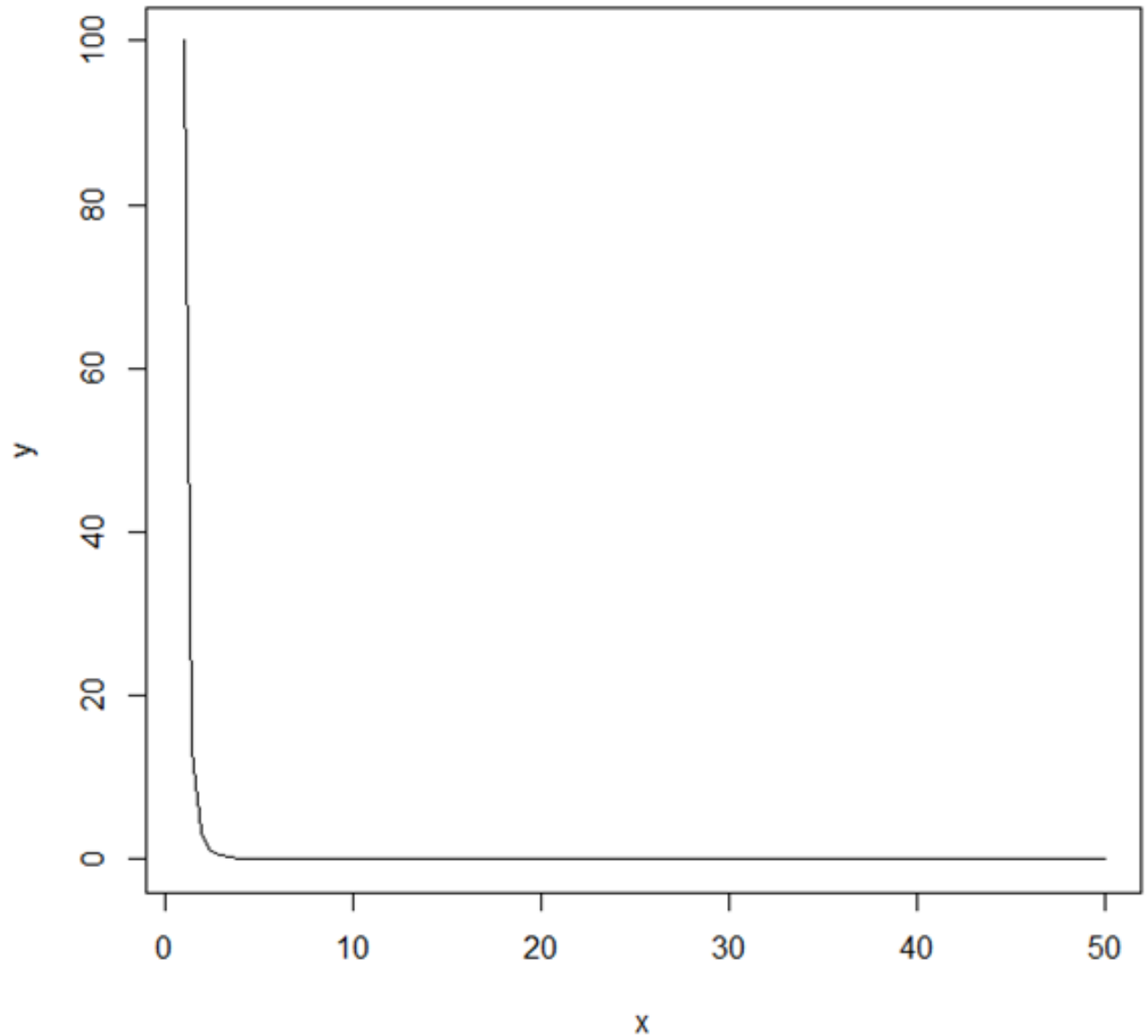
b = 80



Model – mocninná funkce

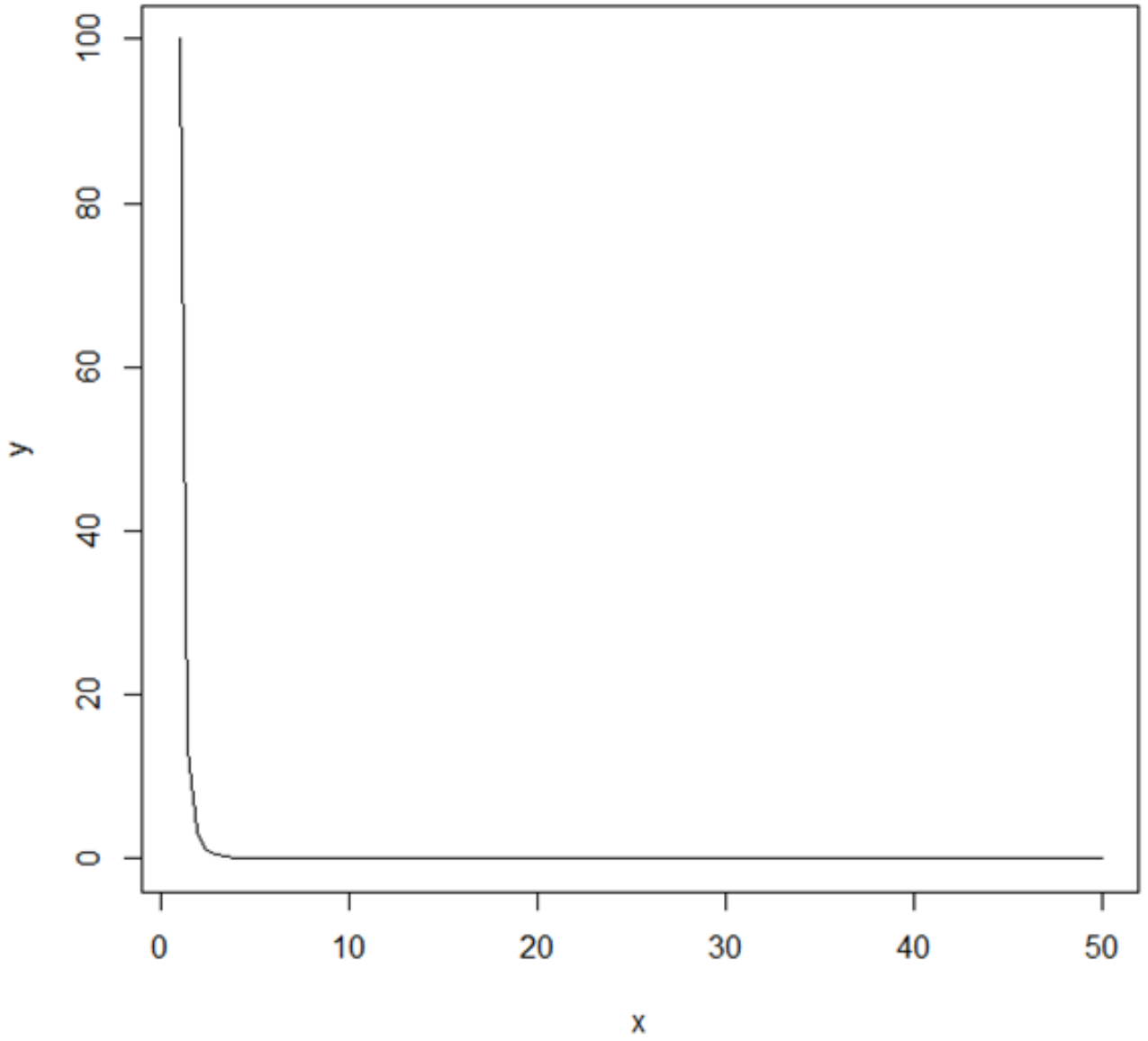
$$y = ax^{-b}$$

$$y = 100x^{-5}$$

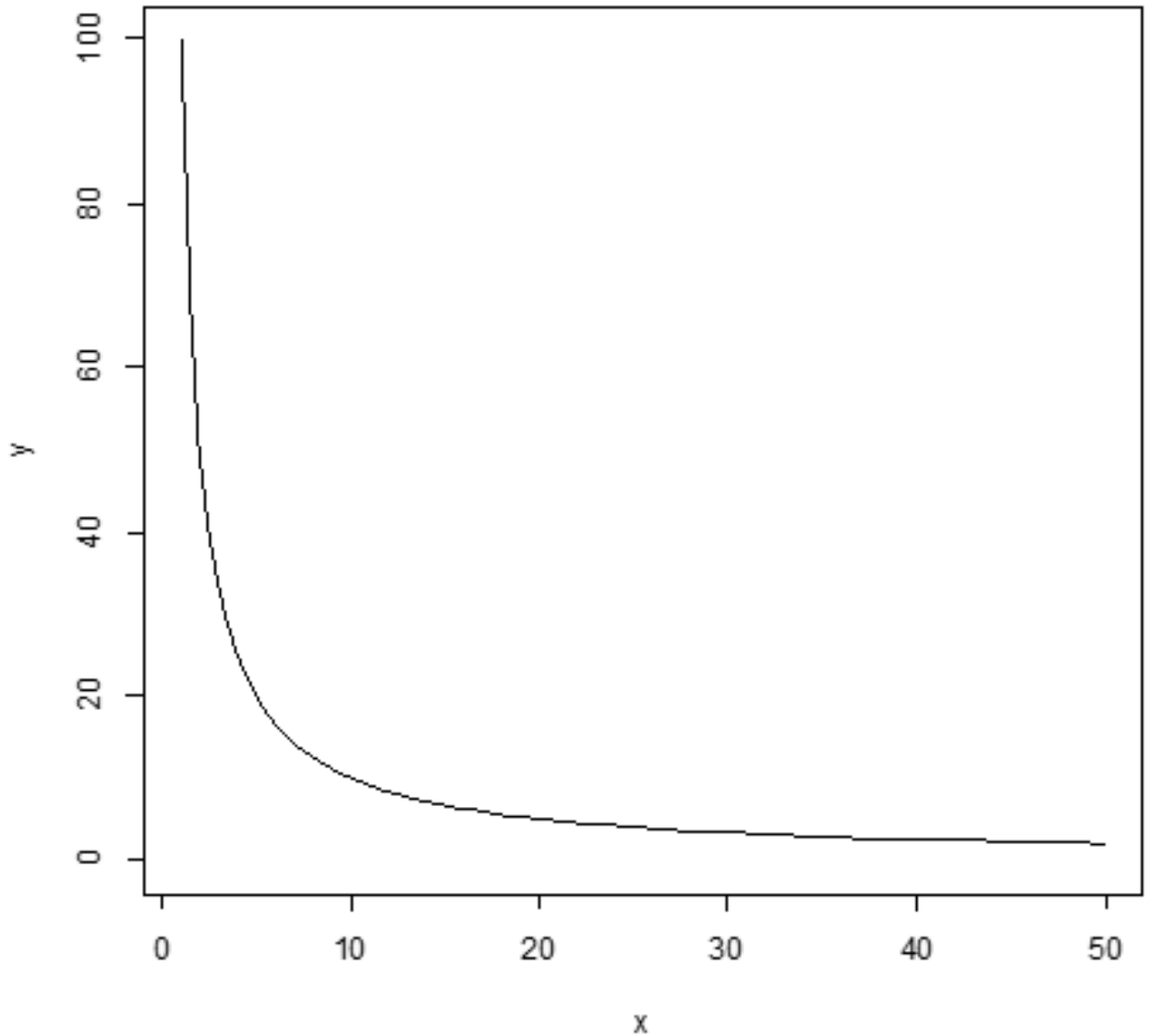


$$y = 100x^{-5}$$

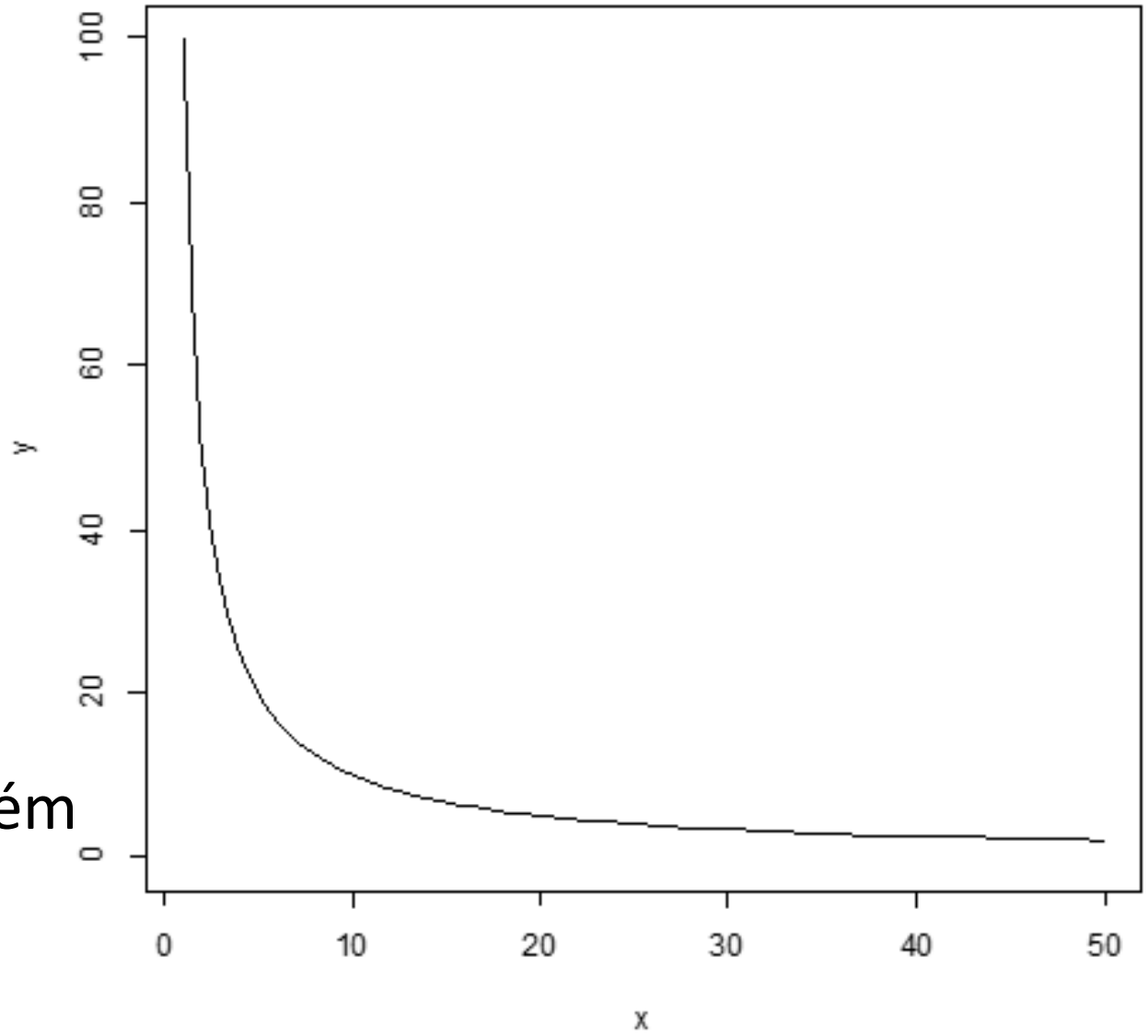
- diverzifikovaný systém



$$y = 100x^{-1}$$

