

ANALÝZA DAT V R

3. POPISNÉ STATISTIKY, NÁHODNÁ VELIČINA

Mgr. Markéta Pavlíková

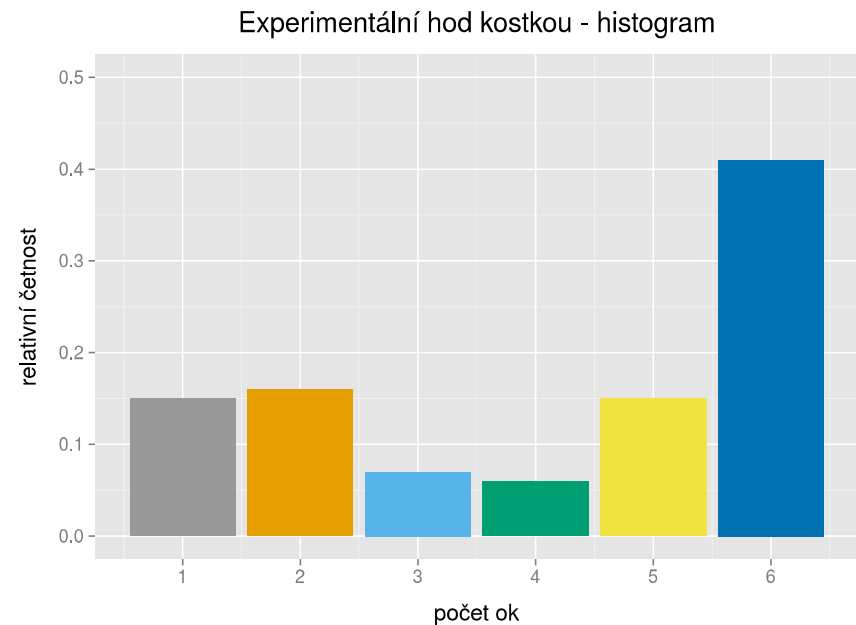
Katedra pravděpodobnosti a matematické statistiky MFF UK

www.biostatisticka.cz

POPISNÉ STATISTIKY - OPAKOVÁNÍ

- jedna kvalitativní (kategoriální) proměnná
 - absolutní četnost, relativní četnost, sloupcový graf

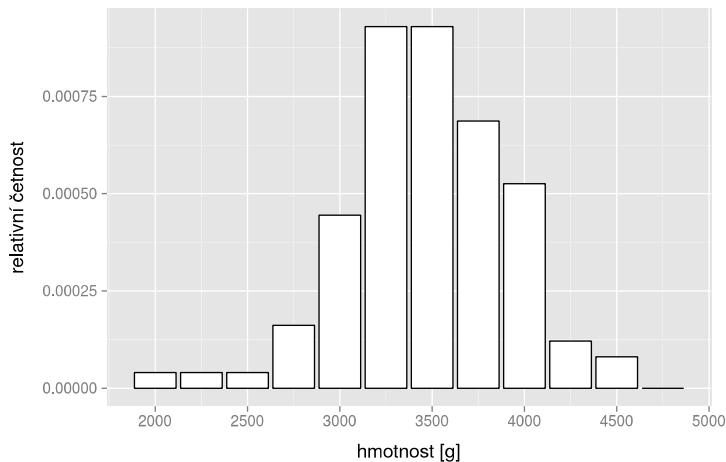
počet ok	absolutní četnost	relativní četnost
1	15	0.15
2	16	0.16
3	7	0.07
4	6	0.06
5	15	0.15
6	41	0.41
celkem	100	1.00



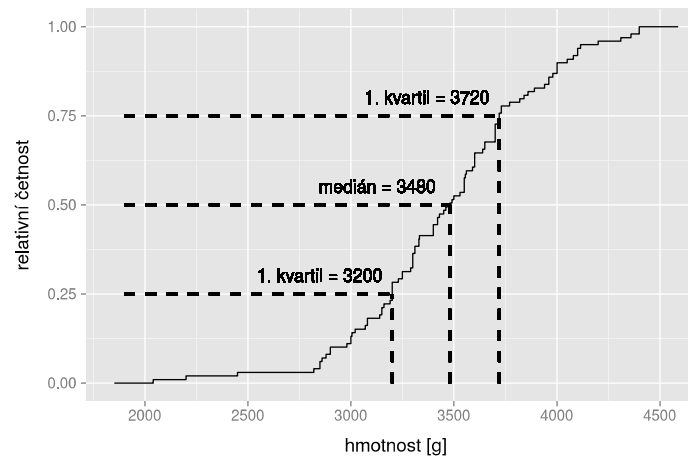
POPISNÉ STATISTIKY - OPAKOVÁNÍ

- jedna kvantitativní (spojitá) proměnná
 - poloha: minimum, maximum, medián, průměr, kvantily
 - variabilita: rozsah, IQR, rozptyl, SD
 - histogram, empirická distribuční funkce, boxplot

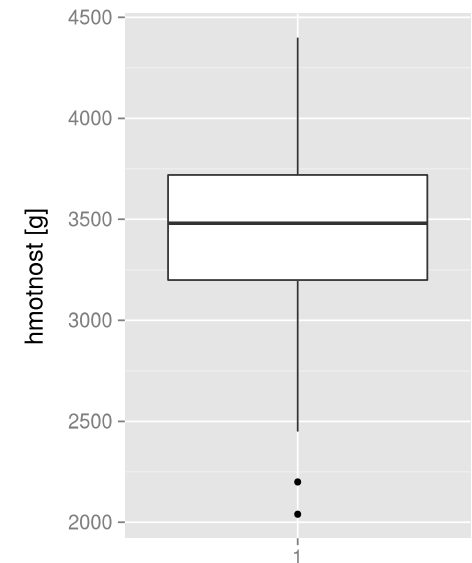
Experimentální rozložení porodní hmotnosti:
histogram



Experimentální rozložení porodní hmotnosti:
empirická distribuční funkce



Rozložení porodní hmotnosti:
boxplot



Z-SKÓRY

- normovaná veličina

$$z_i = \frac{x_i - \bar{x}}{s_x}, \quad i = 1, 2, \dots, n$$

- hodnoty z_i ztratily informaci o poloze a variabilitě:
 - průměr je vždy 0
 - rozptyl a SD jsou vždy 1
- nemění se přičtením konstanty ani násobením konstantou
- umožňují porovnatelnost → základ testování
- R: `scale(x)` nebo `(x - mean(x)) / sd(x)`

POPISNÉ STATISTIKY

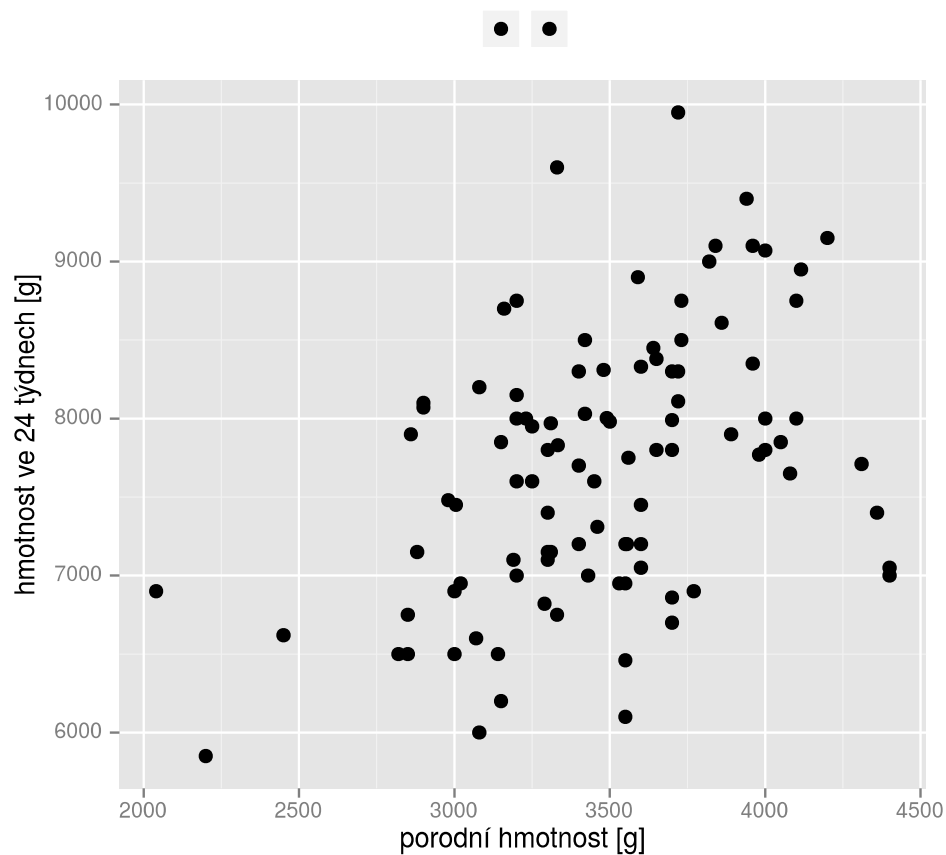
Dvě proměnné: popis vztahu, hledání (ne)závislosti

