ANALÝZA DAT V R

2. POPISNÉ STATISTIKY

Mgr. Markéta Pavlíková
Katedra pravděpodobnosti a matematické statistiky MFF UK

www.biostatisticka.cz

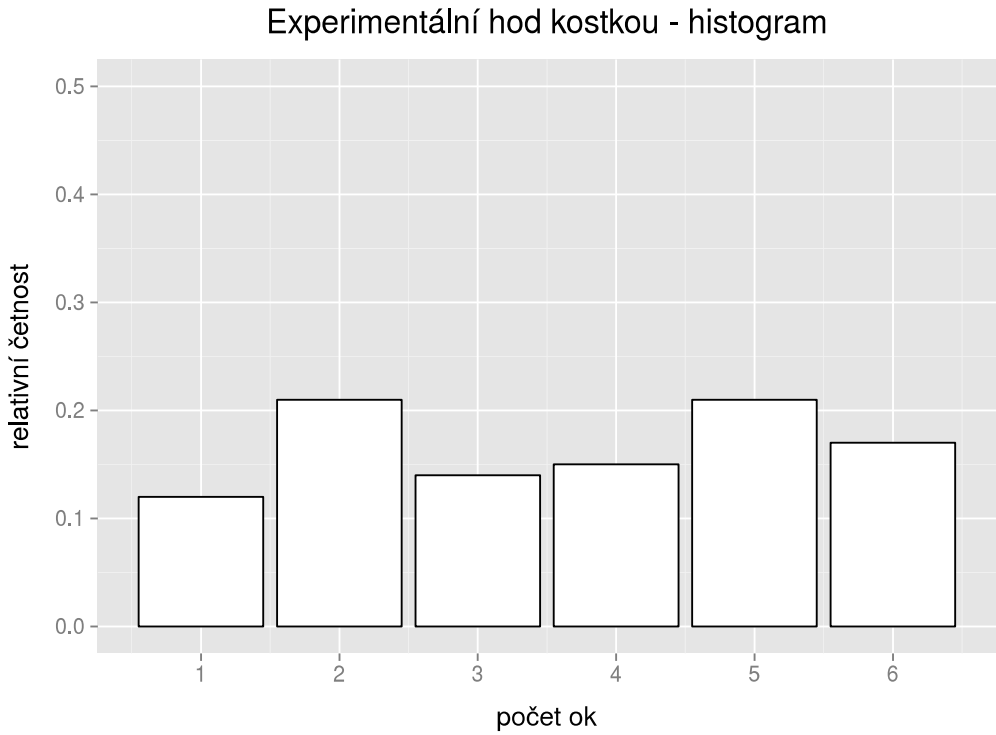
CO SE SKRÝVÁ V DATECH

- data sbíráme proto, abychom porozuměli skutečnosti, principům, zákonitostem
- musíme umět data nějak popsat, zestručnit
  - rozsah hodnot
  - rozložení hodnot
  - jsou některé hodnoty významnější, častější než jiné?
- klademe si otázky po podstatě
  - jsou naměřené hodnoty stejné jako naměřené dříve / jinde?
  - existuje nějaké pravidlo, podle kterého měřené hodnoty vznikají?

CO SE SKRÝVÁ V DATECH

Házíme 100x kostkou

j		n_j	$f_j = n_j/n$
1	𝍸 𝍸 ||	12	0,12
2	𝍸 𝍸 𝍸 |	21	0,21
3	𝍸 𝍸 |||	14	0,14
4	𝍸 𝍸 𝍸	15	0,15
5	𝍸 𝍸 𝍸 |	21	0,21
6	𝍸 𝍸 ||	17	0,17
	$n = 100$		1,00