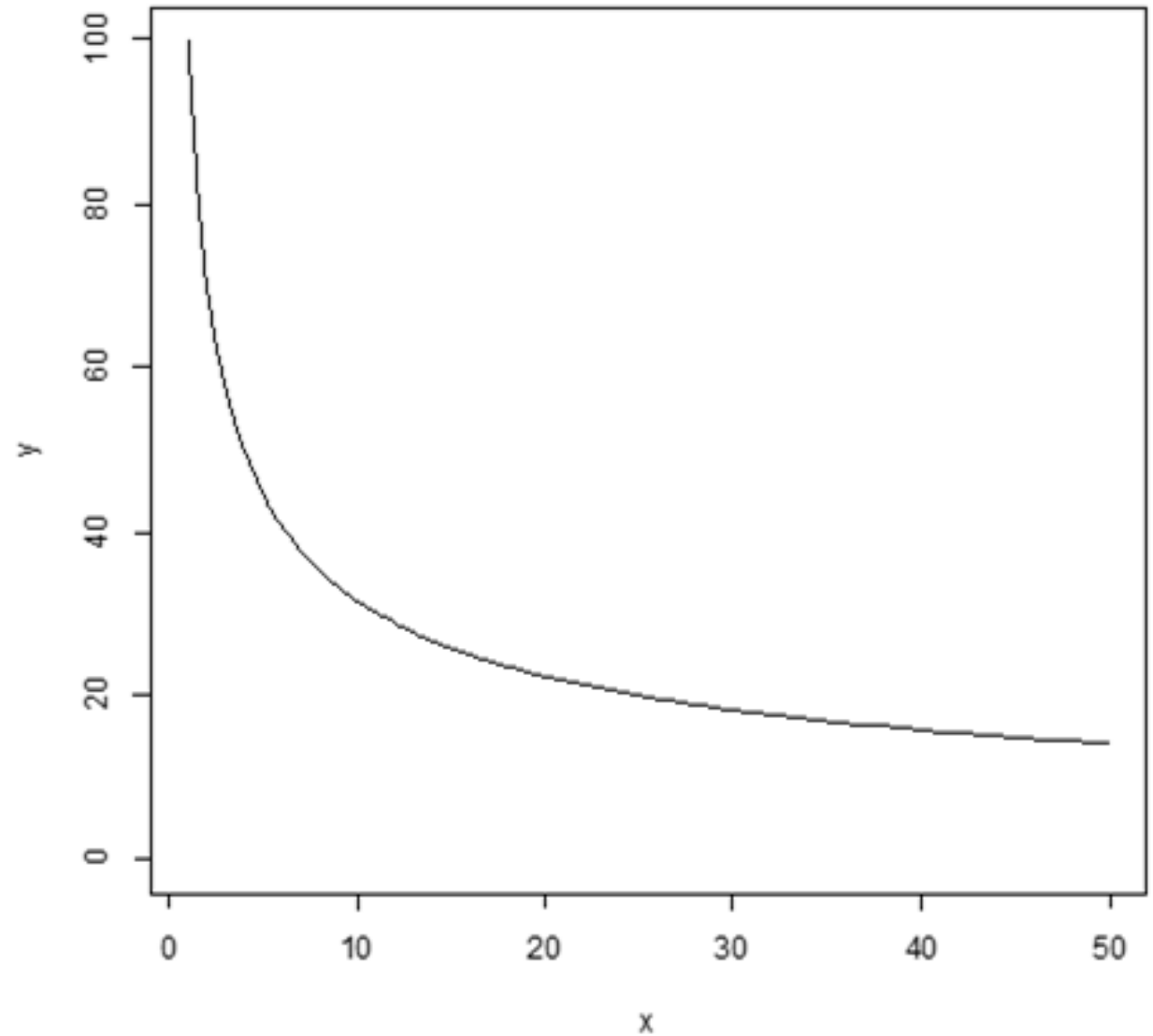


$$y = 100x^{-1}$$

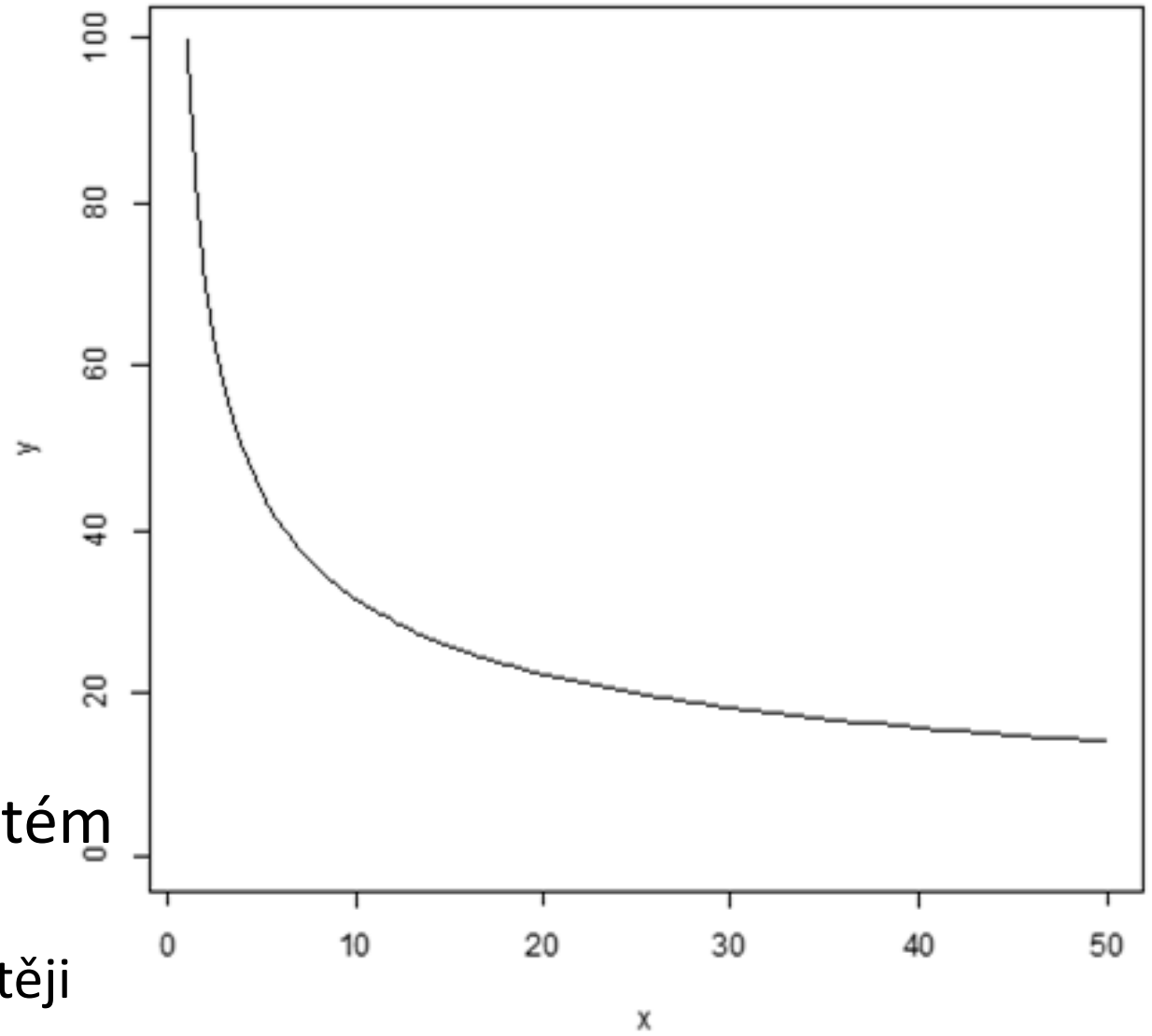


- méně diverzifikovaný systém
  - jednotky se častěji opakují

$$y = 100x^{-0.5}$$



$$y = 100x^{-0.5}$$



- nejméně diverzifikovaný systém (z prezentovaných příkladů)
  - jednotky se opakují ještě častěji