Podle typu proměnných

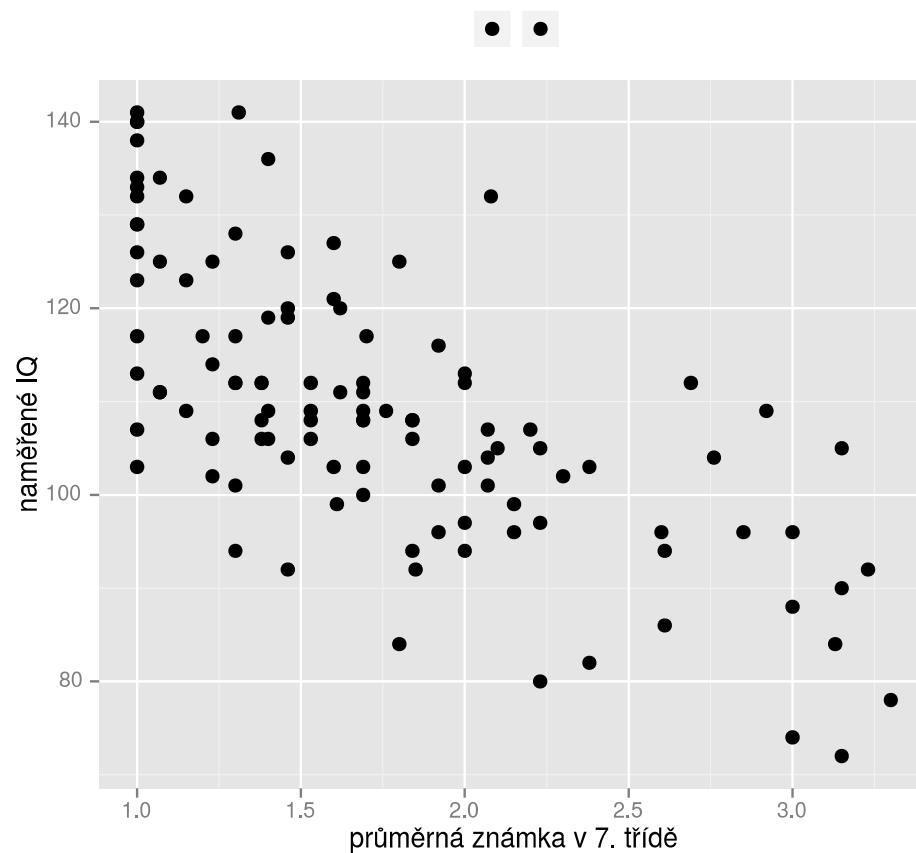
- kvantitativní a kvantitativní
 - bodový graf (scatterplot), korelační koeficient → regrese
- kvantitativní a kvalitativní
 - krabicový graf (boxplot), RSS → t-test, ANOVA, Wilcoxonův test, ...
- kvalitativní a kvalitativní
 - kontingenční tabulka, sloupcový graf
 - $\rightarrow \chi^2$ -test, Fisherův test

KVANITATIVNÍ A KVANTITATIVNÍ

Vztah hmotnosti v 6 měsících a porodní hmotnosti



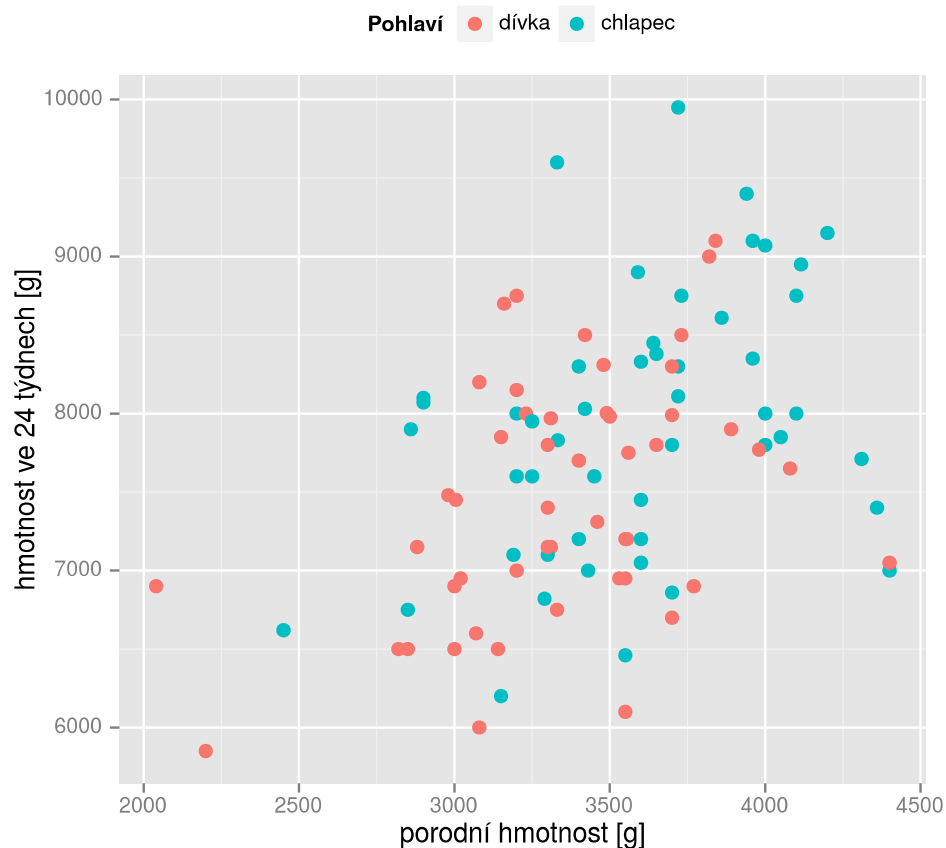
Vztah průměrné známky v 7. třídě a IQ



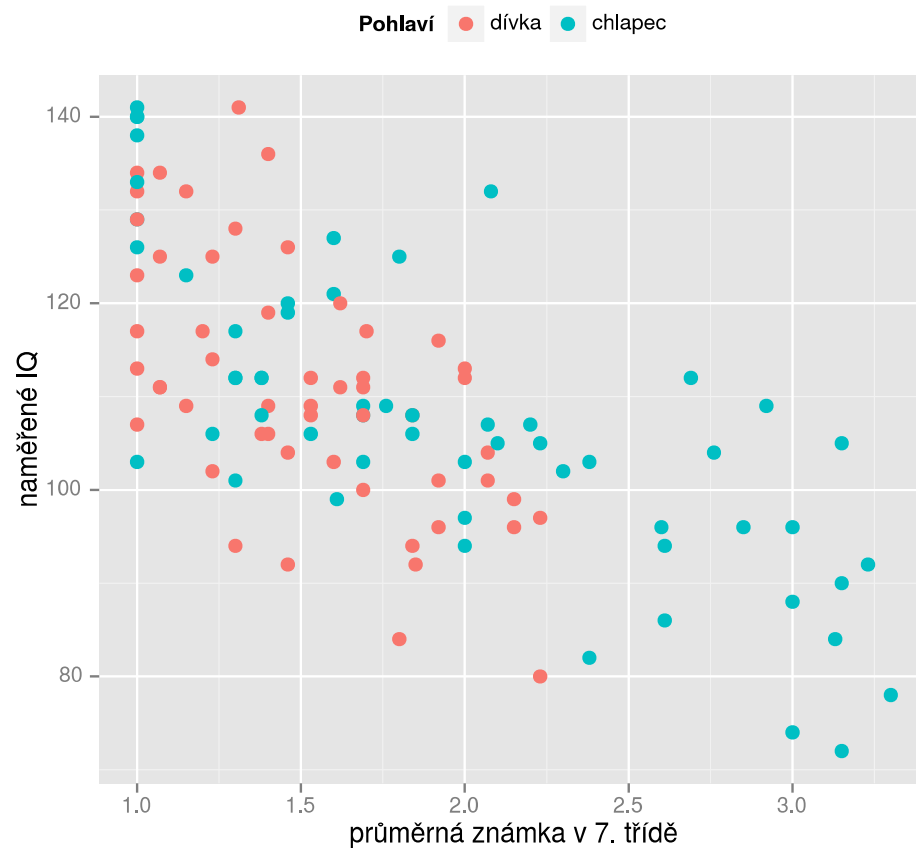
Data Kojeni.csv (vlevo) a data Iq3.csv (vpravo)

KVANITATIVNÍ A KVANTITATIVNÍ

Vztah hmotnosti v 6 měsících a porodní hmotnosti



Vztah průměrné známky v 7. třídě a IQ



pozn: pokud záleží na směru závislosti, pak vysvětlovanou (závislou) proměnnou dáme na osu y

VÝBĚROVÝ KORELAČNÍ KOEFICIENT

- **korelační koeficient** vyjadřuje sílu a směr **vzájemné** závislosti

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}, \quad \text{kde} \quad s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- **[cor(x,y)]**

- s_{xy} – výběrová **kovariance**

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

- pomocí z-skórů ($\Rightarrow$ nezávislost na poloze a měřítku)