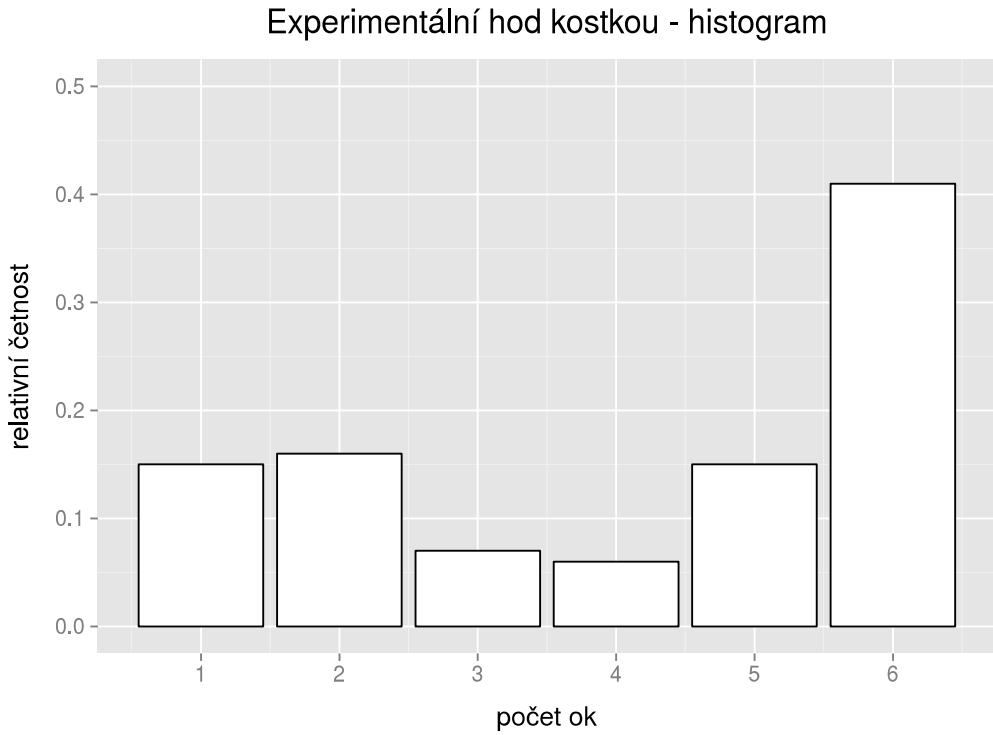
Popis: Absolutní četnost a relativní četnost, tabulka, histogram

Podstata: Existuje nějaké pravidlo za výsledky?

CO SE SKRÝVÁ V DATECH

Házíme 100x kostkou

j		n_j	$f_j = n_j/n$
1	𝍸 𝍸 𝍸	15	0,15
2	𝍸 𝍸 𝍸 I	16	0,16
3	𝍸 II	7	0,07
4	𝍸 I	6	0,06
5	𝍸 𝍸 𝍸	15	0,15
6	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 I	41	0,41
		$n = 100$	1,00