Model vs. realita

- model
  - předpokládá působení mechanismu
  - ideální stav

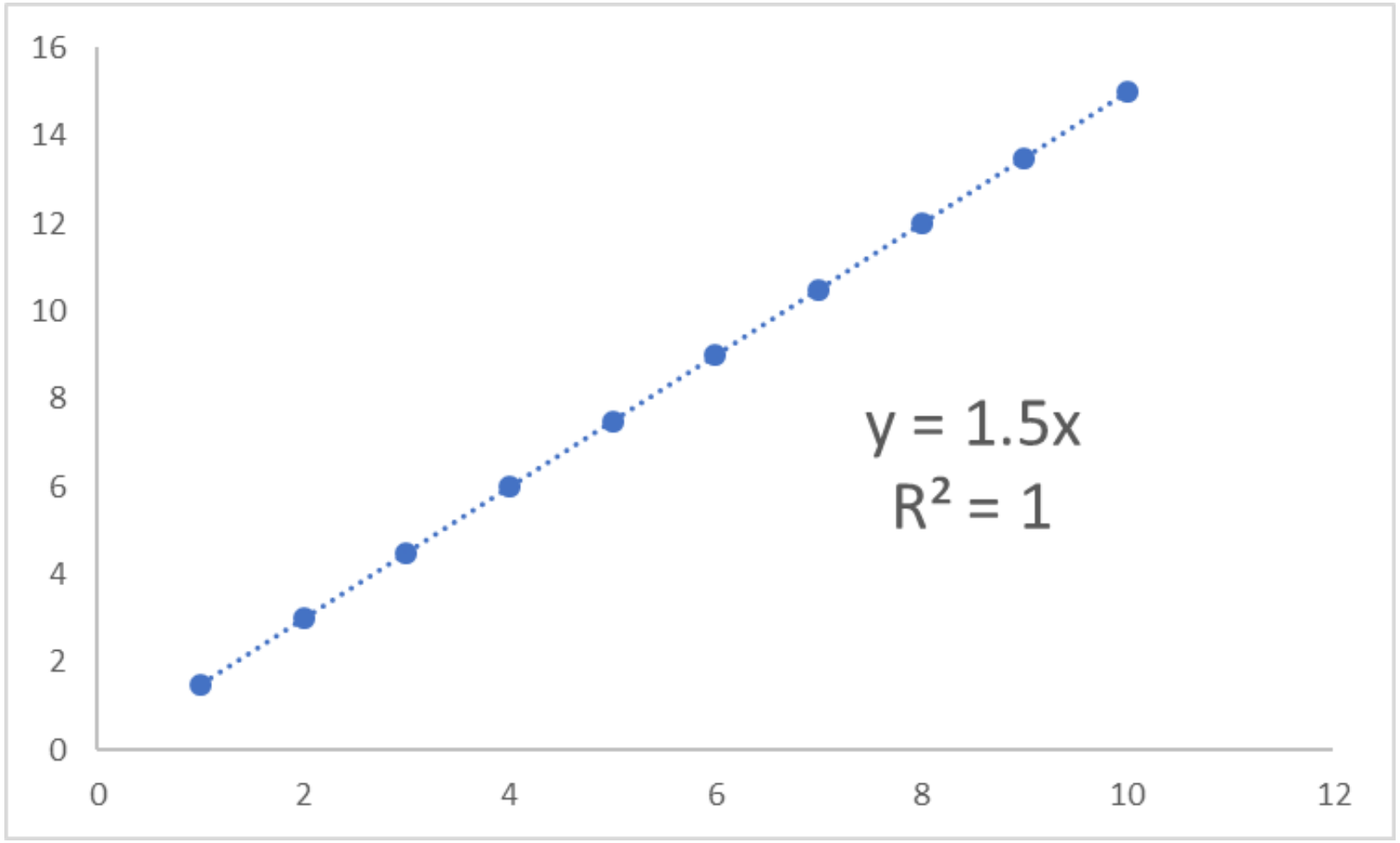
- realita
  - mechanismus ovlivněn různými faktory
  - fluktuace
  - náhodné jevy

Model vs. realita

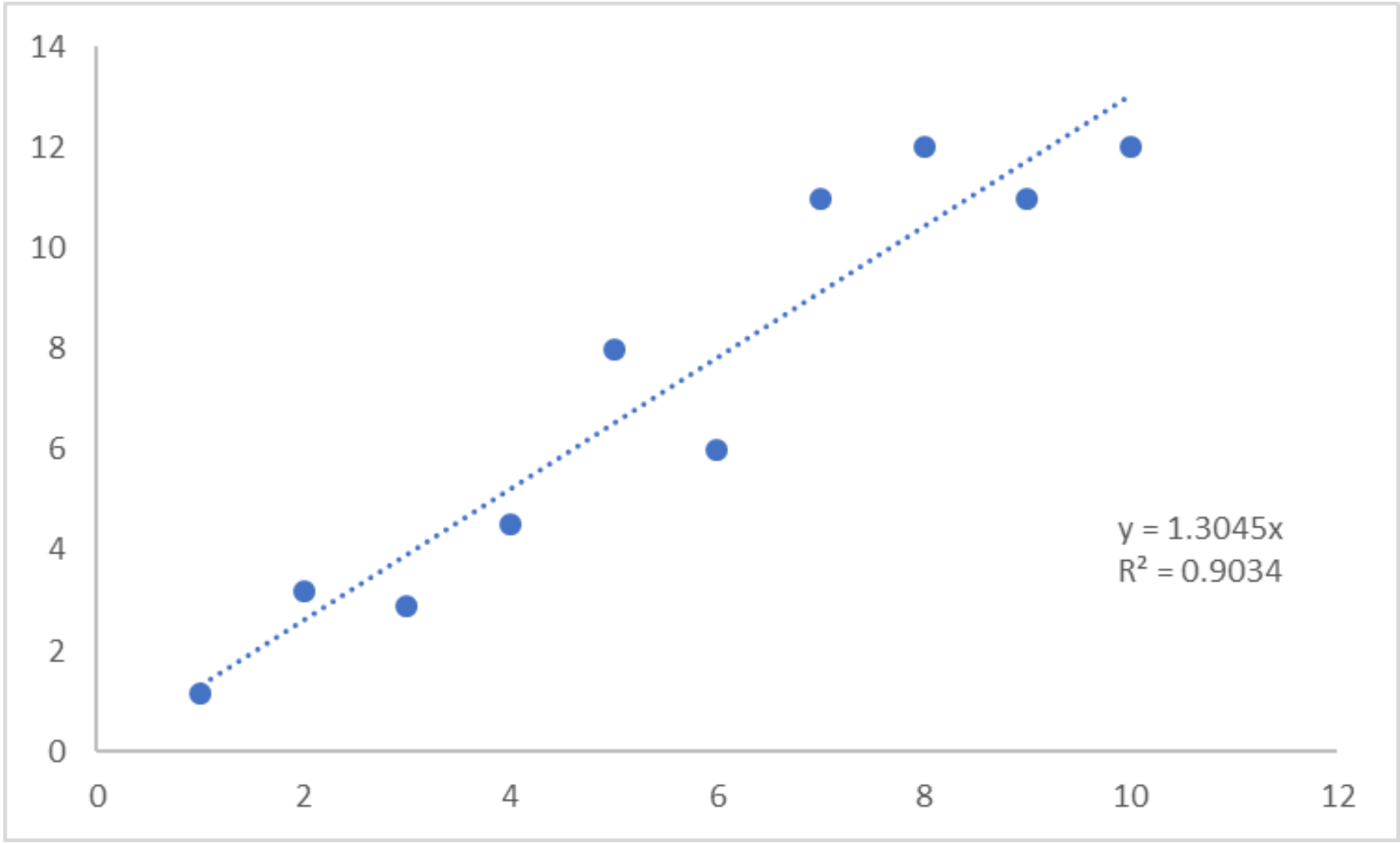
- postup

- model predikuje chování systému
- porovnáváme model s daty
- je možné vyjádřit míru modelu s daty