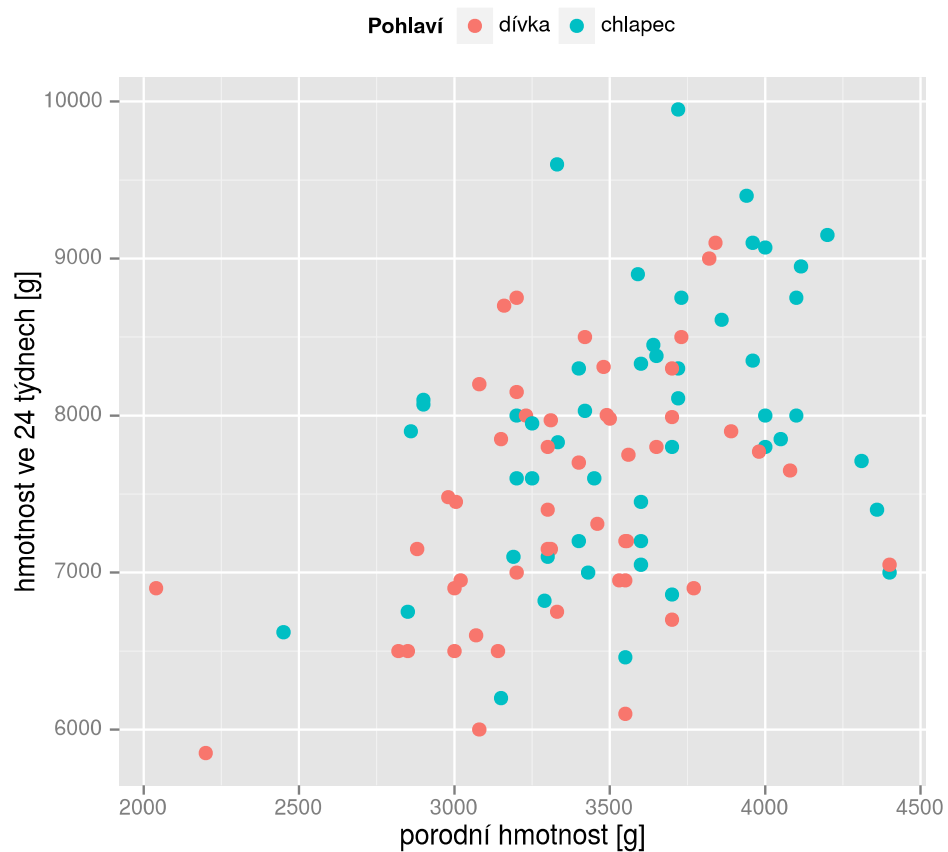
$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- pro $r_{xy} > 0$ s rostoucím x v průměru roste y
pro $r_{xy} < 0$ s rostoucím x v průměru klesá y
nezávislosti odpovídají hodnoty r_{xy} blízké nule