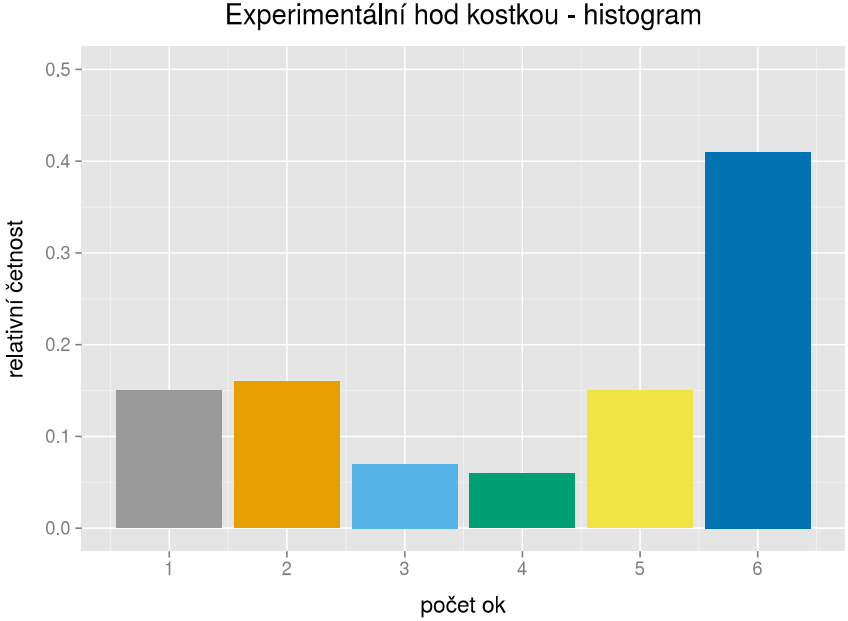


Popis: Absolutní četnost a relativní četnost, tabulka, histogram

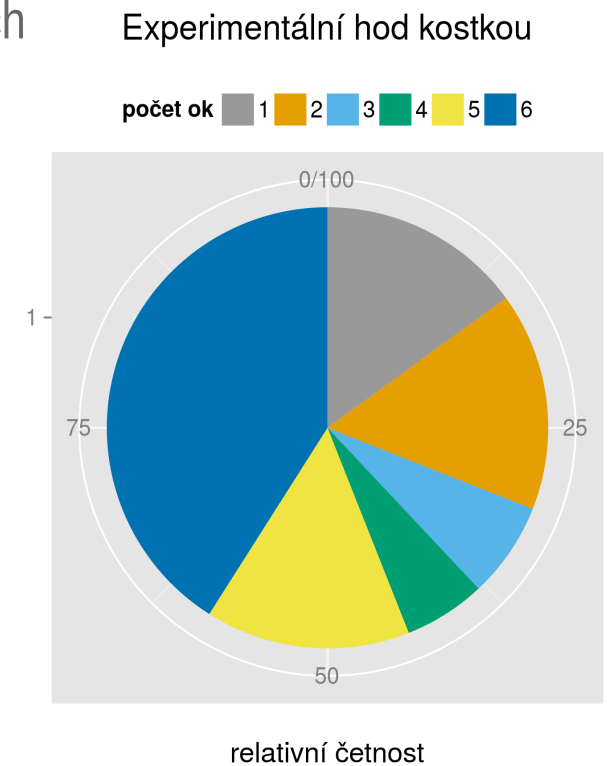
Podstata: Jsou kostky stejné? Liší se od předpokládaného principu rovnoměrného výskytu?

KATEGORIÁLNÍ PROMĚNNÁ

počet ok	absolutní četnost	relativní četnost
1	15	0.15
2	16	0.16
3	7	0.07
4	6	0.06
5	15	0.15
6	41	0.41
celkem	100	1.00



- tabulka ukazuje konkrétní data
- sloupcový graf / histogram vizualizuje četnosti
  - umožňuje okometrické srovnání
- koláčový graf
  - vypadá pěkně, ale nedává reálnou představu o konkrétních hodnotách
  - jen velmi výjimečně