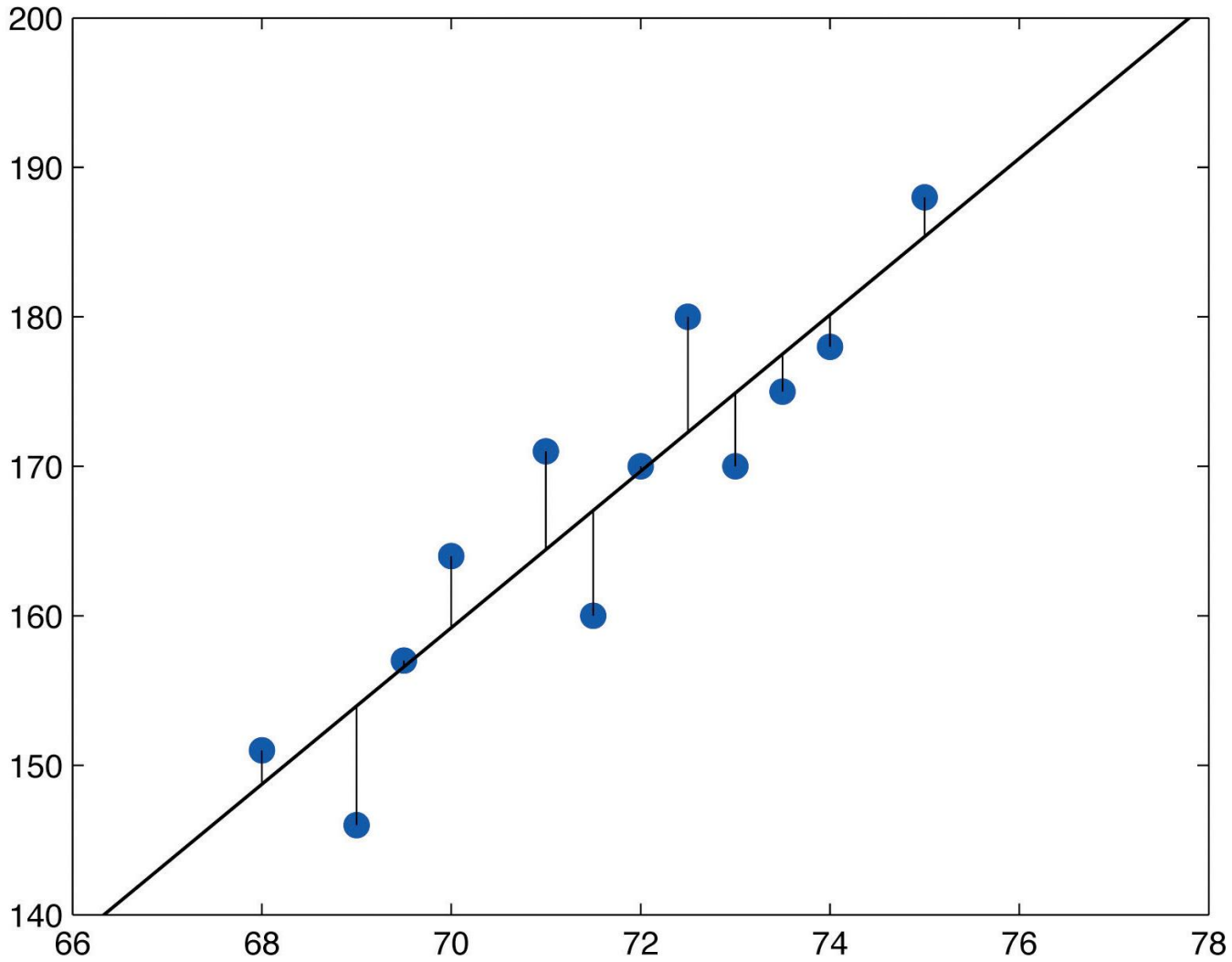
Model vs. realita



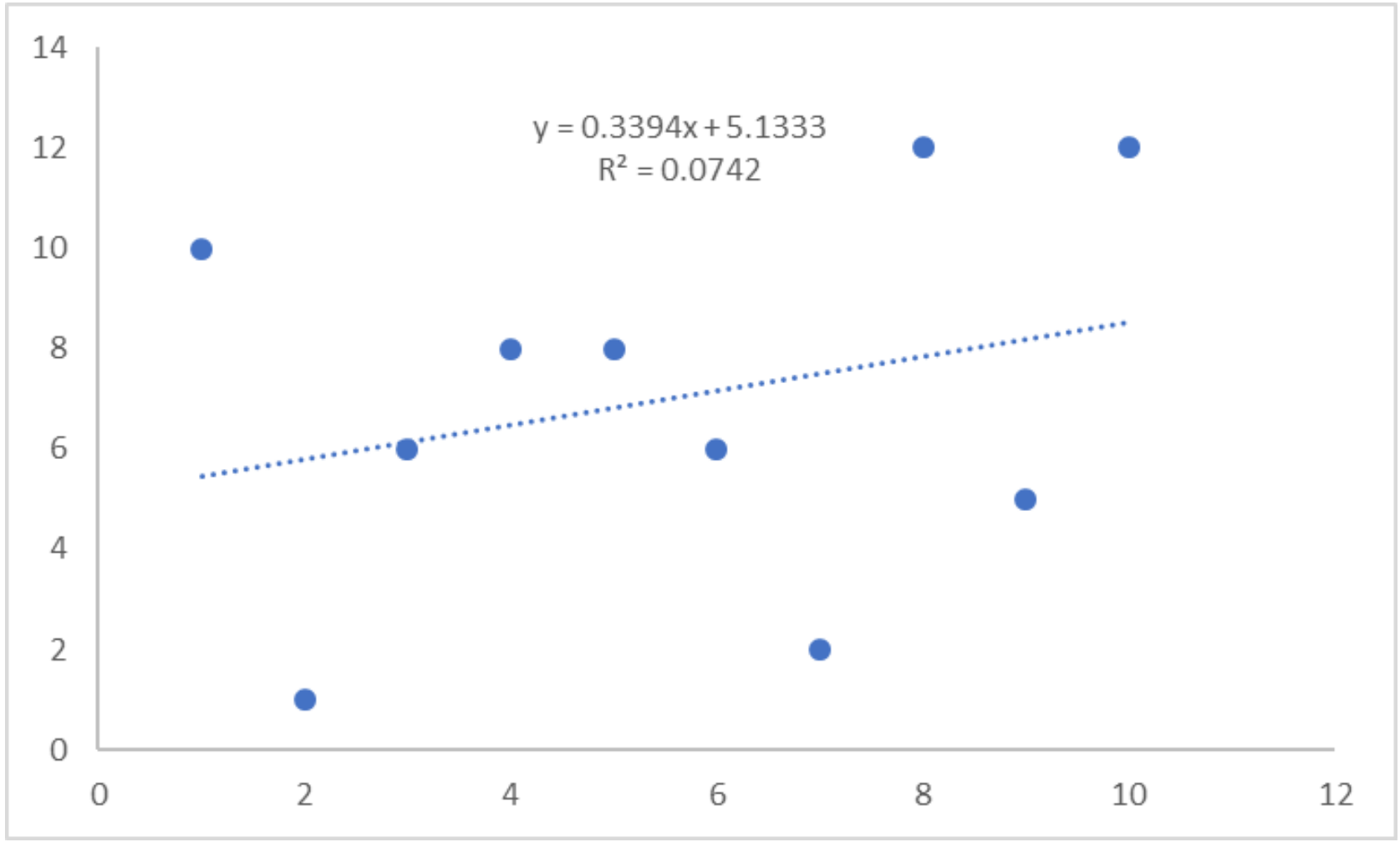
Model vs. realita



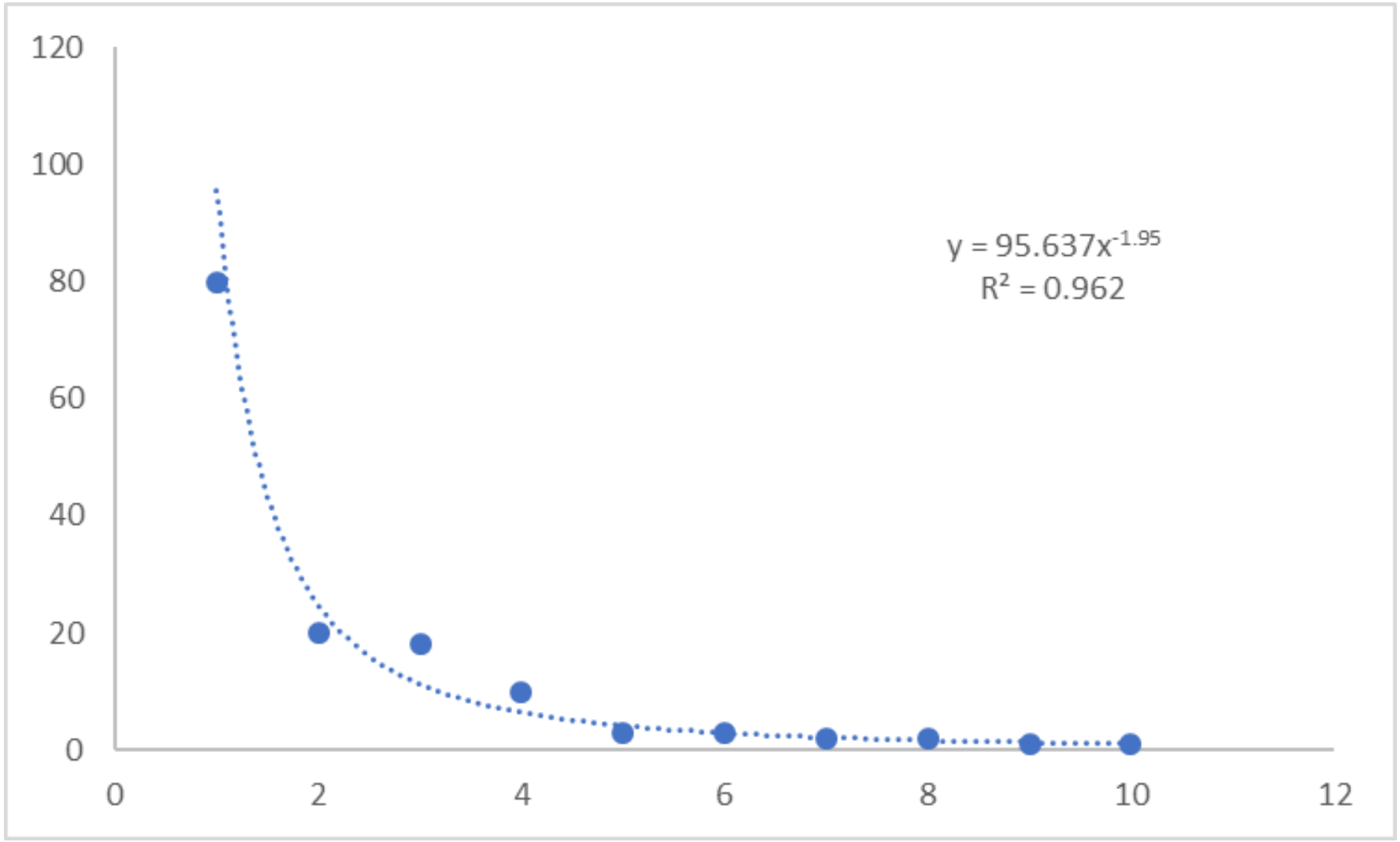
Model vs. realita



Model vs. realita



Model vs. realita



Model vs. realita

- modelujte mocninnou funkci zelená a oranžová data

Příklad – modelování frekvenční distribuce tzv. dependenčních rámců

- Čech, R., Milička, J., Mačutek, J., Koščová, M., Lopatková, M. (2018). Quantitative Analysis of Syntactic Dependency in Czech. In Jiang, J., Liu, H. (eds.). Quantitative Analysis of Dependency Structures. De Gruyter, 53-70.