$$-1 \leq r_{xy} \leq 1$$

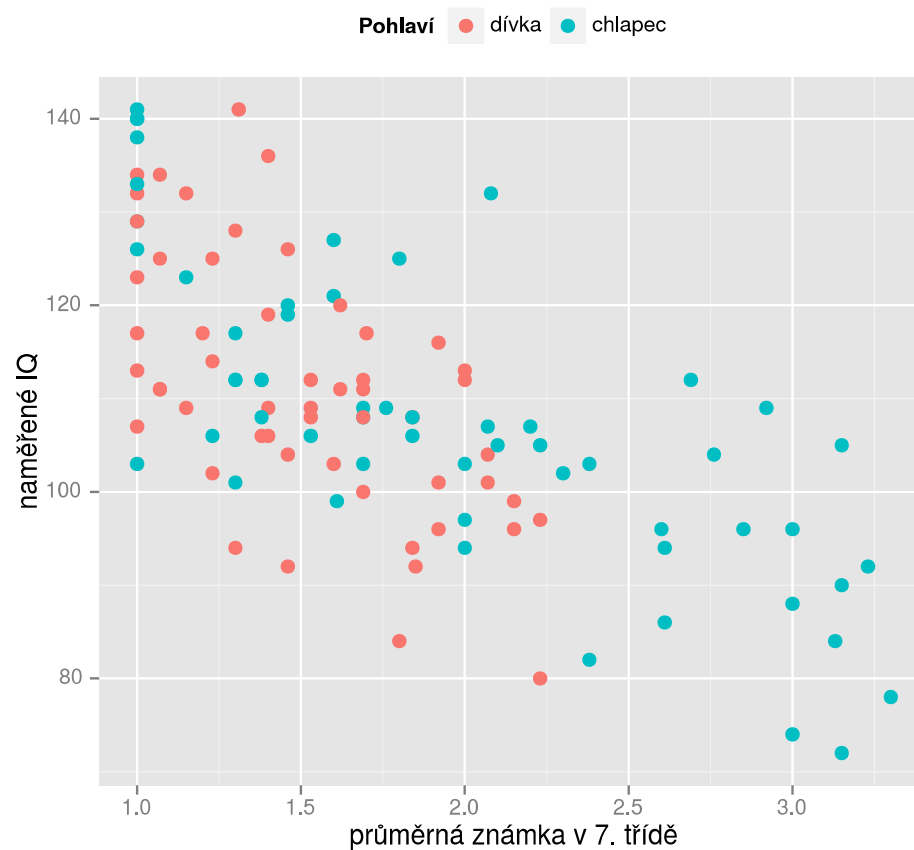
KVANITATIVNÍ A KVANTITATIVNÍ

Vztah hmotnosti v 6 měsících a porodní hmotnosti



$$r = 0.429$$

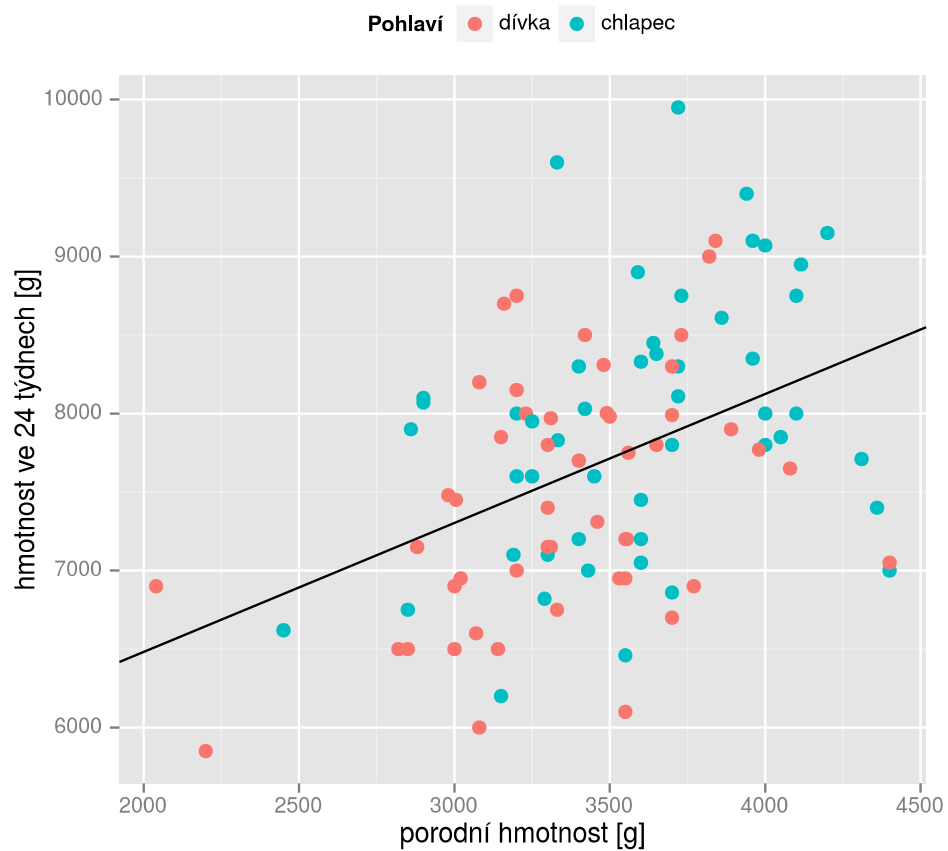
Vztah průměrné známky v 7. třídě a IQ



$$r = -0.688$$

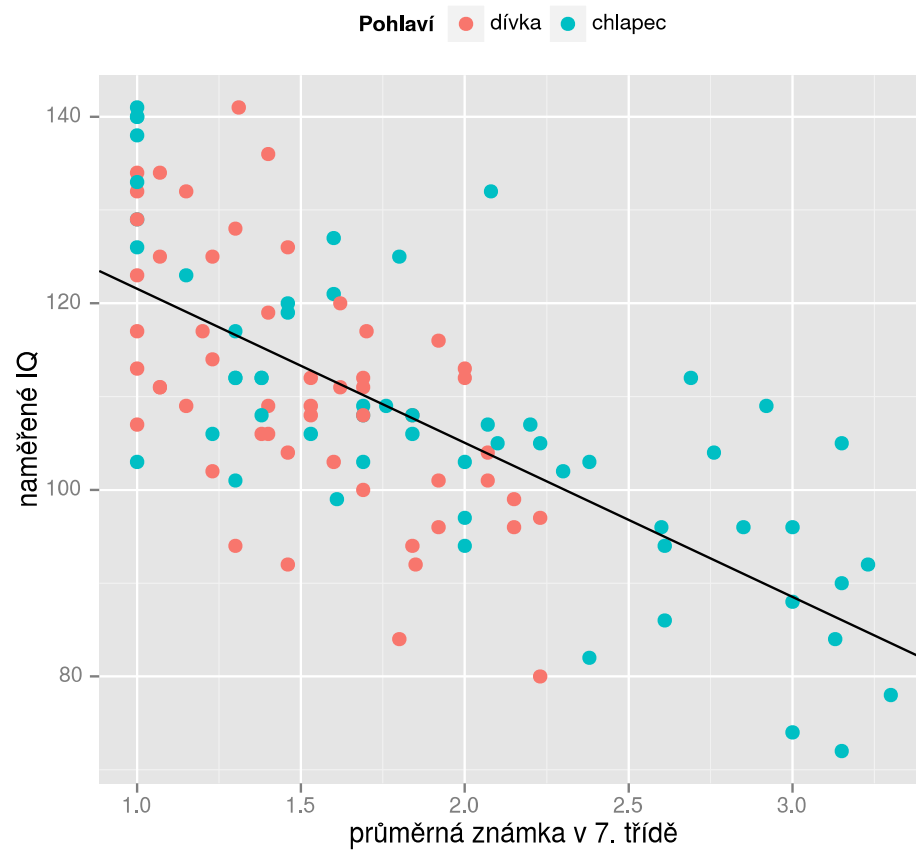
KVANITATIVNÍ A KVANTITATIVNÍ

Vztah hmotnosti v 6 měsících a porodní hmotnosti



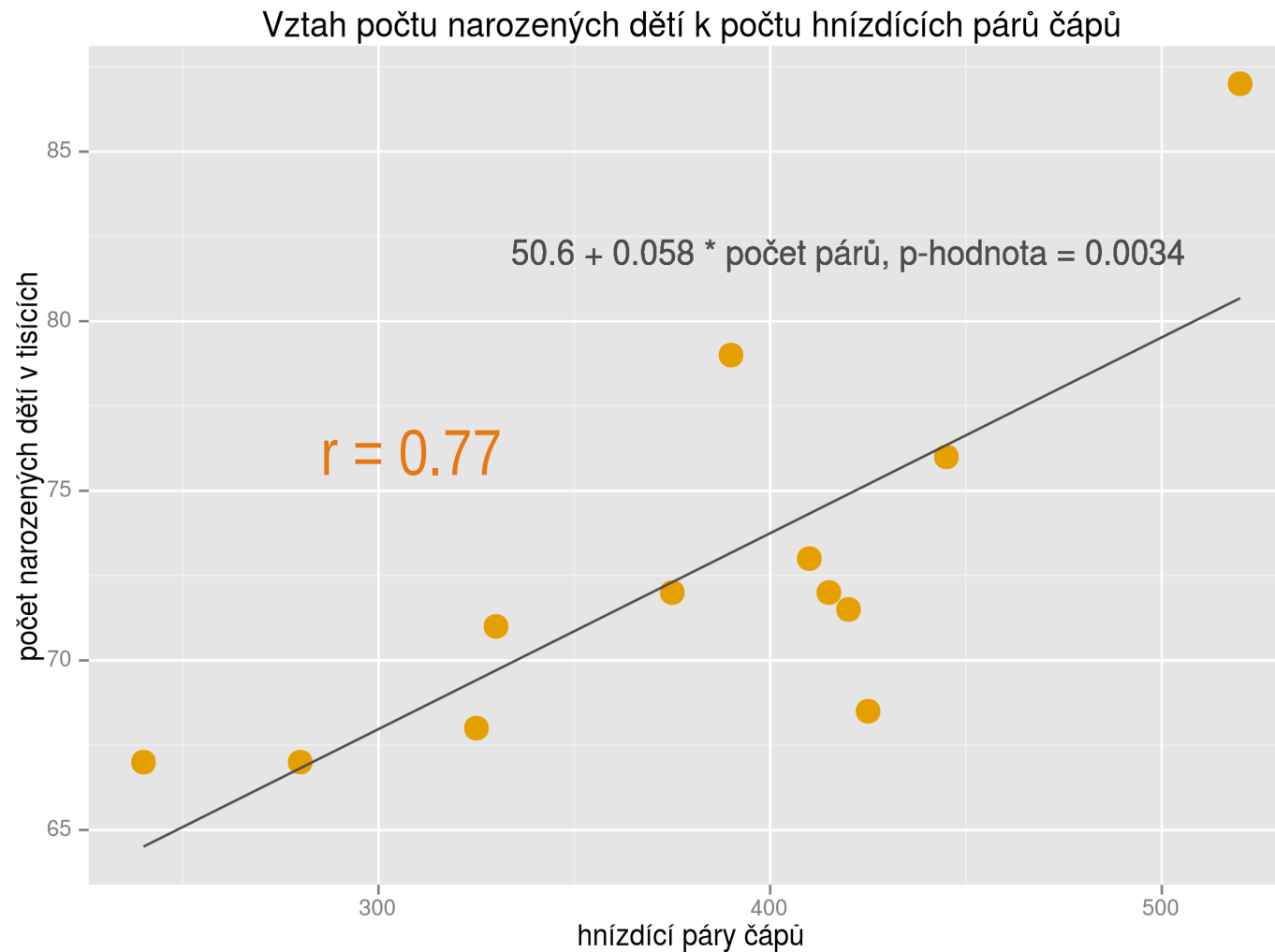
$$r = 0.429$$

Vztah průměrné známky v 7. třídě a IQ



$$r = -0.688$$

KORELACE vs. KAUZALITA

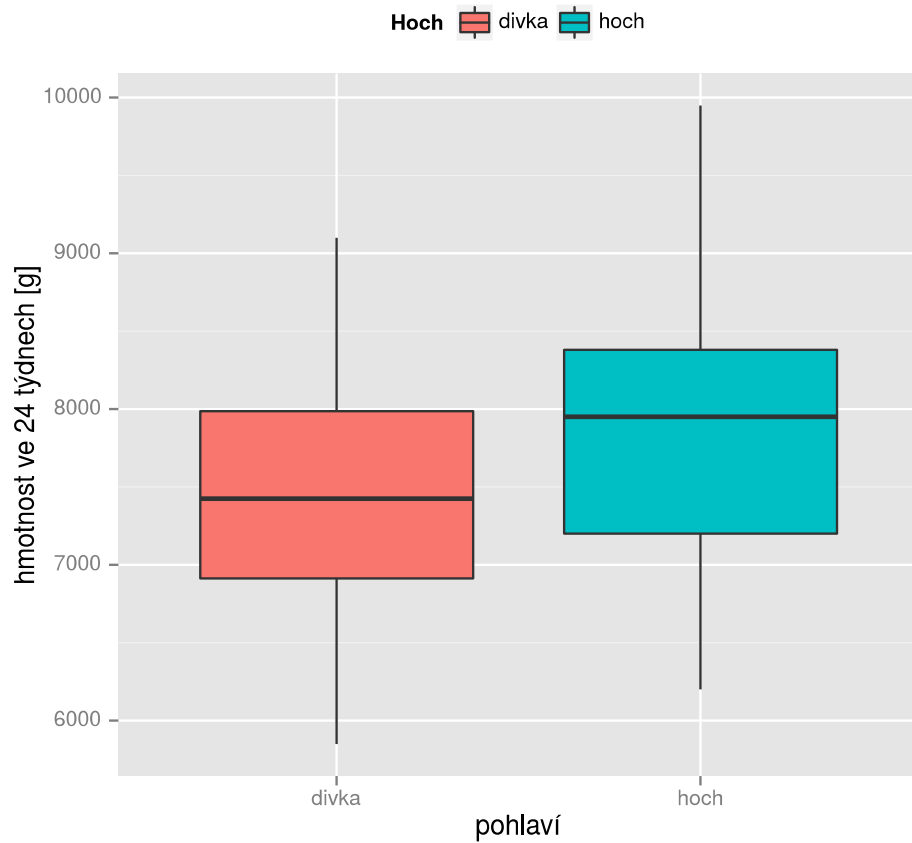


KVANITATIVNÍ A KVALITATIVNÍ

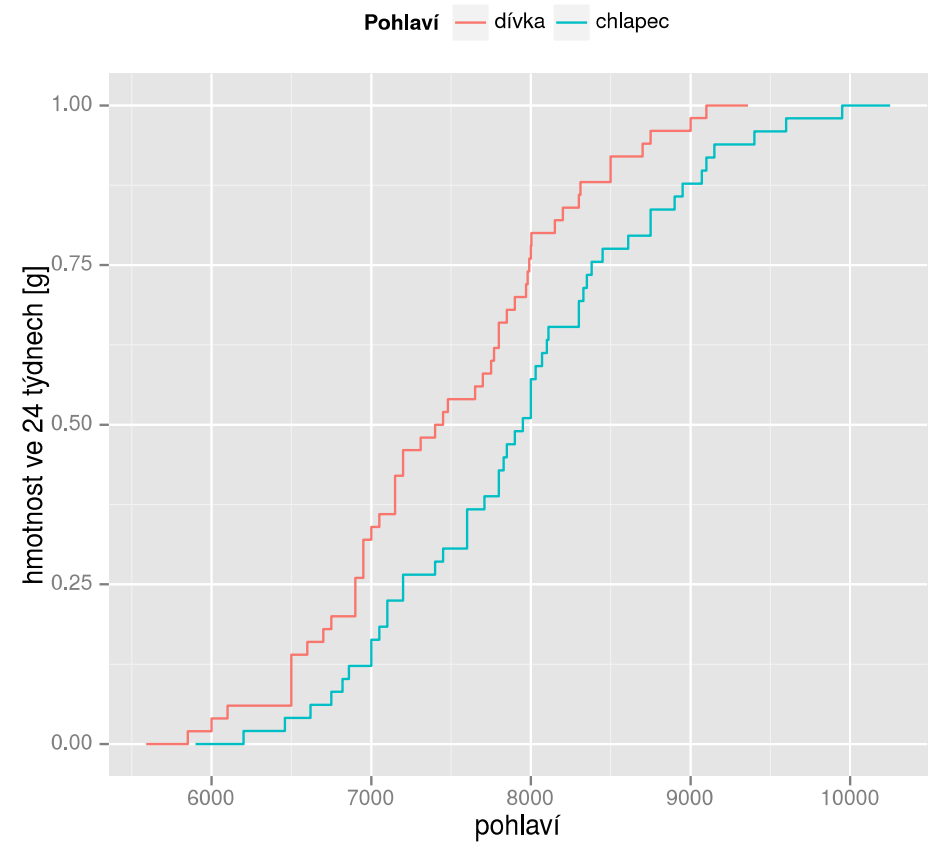
- máme několik kategorií (skupin) a v nich mají jednotlivci naměřené kvantitativní hodnoty
- srovnáváme soubory dat v těchto skupinách
- lze chápat jako (ne)závislost kvantitativní na kvalitativní
- zobrazení boxplotem nebo empirické distribuční funkce
- nezávislost pokud krabice podobně umístěné