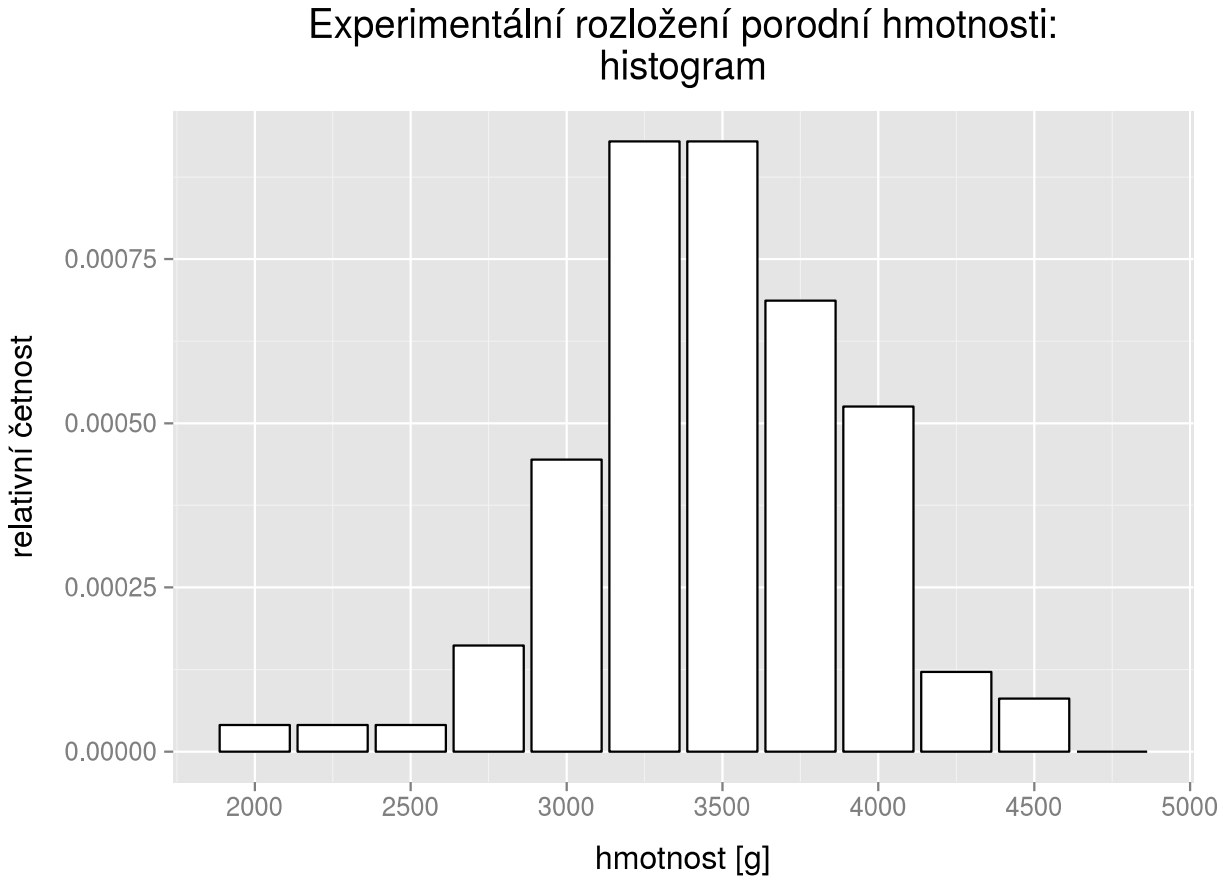


SPOJITÁ PROMĚNNÁ

Vzorek 99 novorozených dětí, porodní hmotnost

Jak data popsat?

- rozsah
- nejčastější hodnota
- „prostřední“ hodnota
- symetrie, „sešikmenost“
- „rozpláclost“
- ...



POPISNÉ STATISTIKY - POLOHA

- minimum, maximum
  - odkud kam jdou data
  - dávají hodnoty smysl? pokrývají „správný“ rozsah?
  - jsou tam nějaká extrémní měření? jsou správně?

- aritmerický průměr
  - součet všech hodnot / počet všech hodnot

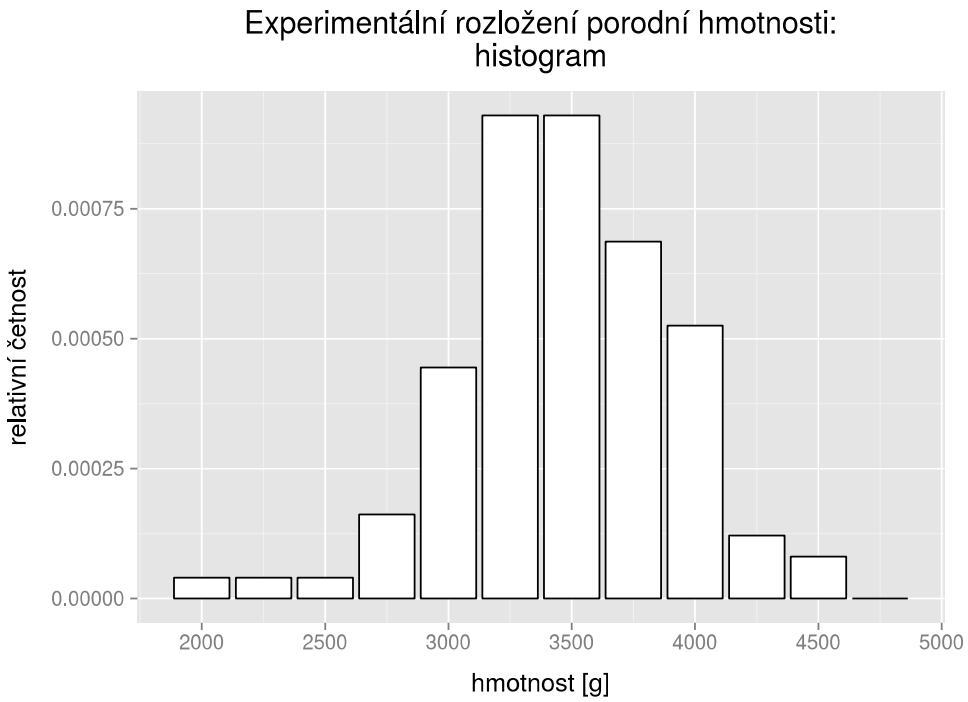
$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