- http://www.cechradek.cz/publ/2018_Cech_etal_Quantitative_Analysis_Syntactic_Dependency.pdf

Distribuce příslovečných určení

- Čech, Uhlířová (2014)

Adverbial	*r*	*f*	f_r
Place	1	273	27.3
Time	2	204	20.4
Manner	3	172	17.2
Means	4	68	6.8
Aspect	5	61	6.1
Condition	6	59	5.9
Measure	7	52	5.2
Cause	8	30	3.0
Result	9	18	1.8
Origin	10	18	1.8
Purpose	11	17	1.7
Concession	12	16	1.6
Originator	13	12	1.2
∑		1 000	100

Adverbial	Noun	Adverb	Clause
Place	263	9	1
Time	96	104	4
Manner	79	75	18
Means	68	-	-
Aspect	46	13	2
Condition	30	-	29
Measure	21	30	1
Cause	11	-	19
Result	18	-	-
Origin	18	-	-
Purpose	10	-	7
Concession	4	-	12
Originator	12	-	-
∑	676	231	93
R^2	0.98	1	0.96

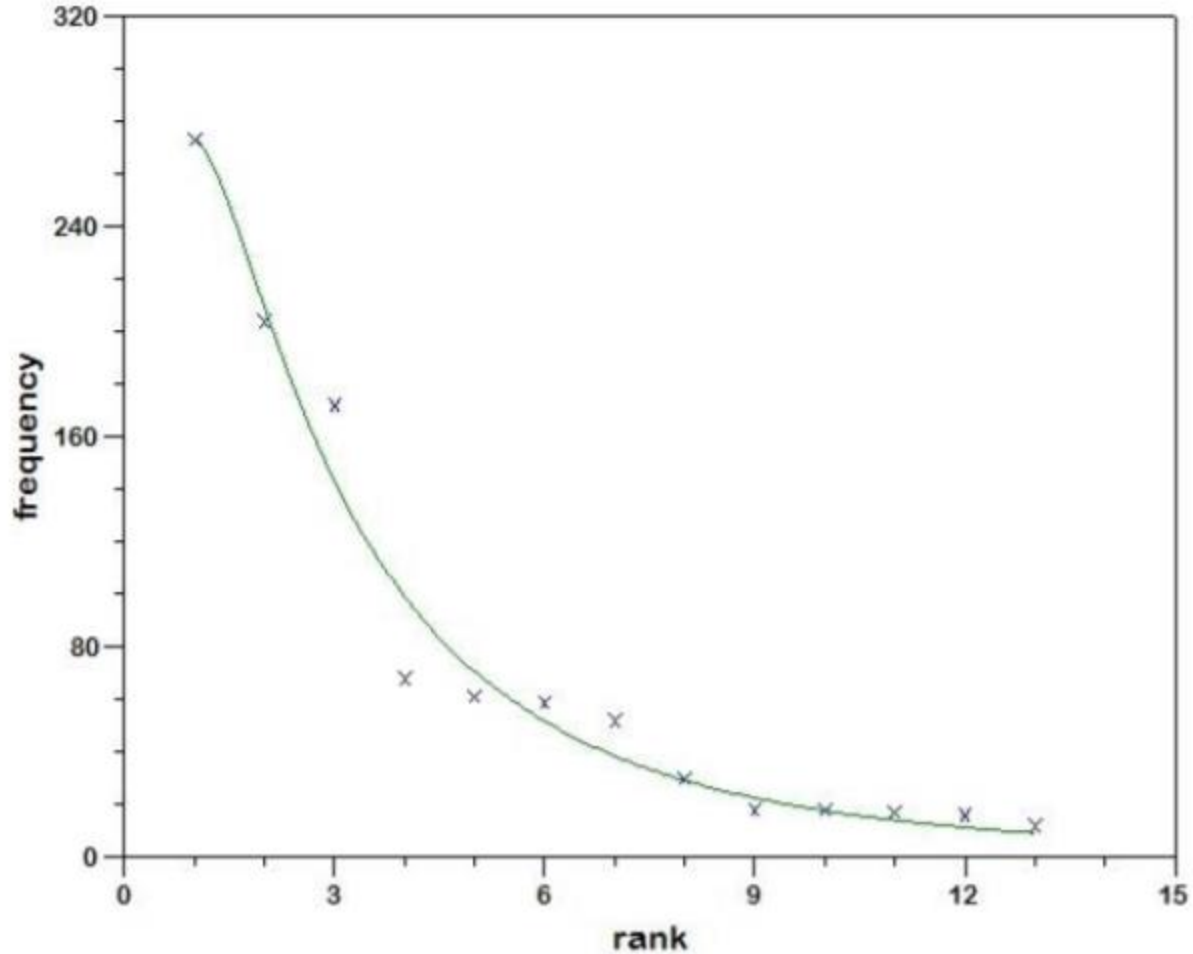


Figure 1. The distribution of all adverbials and the result of the fitting of the Zipf-Alekseev function to the data.

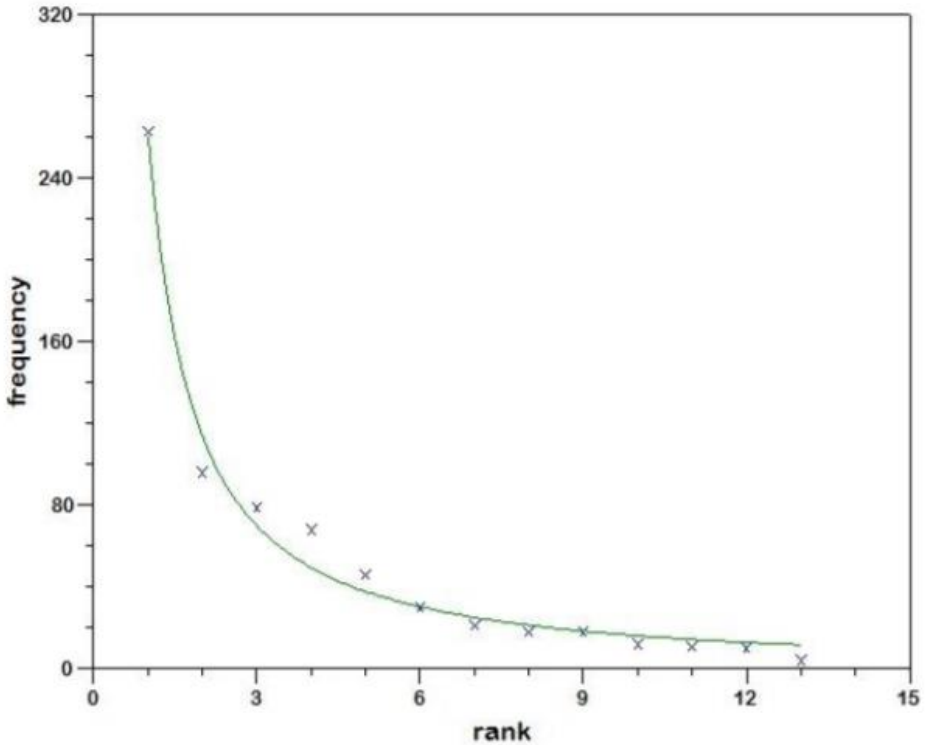


Figure 2. The distribution of adverbials expressed by nouns and the resu fitting of the Zipf-Alekseev function to the data.

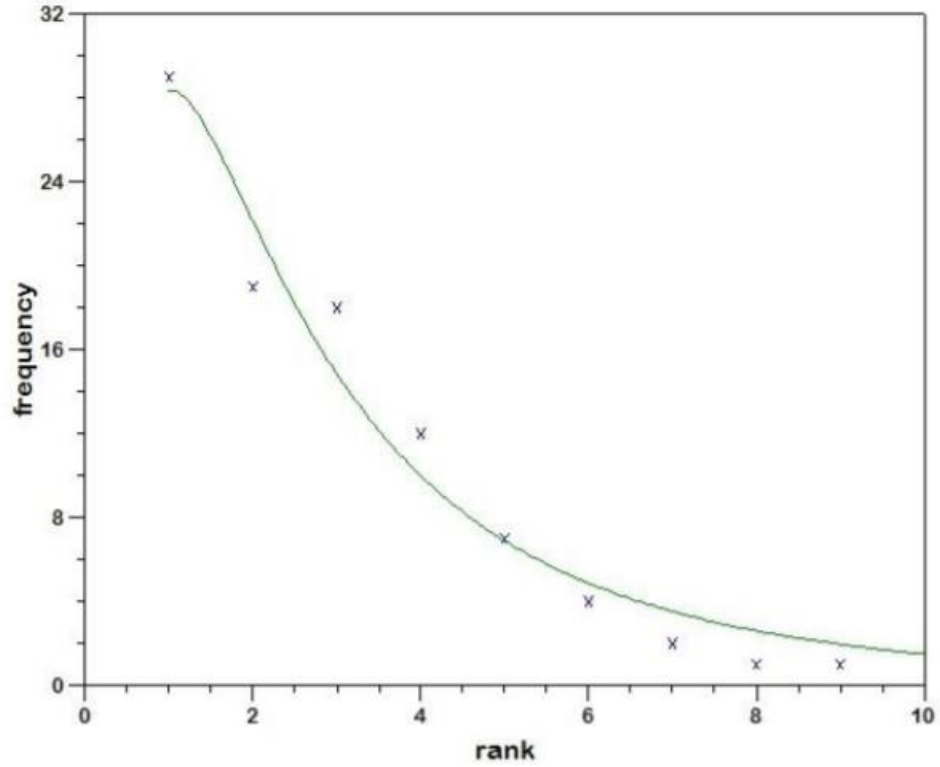


Figure 4. The distribution of adverbials expressed by clauses and the result of the fitting of the Zipf-Alekseev function to the data.