KVANITATIVNÍ A KVALITATIVNÍ

Vztah hmotnosti v 6 měsících a pohlaví



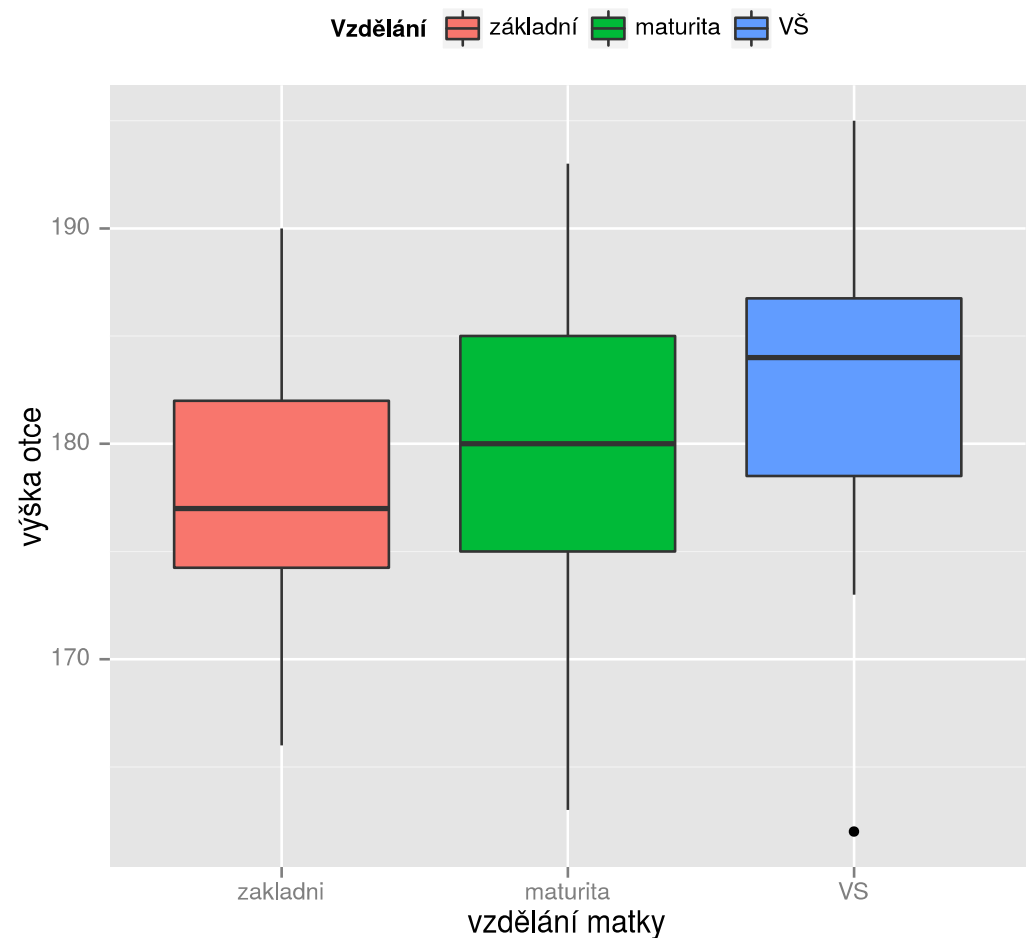
Vztah hmotnosti v 6 měsících a pohlaví



KVANITATIVNÍ A KVALITATIVNÍ

- vidíme dokonce i jistý trend
- lze posuzovat jak různost výšky otců mezi skupinami navzájem, tak možnost, že každý stupeň vzdělání výšku někam posune výš

Vztah výšky otce a vzdělání matky

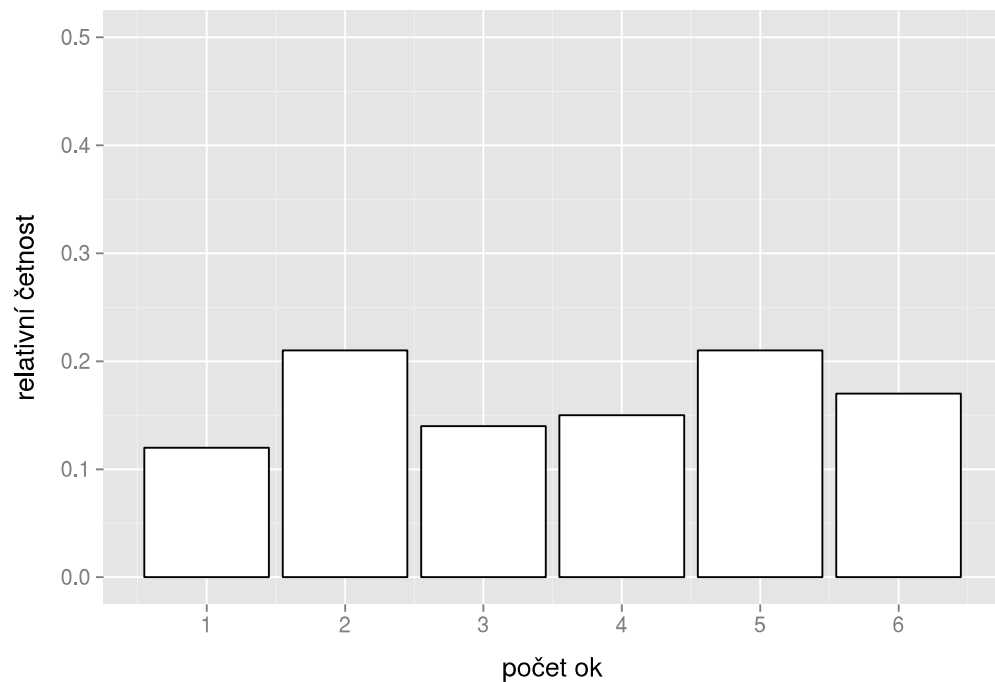


NÁHODNÁ VELIČINA

PLATONOVA JESKYNĚ A STÍN NA ZDI

Naházeli jsme si 100 výsledků.

Experimentální hod kostkou - histogram

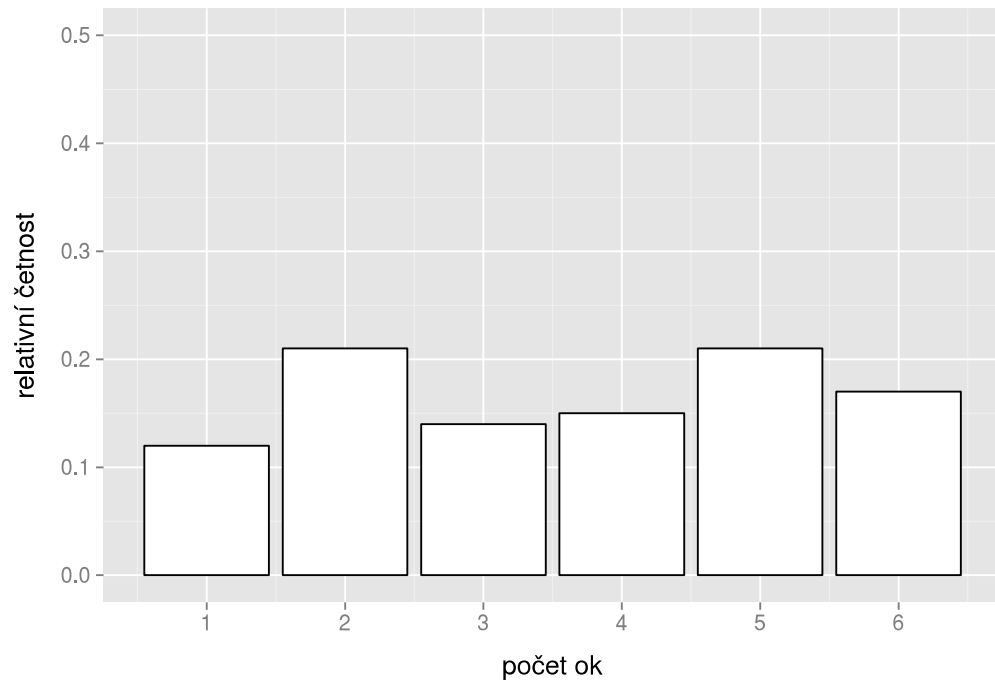


O kostce ale už principiálně něco předem předpokládáme.