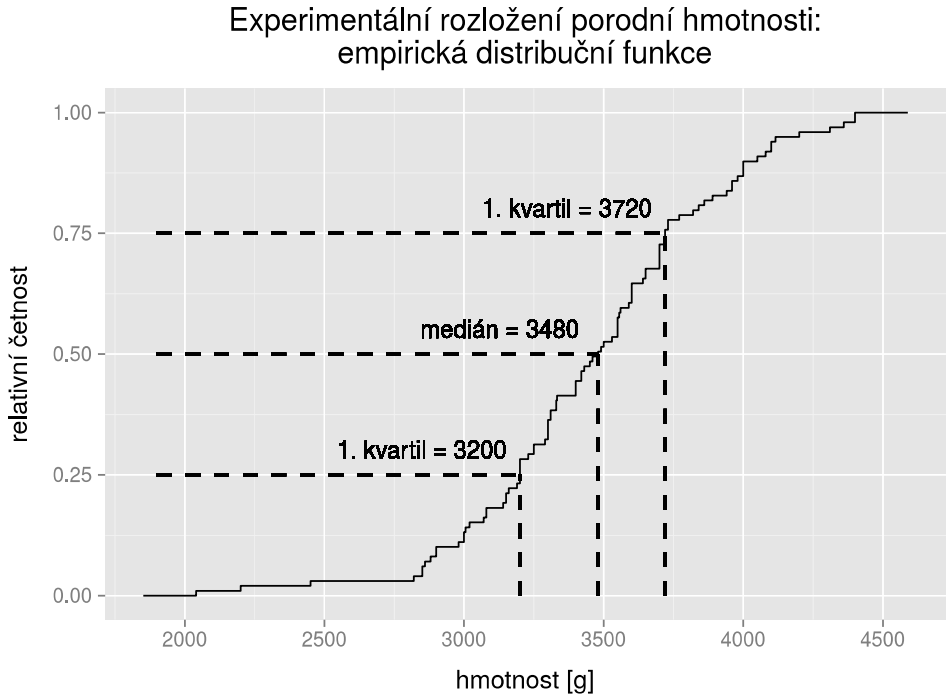
POPISNÉ STATISTIKY - POLOHA

- modus
  - nejčastější hodnota, může jich být i víc (více hrbů)
- medián
  - dělí soubor na polovinu
  - pod ním je polovina dat, nad ním je polovina dat
  - často výrazně spolehlivější informace než průměr
- kvartily
  - dělí soubor na čtvrtiny
  - pod 1. kvartilem leží čtvrtina dat
  - nad 3. kvartilem leží čtvrtina dat