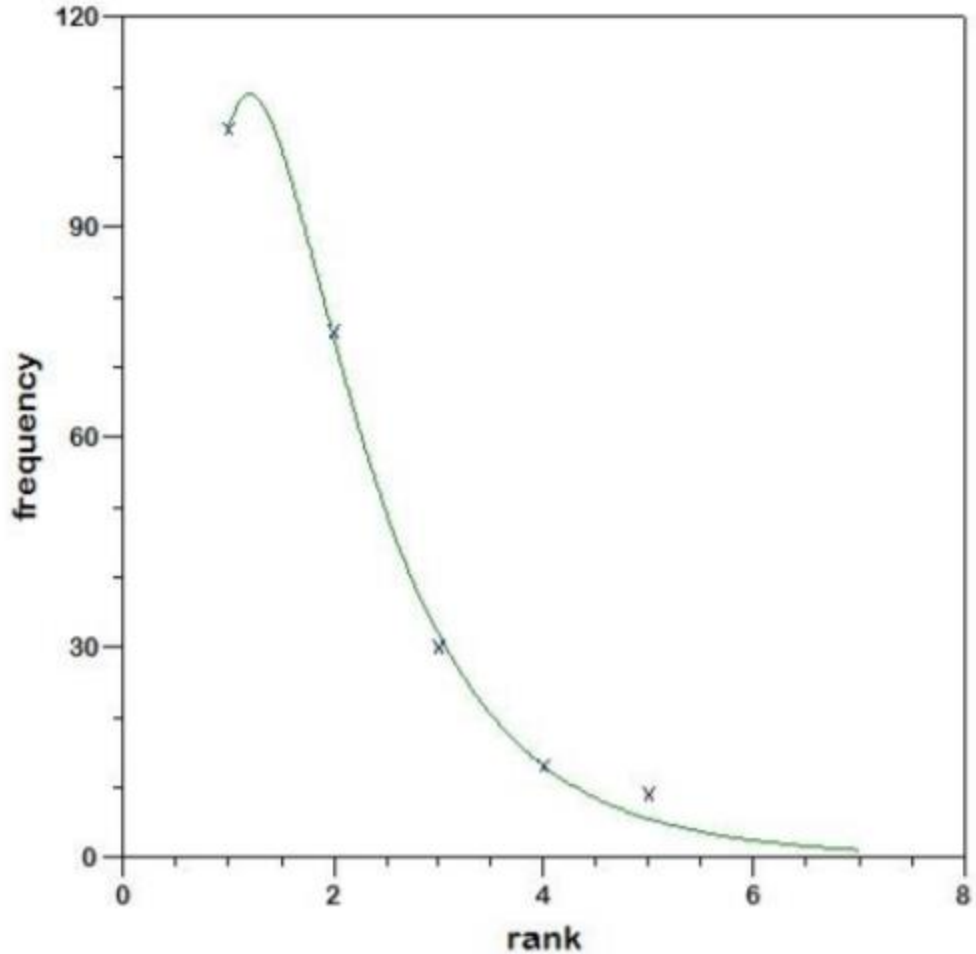
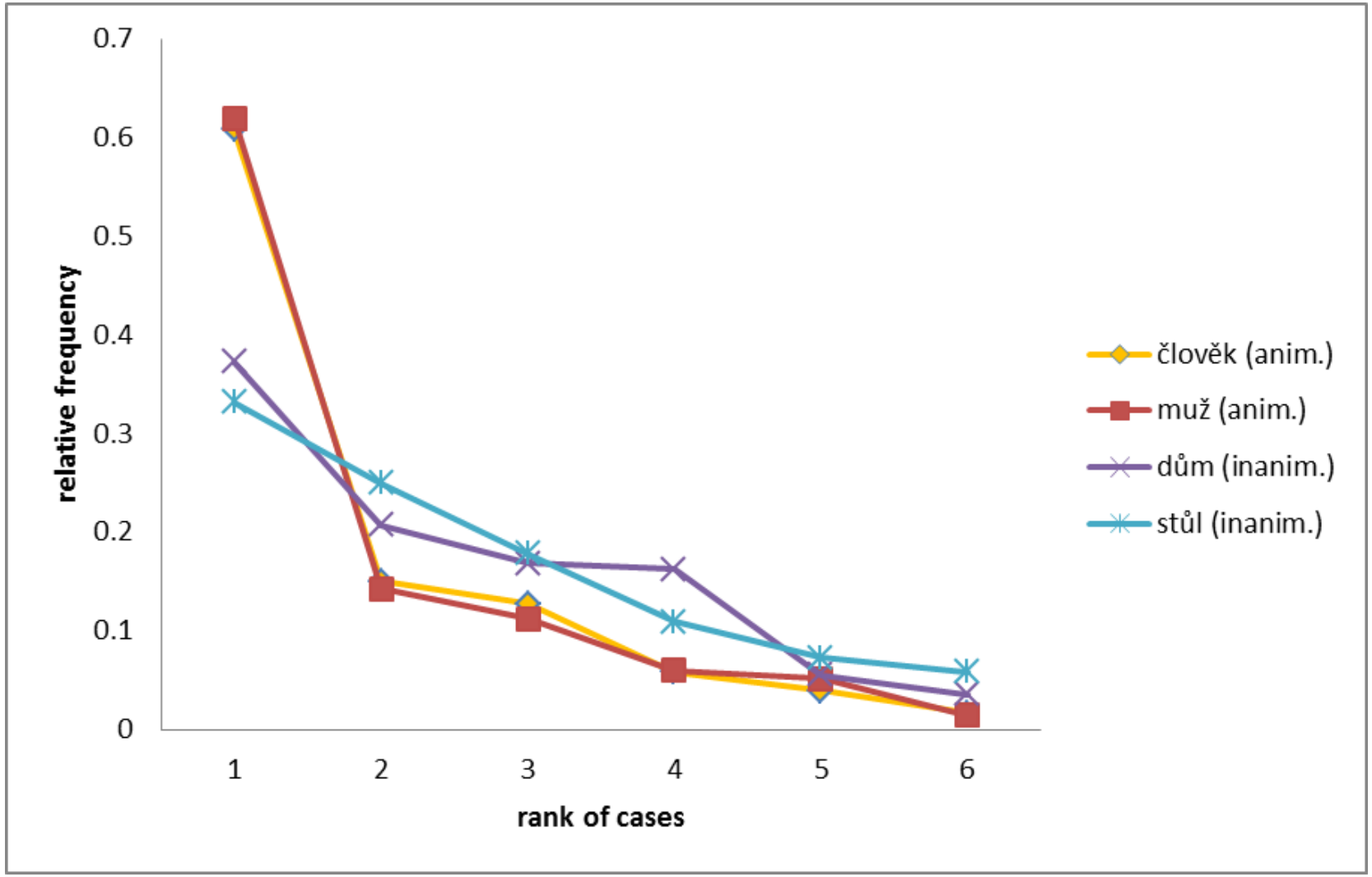


Figure 3. The distribution of adverbials expressed by adverbs and the result of the fitting of the Zipf-Alekseev function to the data.

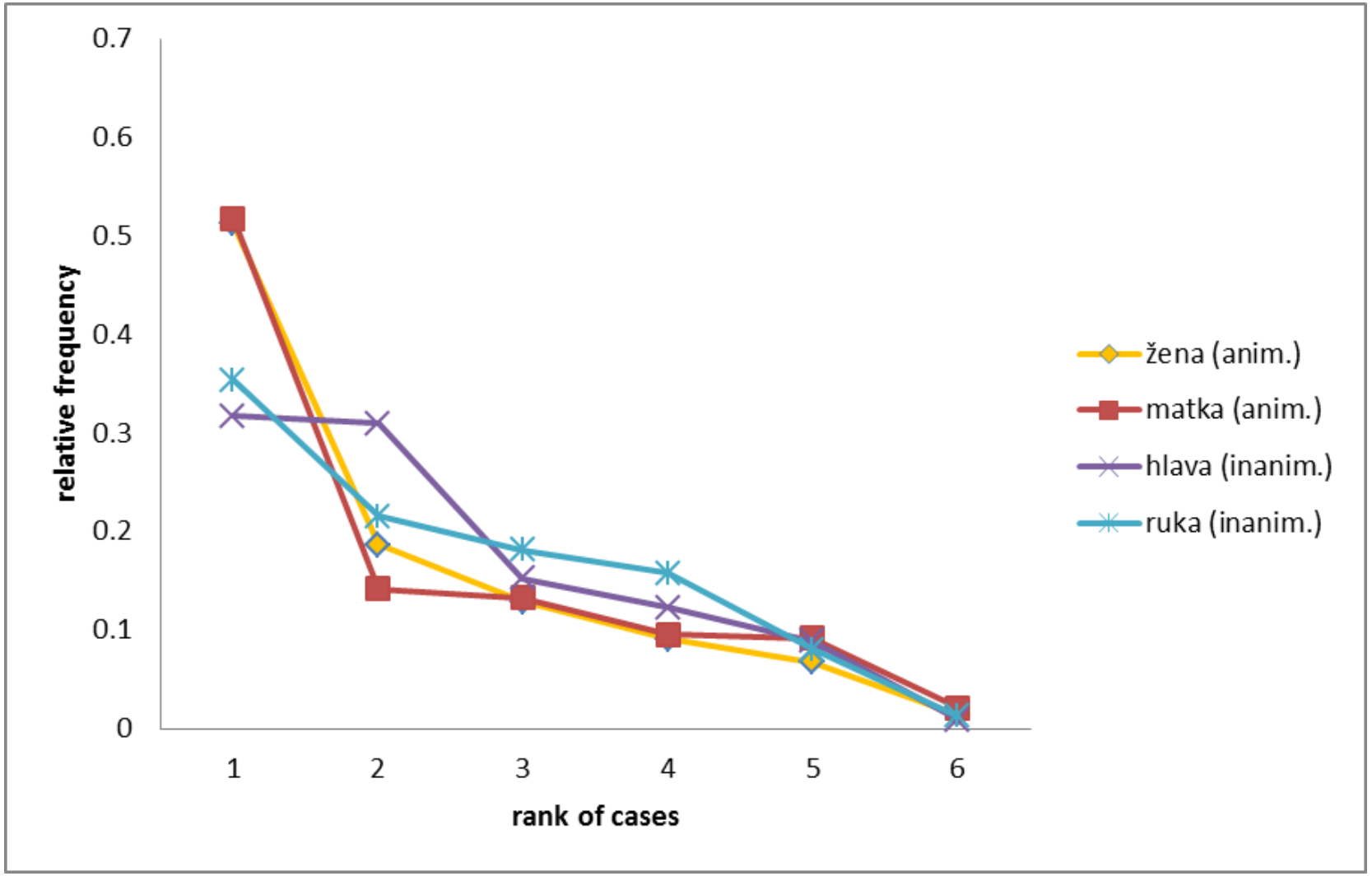
Case study

- Radek Čech, Emmerich Kelih, Jan Mačutek: Impact of semantics on case diversification