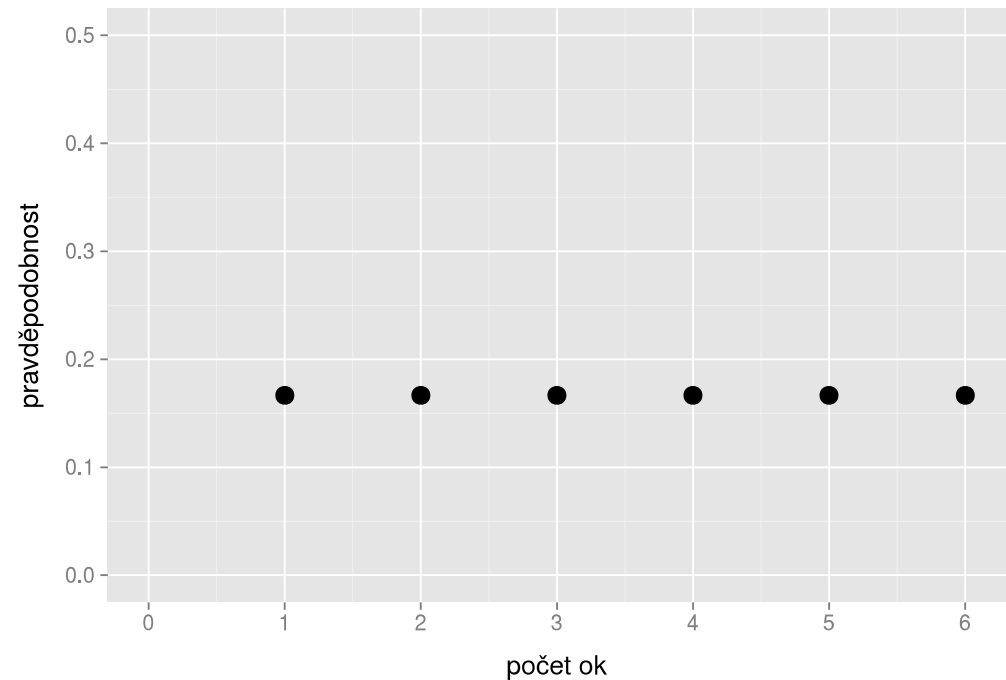
PLATONOVA JESKYNĚ A STÍN NA ZDI

V hlavě chaos měření přibližujeme teoretickou „křivkou“

Experimentální hod kostkou - histogram



Teoretická hustota hodu kostkou



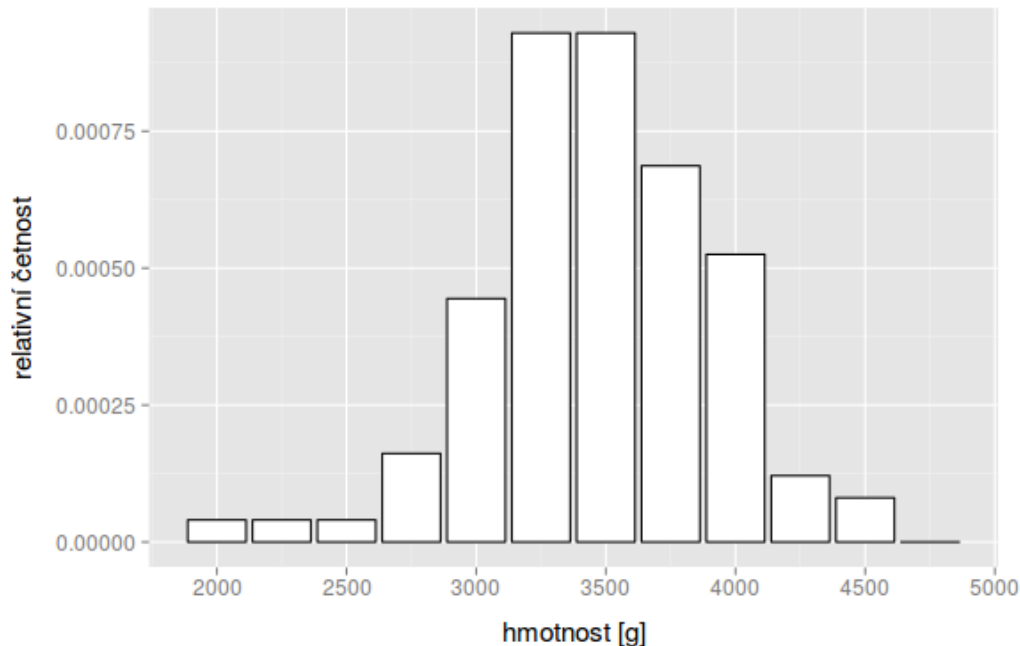
Předpokládáme rovnoměrnost rozložení (nefalešná kostka)

Máme pro to i logické odůvodnění, proč by to tak mělo být.

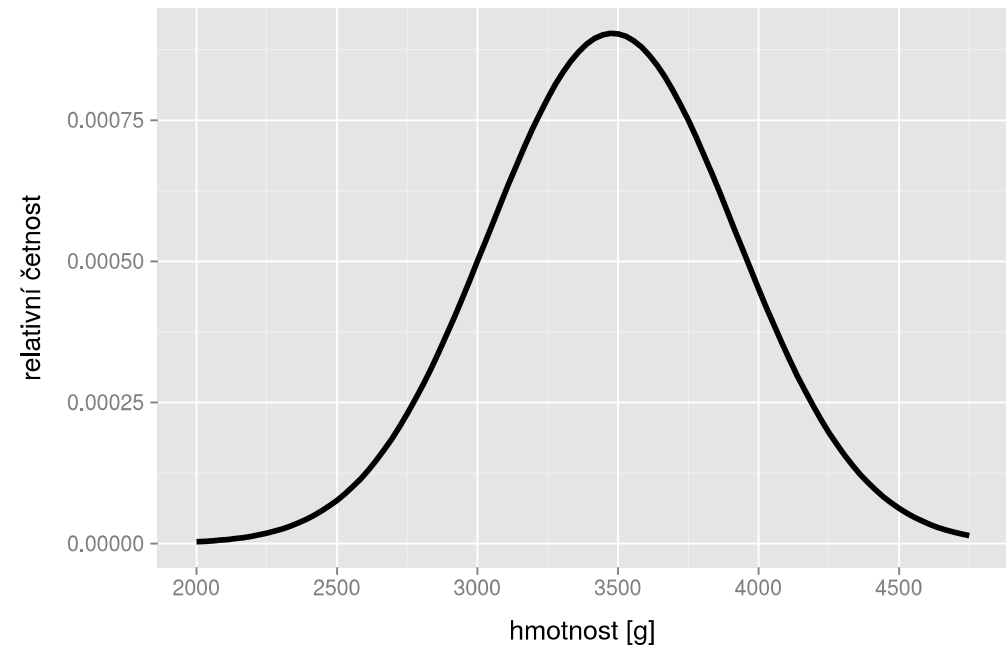
PLATONOVA JESKYNĚ A STÍN NA ZDI

Chaos měření přiblížíme teoretickou křivkou

Experimentální rozložení porodní hmotnosti:
histogram



Teoretické rozložení porodní hmotnosti:
hustota normálního rozdělení



Předpokládáme normální rozdělení, ale nemáme pro to žádný teoretický důvod.

NÁHODNÁ VELIČINA

- to, co máme naměřeno, vnímáme jako **konkrétní realizaci** nějaké **skryté vlastnosti** – říkáme jí **náhodná veličina**
- příklad náhodné veličiny
 - číslo, které padne na kostce
 - pohlaví narozeného dítěte
 - porodní hmotnost narozeného dítěte
- teoretický model náhodné veličiny slouží k
 - srozumitelnému zjednodušenému popisu možného principu
 - k tvorbě testů, modelů a analytických technik
 - bez zjednodušení se nikam nehneme

NÁHODNÁ VELIČINA

Náhodná veličina může nabývat různých hodnot

- hod kostkou: 1,2,3,4,5,6
- pohlaví novorozence: dívka, chlapec, nerozhodnutelné/nestandardní
- hmotnost novorozence: cca 400g až 10kg