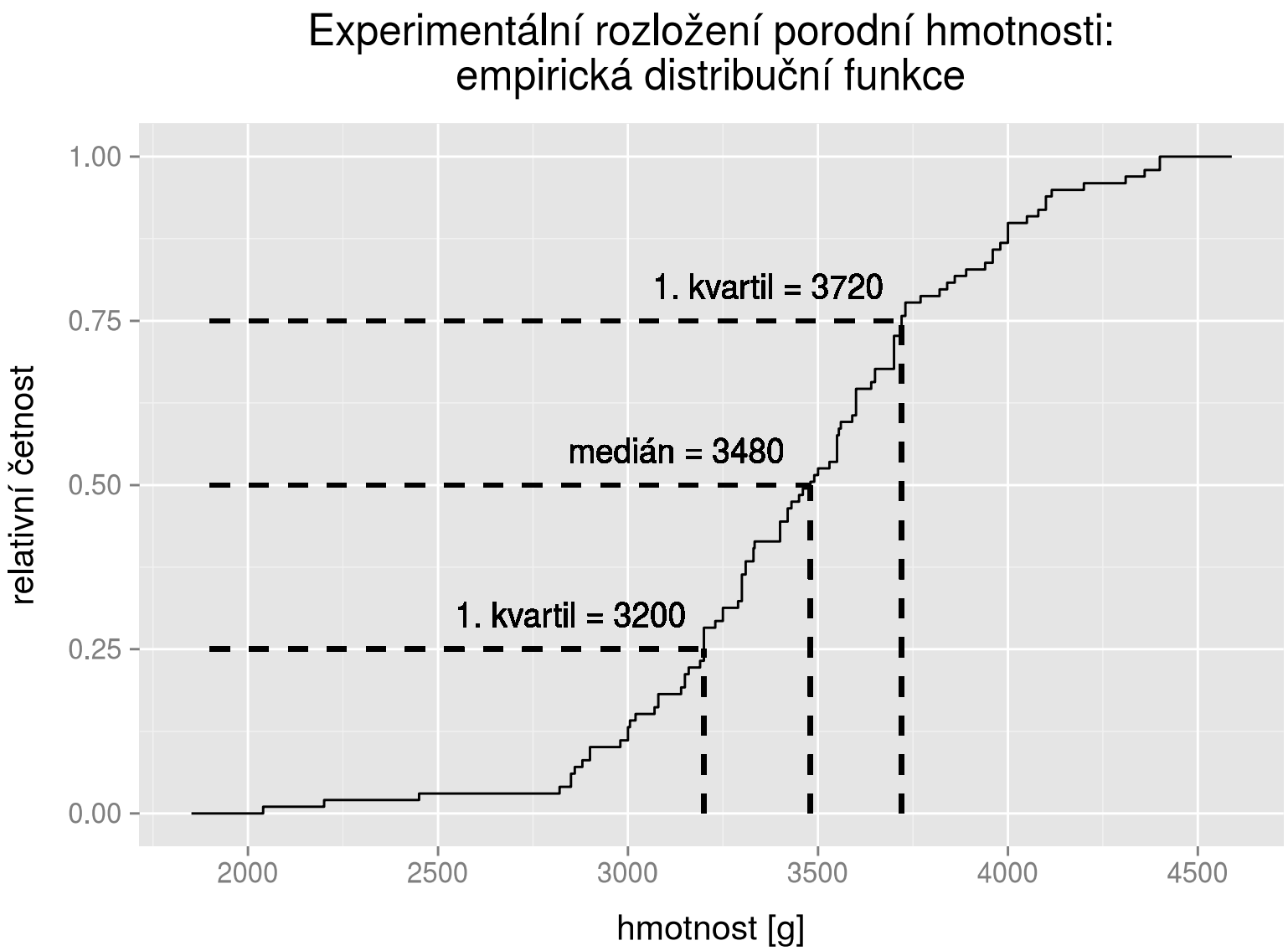
MEDIÁN, KVARTILY – JAK VYPOČÍST

- seřadíme data podle velikosti (novorozenci: 99 hodnot)
- najdeme polovinu (novorozenci: 50. hodnota)
- pokud sudý počet: vezmeme průměr z těch dvou uprostřed