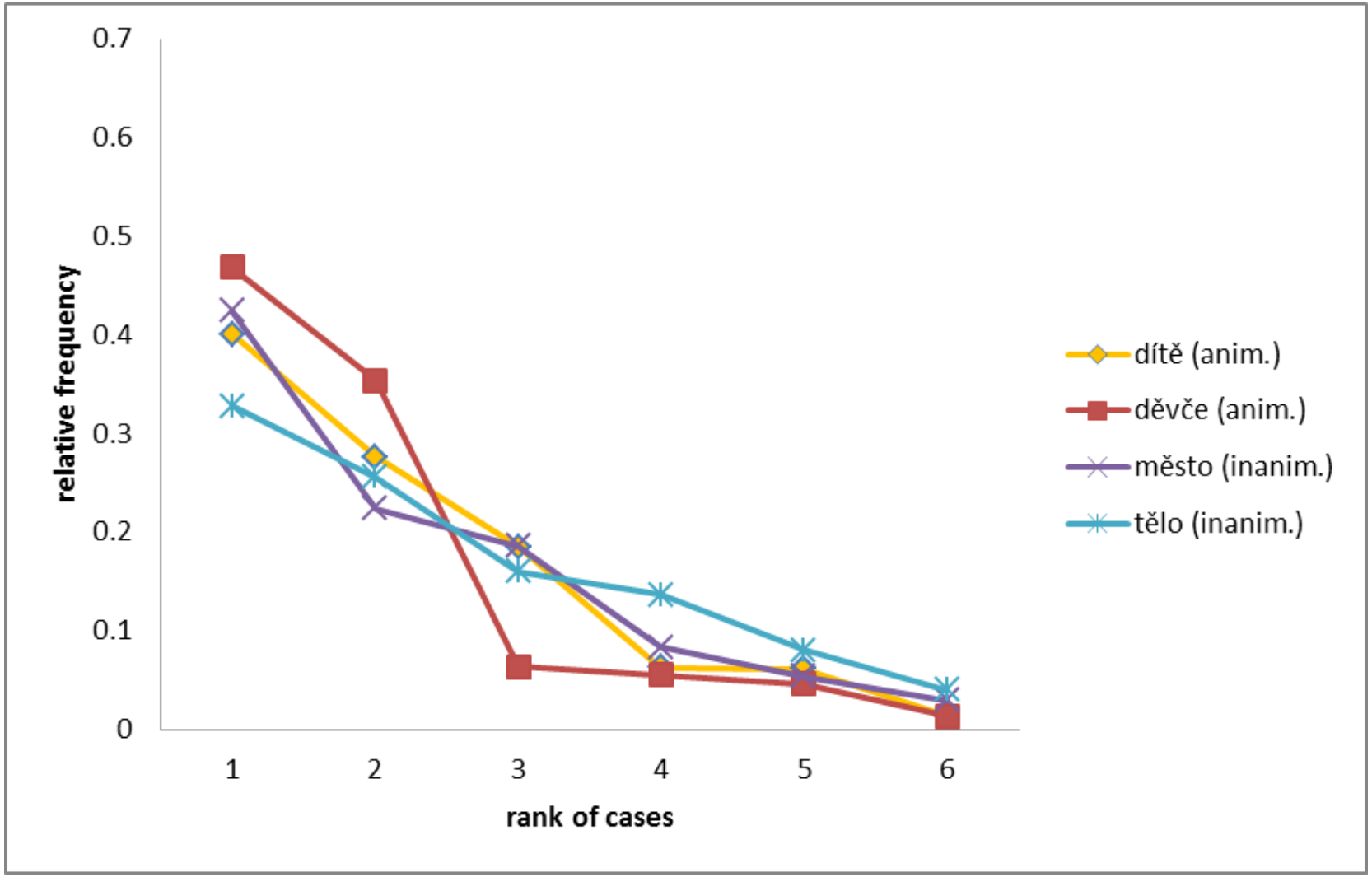
Rankové frekvenční distribuce (mask.)



Rankové frekvenční distribuce (fem.)



Rankové frekvenční distribuce (neut.)



Model

$$y = ae^{-bx}$$

x … pořadí pádu

y … frekvence pádu

a, b … parametry

speciální případ Wimmerova-Altmannova modelu

Výsledky aplikace modelu na data (mask.)

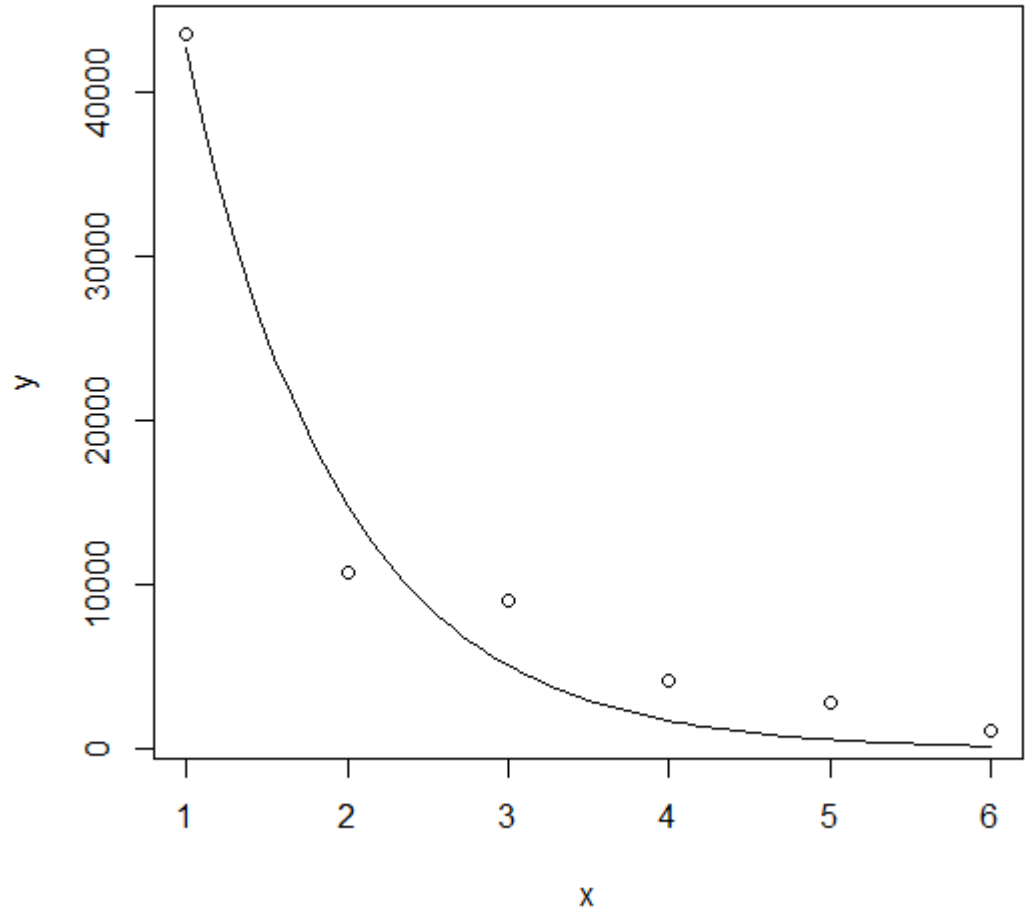
mask. anim.

lemma	*a*	*b*	R^2
člověk	123208.0	1.059	0.9723
muž	91962.4	1.149	0.9711
pan	39832.5	0.685	0.9887
otec	48060.8	1.014	0.9476
ředitel	46335.5	1.142	0.9950

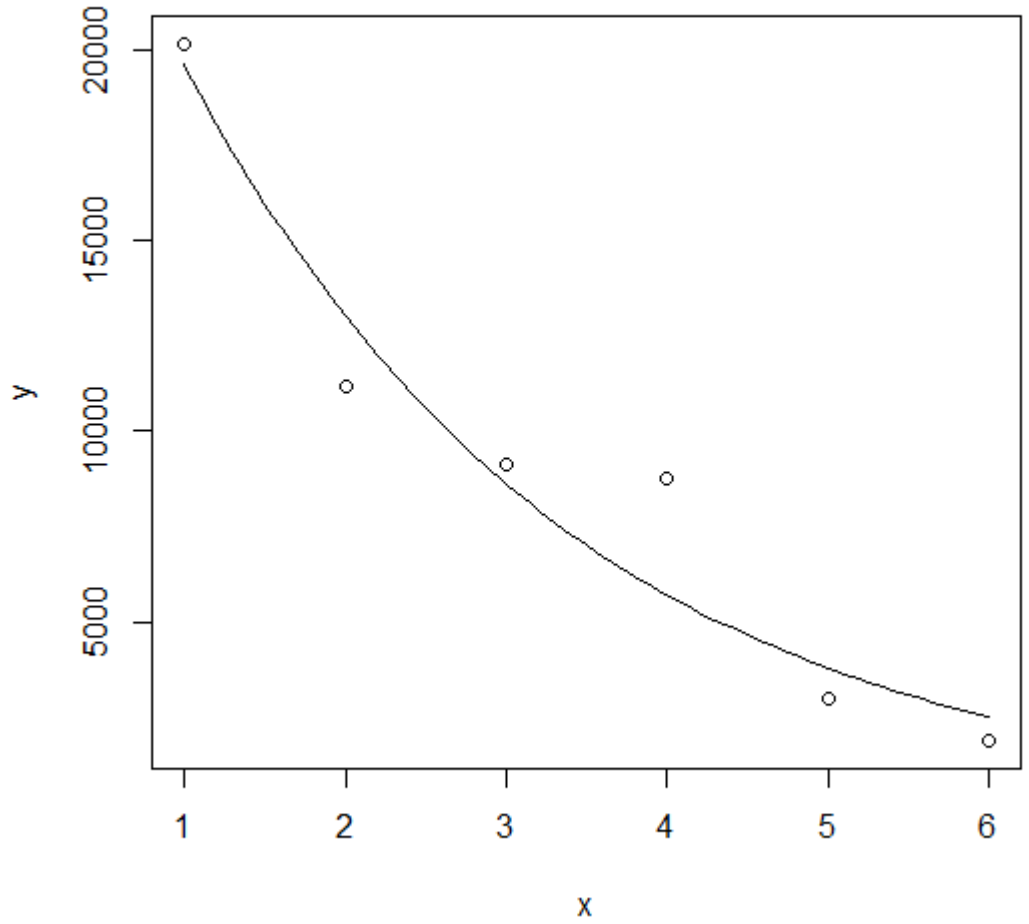
mask. inanim.

lemma	*a*	*b*	R^2
dům	29591.9	0.411	0.9400
stůl	11730.4	0.375	0.9772
měsíc	18996.9	0.738	0.9852
vzduch	8805.4	0.374	0.9268
byt	10488.9	0.456	0.9221

Aplikace modelu na lemma „člověk“



Aplikace modelu na lemma „dům“



Výsledky aplikace modelu na data (fem.)

fem. anim.

lemma	*a*	*b*	R^2
žena	46525.9	0.723	0.9664
matka	32093.7	0.772	0.9108
paní	56238.4	1.247	0.9667
dívka	17406.2	0.889	0.9771
dcera	8179.4	0.471	0.9757

fem. inanim.

lemma	*a*	*b*	R^2
hlava	32671.4	0.380	0.9046
ruka	25636.7	0.385	0.9366
škola	25894.7	0.506	0.9778
ulice	31193.3	0.772	0.9802
tvář	11521.9	0.308	0.8232

Výsledky aplikace modelu na data (neut.)

neut. anim.