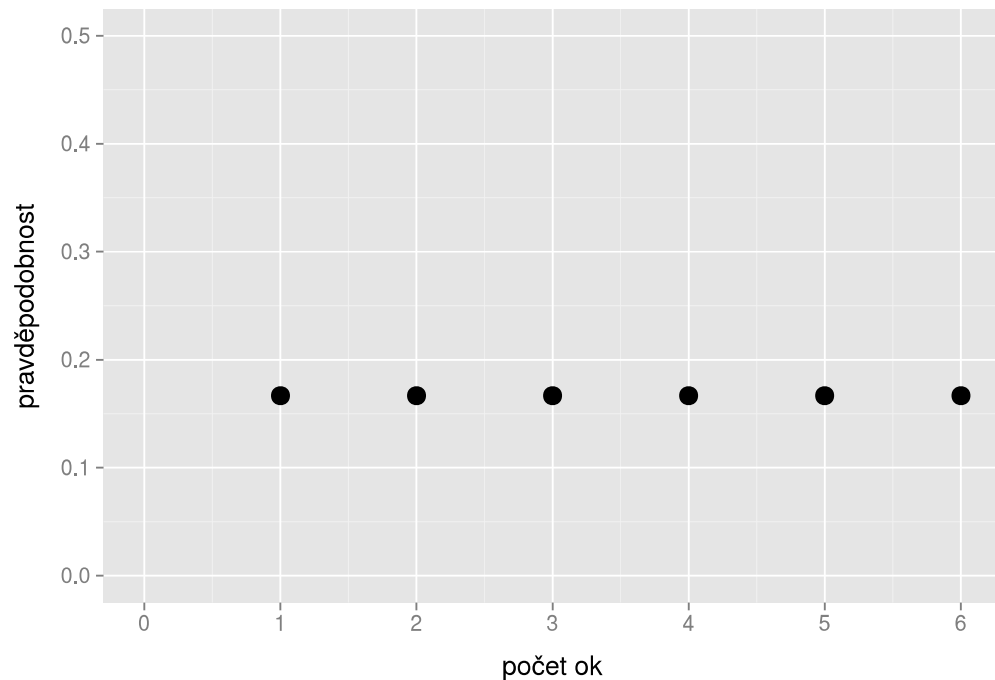
Teoretická pravděpodobnost výskytu konkrétní realizace:

- hod nefalešnou kostkou: $1/6$
- pohlaví novorozence
 - klasický model $1/2$ a $1/2$
 - korigovaný model: něco pod $1/2$, něco nad $1/2$, cca $1/1000$
- hmotnost novorozence:
 - hodně malé a hodně velké hmotnosti se vyskytnou málo často až výjimečně
 - nejčastěji se narodí dítě s hmotností cca 2.5 kg až 4.5 kg

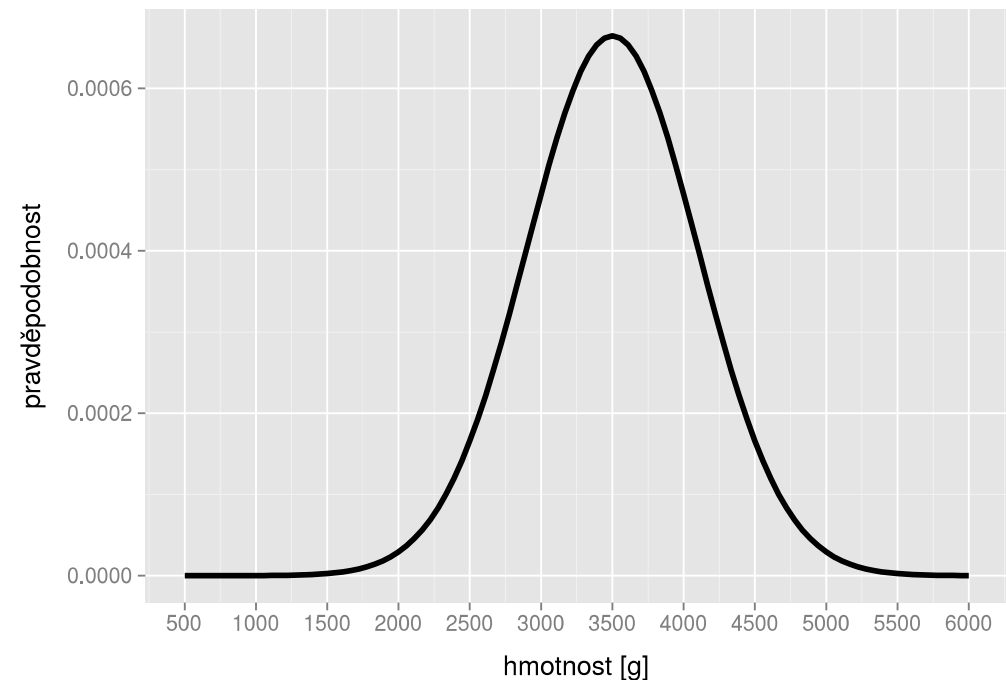
HUSTOTA PRAVDĚPODOBNOSTI

Jak pravděpodobný je výskyt konkrétní realizace?

Teoretická hustota hodu kostkou



Teoretická hustota porodní hmotnosti

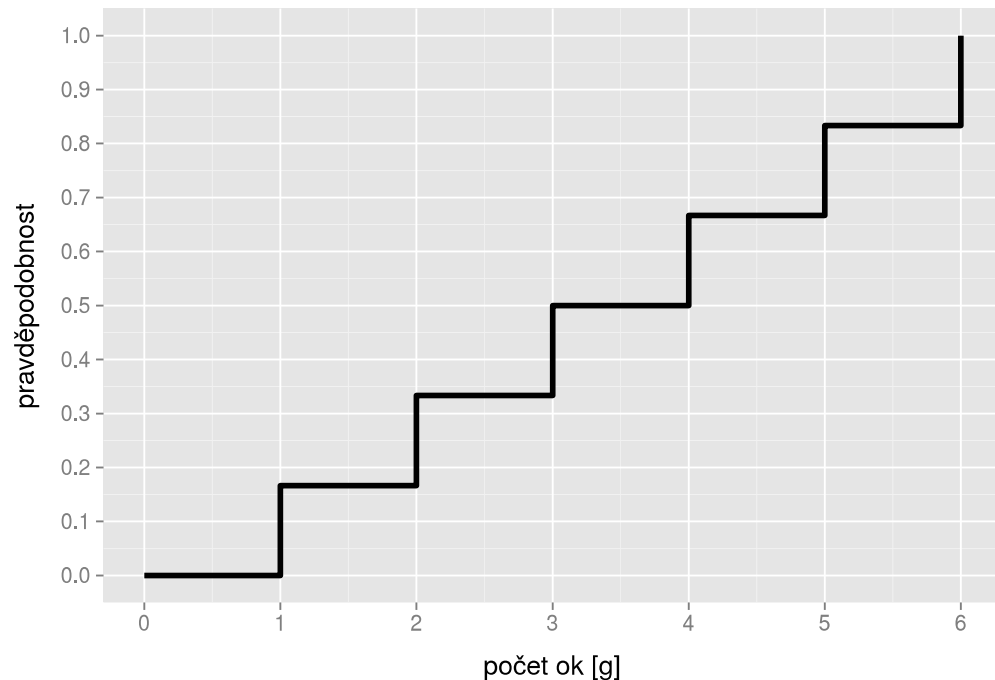


U kostky teoreticky odvozujeme, u hmotnosti tvar předpokládáme

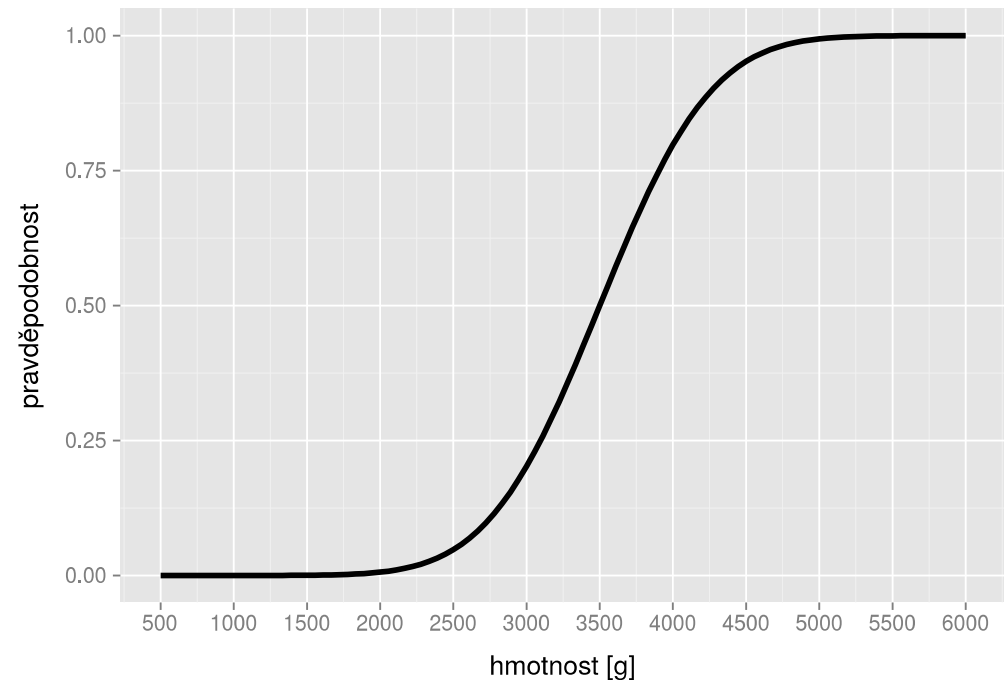
DISTRIBUČNÍ FUNKCE

Jak pravděpodobné je, že konkrétní realizace bude menší (nebo rovna) než nějaká daná hodnota?