MEDIÁN, KVARTILY – JAK VYPOČÍST

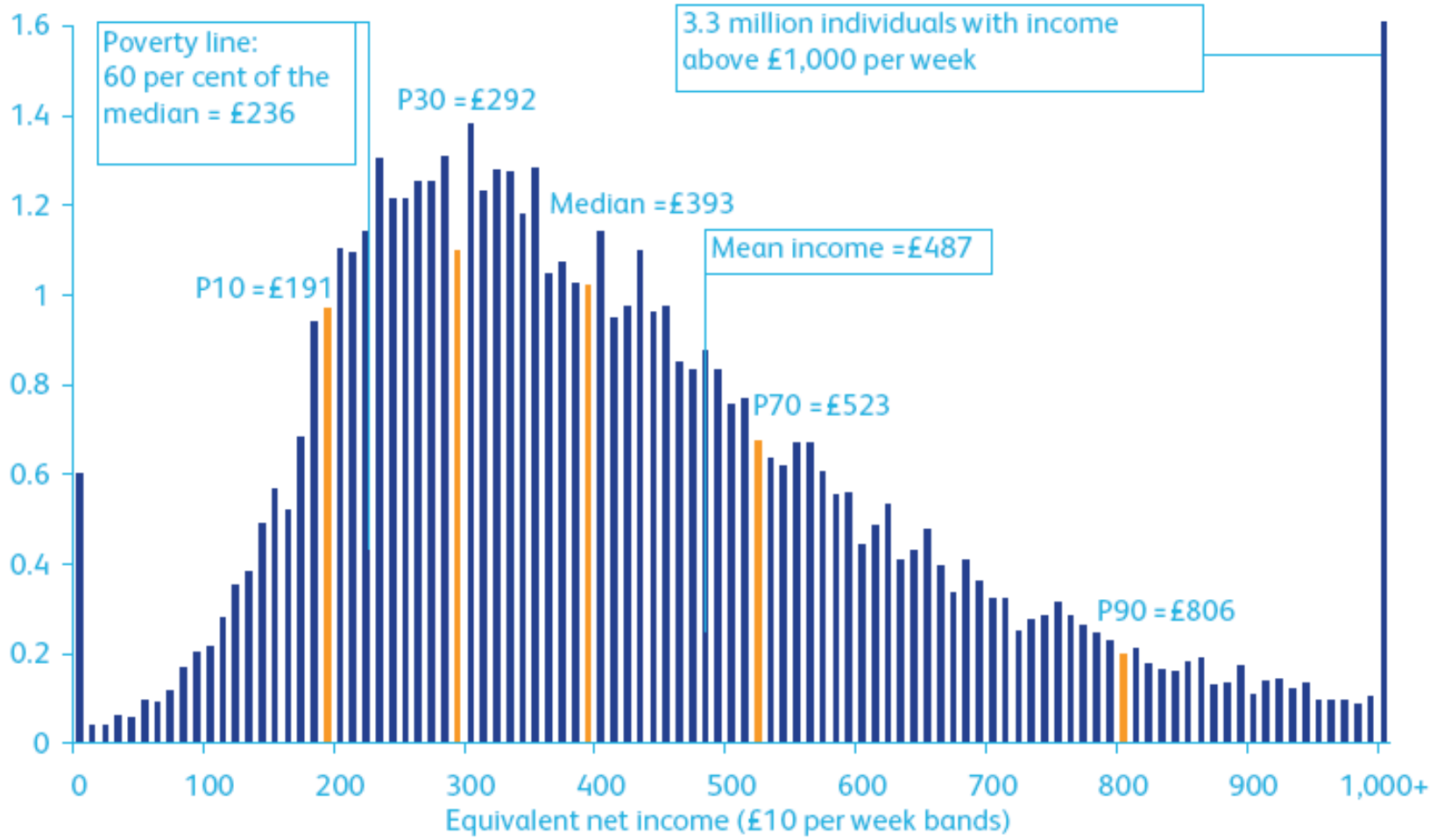


MEDIÁN vs. PRŮMĚR

- průměr
  - dobře určuje „prostředek“ u symetrických dat (novorozenci: 3470g)
  - nejlépe funguje, když je rozložení typu Gaussova křivka
  - nefunguje v případech typu A: 1 kuře B, C: žádné kuře, každý snědl třetinu kuřete
  - klasický příklad: 2/3 zaměstnanců mají podprůměrný plat
  - bohatí táhnou průměr svým směrem
- medián
  - u symetrických dat bude poblíž průměru (novorozenci: 3480g)
  - u nesymetrických dat bude blíže „těžší“ straně
  - ignoruje výši extrémů, bere v potaz jen jejich počet
  - plat pod mediánem má vždy 1/2 zaměstnanců
  - bohatých je málo, a tak hrají ve výpočtu menší roli
  - u kuřat: jaký bude medián?

MEDIÁN vs. PRŮMĚR

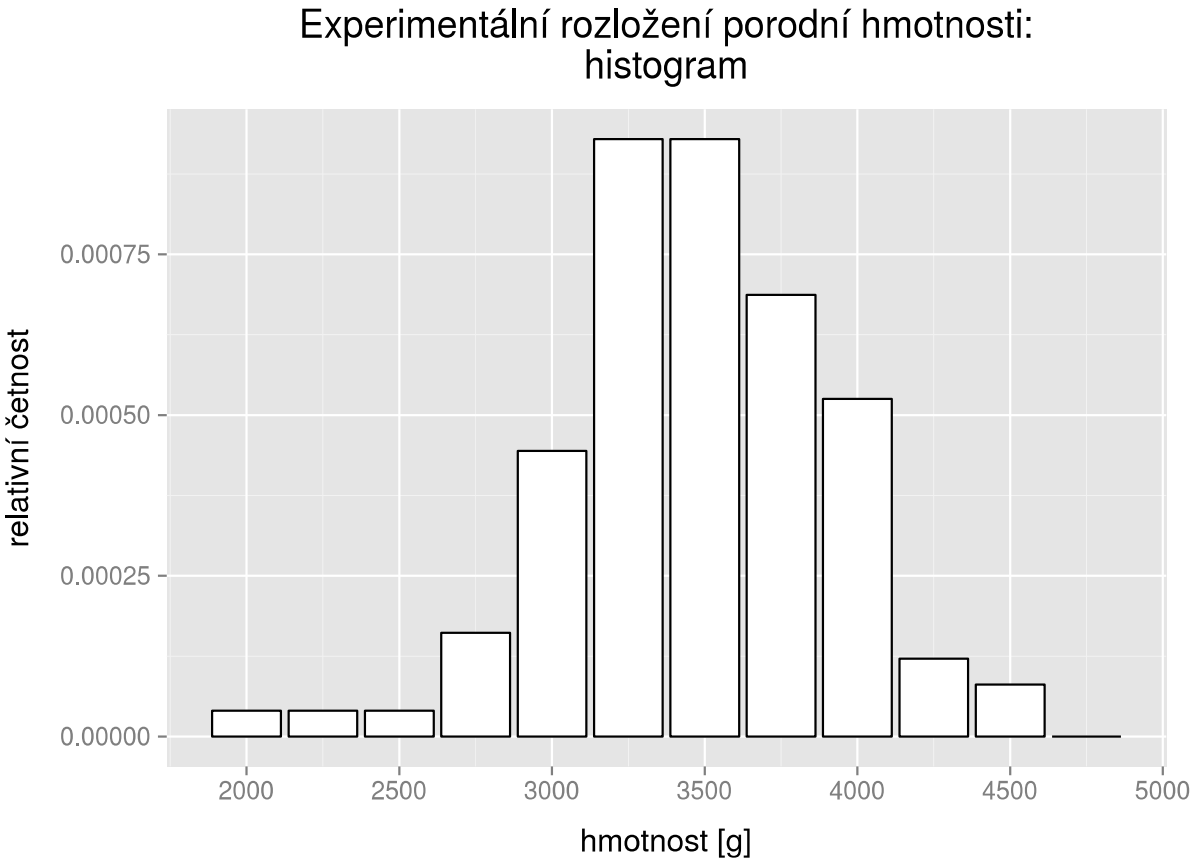
Half of the population has income below and half above £393 per week (adjusted for household size)