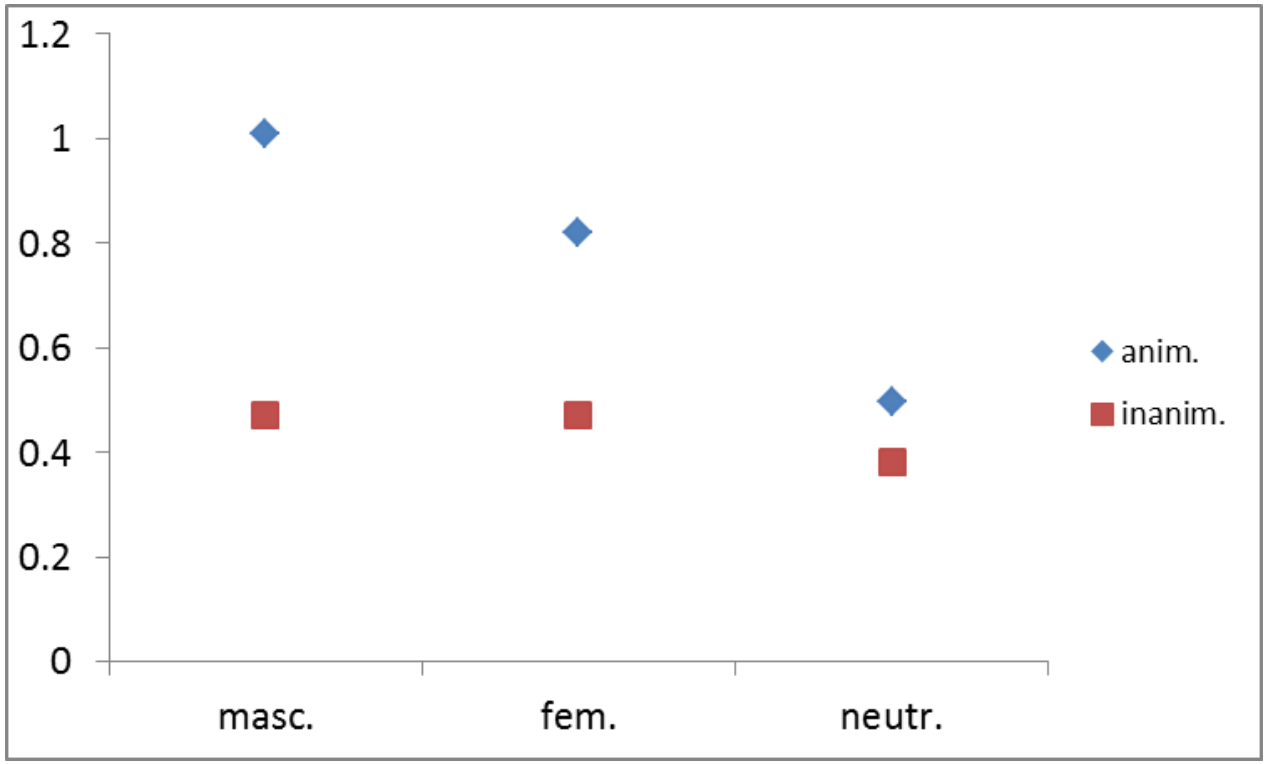
lemma	*a*	*b*	R^2
dítě	19762.8	0.492	0.9743
děvče	2220.8	0.596	0.9281
miminko	1135.3	0.403	0.8960
děcko	587.2	0.430	0.9574
děťátko	691.4	0.555	0.9854

neut. inanim.

lemma	*a*	*b*	R^2
město	53158.3	0.515	0.9819
tělo	15578.0	0.369	0.9673
auto	10164.0	0.299	0.8910
divadlo	11608.6	0.423	0.9291
srdce	7553.4	0.295	0.9777

Průměry parametru *b*

rod	anim.	inanim.
mask.	1.010	0.471
fem.	0.821	0.471
neut.	0.495	0.380



Rozdělení dat

- diskrétní modely
- může nabývat pouze spočetně izolovaných hodnot z množiny

Table 4.4

b) As to Text 2: fitting of the 1-displaced hyper-Poisson distribution to the word lengths (syllables per word) in: Pestalozzi, *Das Menschenvertilgen* (Pestalozzi, *Fabeln*, p. 42f.)

x	n_x	NP_x
1	130	128.99
2	93	87.97
3	21	29.93
4	10	6.79
5	1	1.33
$a = 0.6791$		$X_2^2 = 4.566$
$b = 0.9958$		$P = 0.10$

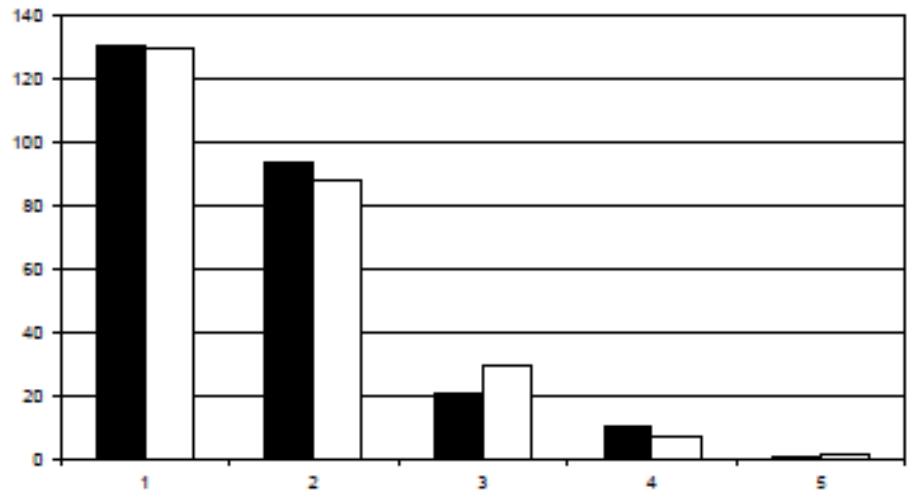


Figure 4.4. Concerning the data in Table 4.4

Table 4.5

As to Text 3: fitting the 1-displaced hyper-Poisson distribution to word length (syllables per word) in: Böll, Brief an E.-A. Kunz, 11.11.52

x	n_x	NP_x
1	226	226.22
2	125	128.09
3	57	51.05
4	13	15.70
5	4	3.93
6	1	1.01
$a = 1.3462$		$X_3^2 = 1.233$
$b = 2.3775$		$P = 0.75$

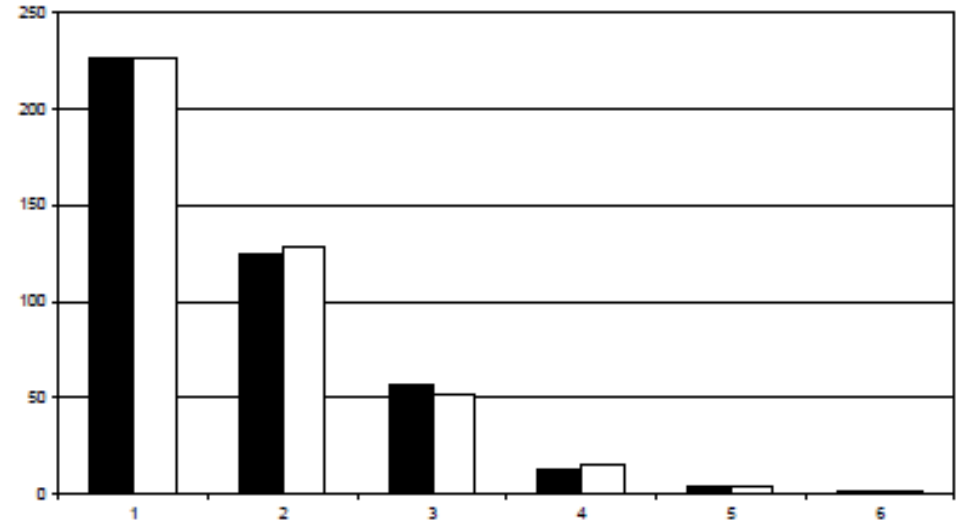


Figure 4.5. Concerning the data in Table 4.5

Table 4.4

b) As to Text 2: fitting of the 1-displaced hyper-Poisson distribution to the word lengths (syllables per word) in: Pestalozzi, *Das Menschenvertilgen* (Pestalozzi, *Fabeln*, p. 42f.)

x	n_x	NP_x
1	130	128.99
2	93	87.97
3	21	29.93
4	10	6.79
5	1	1.33
$a = 0.6791$		$X_2^2 = 4.566$
$b = 0.9958$		$P = 0.10$

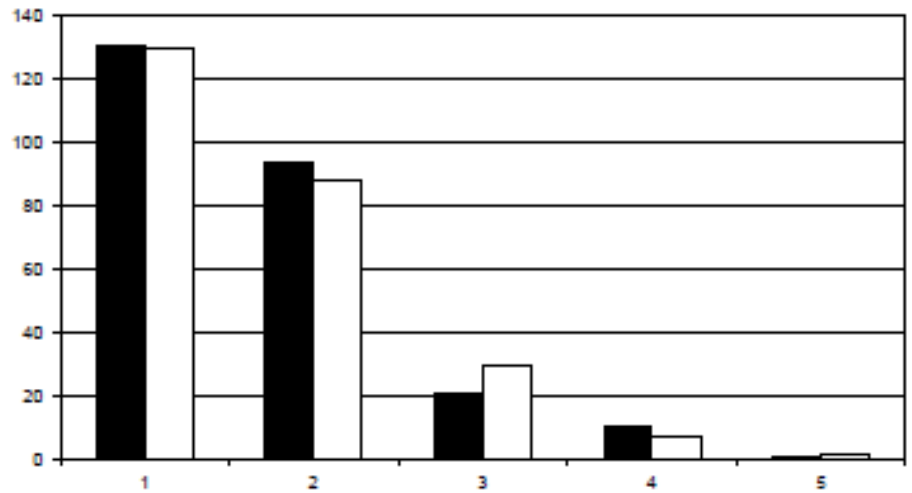


Figure 4.4. Concerning the data in Table 4.4

Table 4.5

As to Text 3: fitting the 1-displaced hyper-Poisson distribution to word length (syllables per word) in: Böll, Brief an E.-A. Kunz, 11.11.52

x	n_x	NP_x
1	226	226.22
2	125	128.09
3	57	51.05
4	13	15.70
5	4	3.93
6	1	1.01
$a = 1.3462$		$X_3^2 = 1.233$
$b = 2.3775$		$P = 0.75$

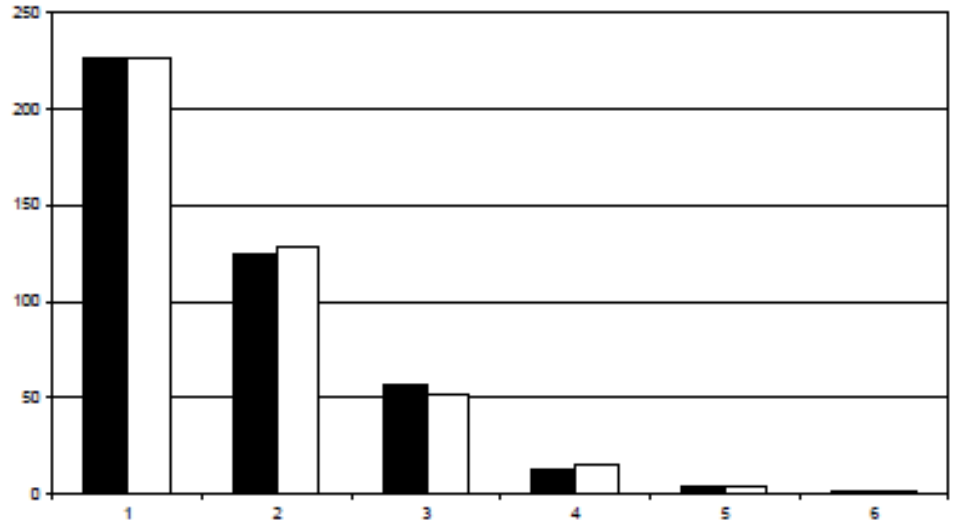


Figure 4.5. Concerning the data in Table 4.5

Quantitative Linguistics, an Invitation

Karl-Heinz Best

Otto Rottmann

2017

RAM-Verlag