POPISNÉ STATISTIKY - POLOHA

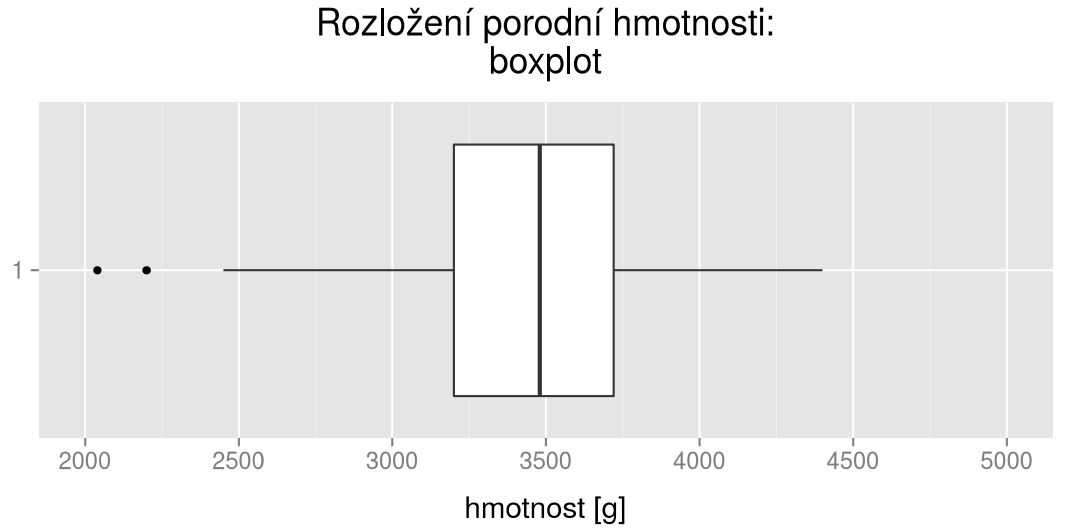
Vlastnosti dobré míry polohy

- přičteme-li ke každé hodnotě *x* stejnou konstantu ***a***, posune se míra polohy také o ***a***
  - příklad: asijské děti jsou o cca 250g menší než evropské
- vynásobíme-li každou hodnotu *x* stejnou konstantou ***b***, míra polohy bude také vynásobena ***b***
  - příklad: zachovává změny jednotek
- tedy citlivost vůči změně posunutí a změně měřítka



Boxplot

- shrnuje míry polohy
- uprostřed medián
- okraj boxu kvartily
- vousky 1.5 x IQR
- tečky jsou „outliers“



POPISNÉ STATISTIKY - VARIABILITA

- rozpětí (range)
  - maximum - minimum
  - základní informace o variabilitě dat
- interkvartilové rozpětí (interquartile range)
  - třetí - první kvartil
- výběrový rozptyl
  - čím dále od průměru tím větší váha
  - mocnina zdůrazňuje vzdálenější
- směrodatná odchylka (SD)
  - odmocnina rozptylu
  - má stejnou jednotku jako data

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

POPISNÉ STATISTIKY - VARIABILITA

Vlastnosti dobré míry variability

- nezávisí na míře polohy
- když celá data posunu o ***a***, tak míra variability zůstane stejná
- když data vynásobím ***b***, tak je míra variability také ***b***-krát vyšší
- tedy citlivost na změnu měřítka, necitlivost na změnu polohy