Teoretická distribuční funkce hodu kostkou



Teoretická distribuční funkce porodní hmotnosti

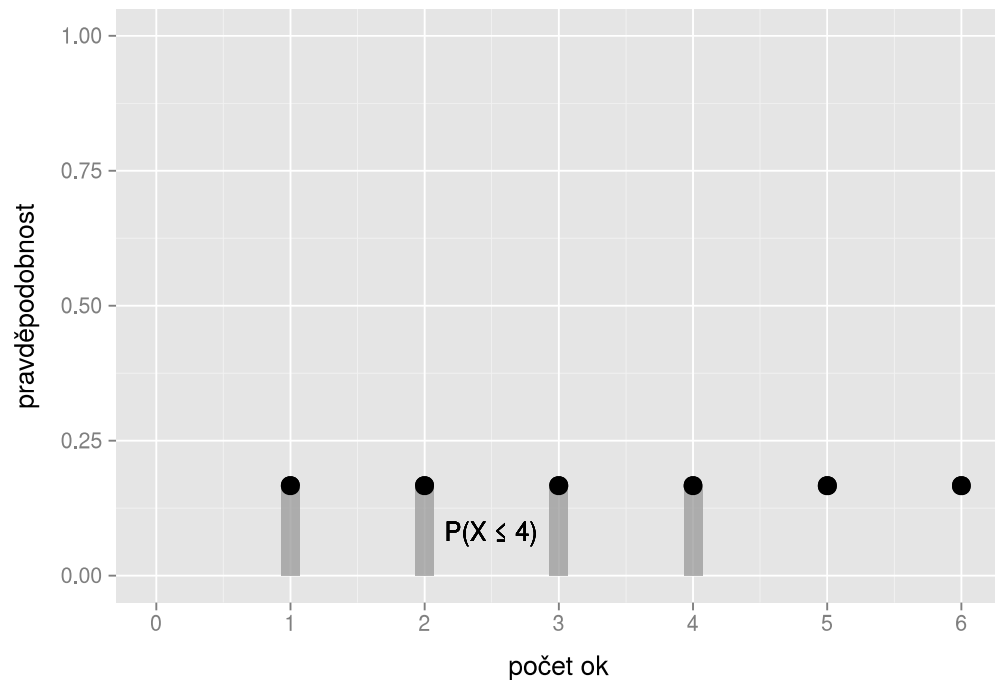


U kostky teoreticky odvozujeme, u hmotnosti předpokládáme

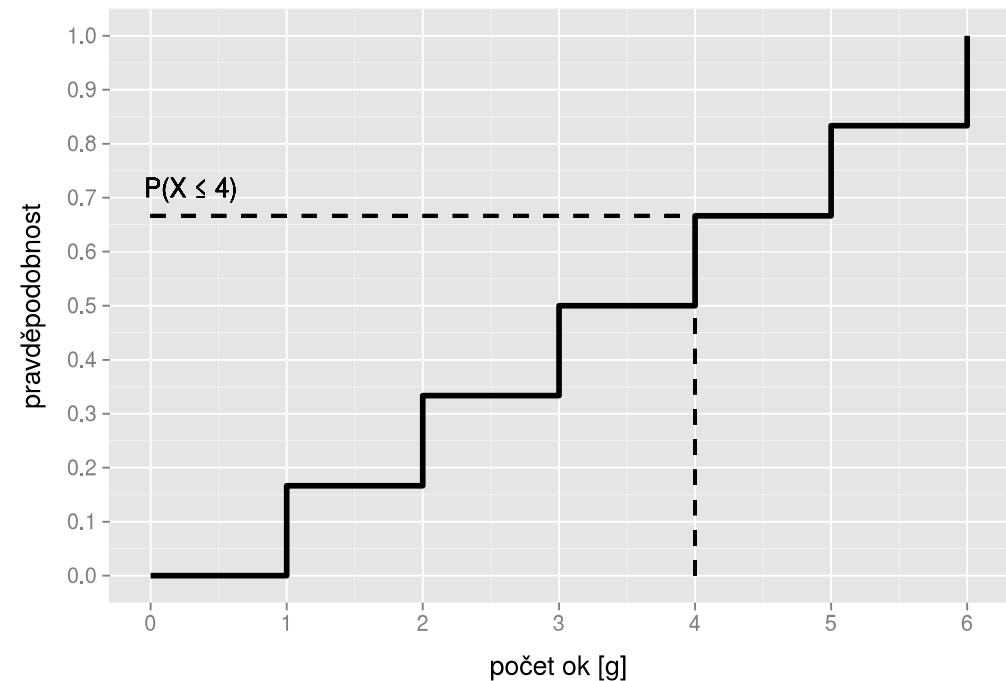
DISTRIBUČNÍ FUNKCE

Jak pravděpodobné je, že padne cokoli menší nebo rovno 4?

Teoretická hustota hodu kostkou



Teoretická distribuční funkce hodu kostkou

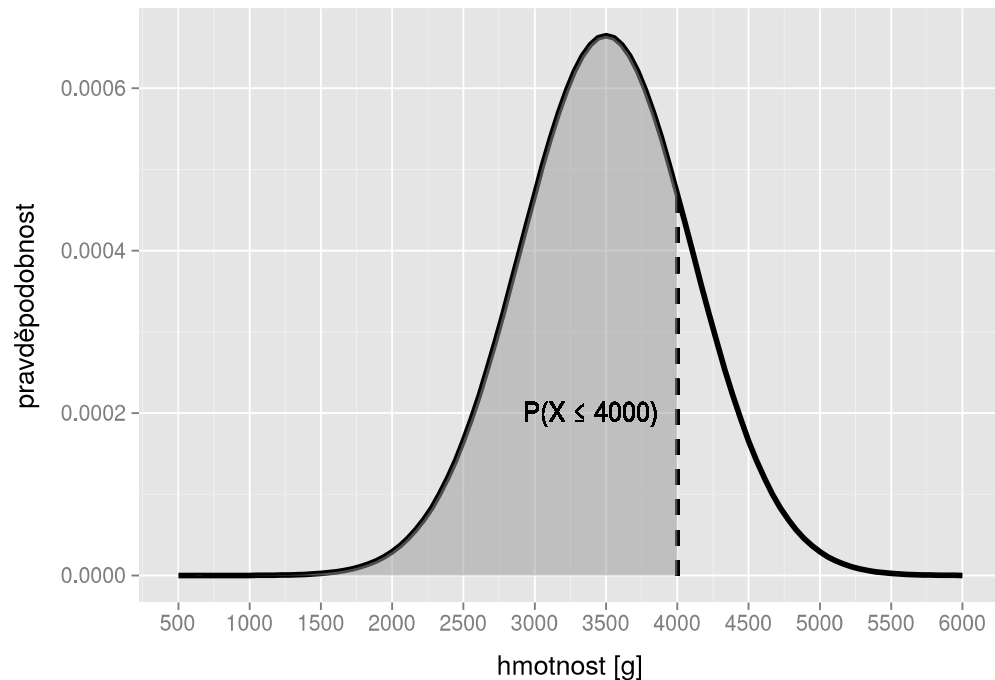


Výsledek čteme přímo z grafu distribuční funkce: $p = 0.67$

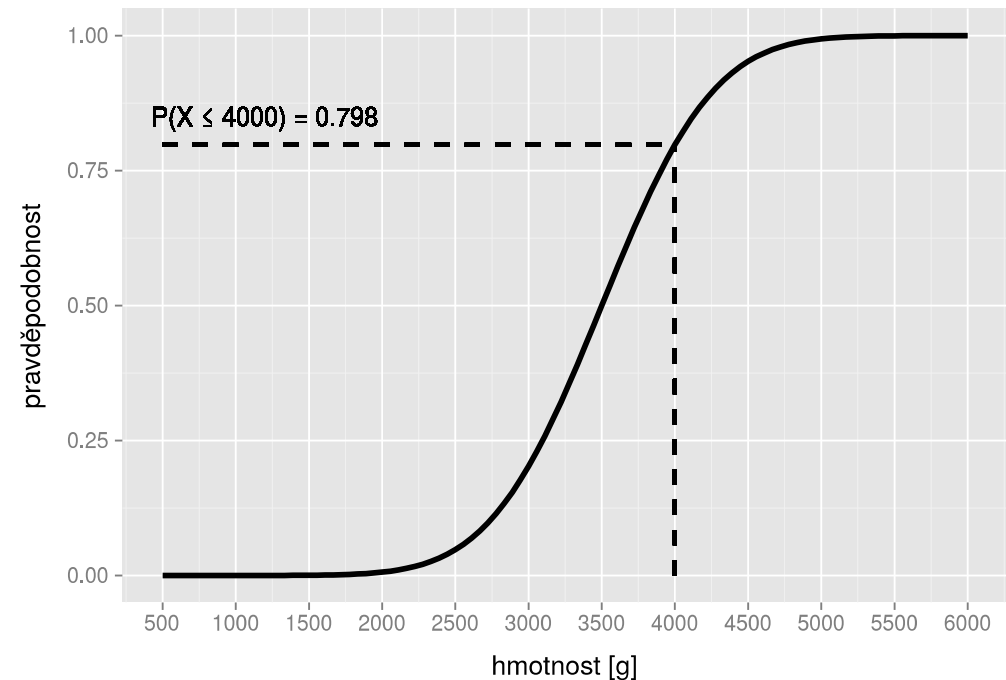
DISTRIBUČNÍ FUNKCE

Jak pravděpodobné je, že se narodí dítě s hmotností 4000g a méně?

Teoretická hustota porodní hmotnosti



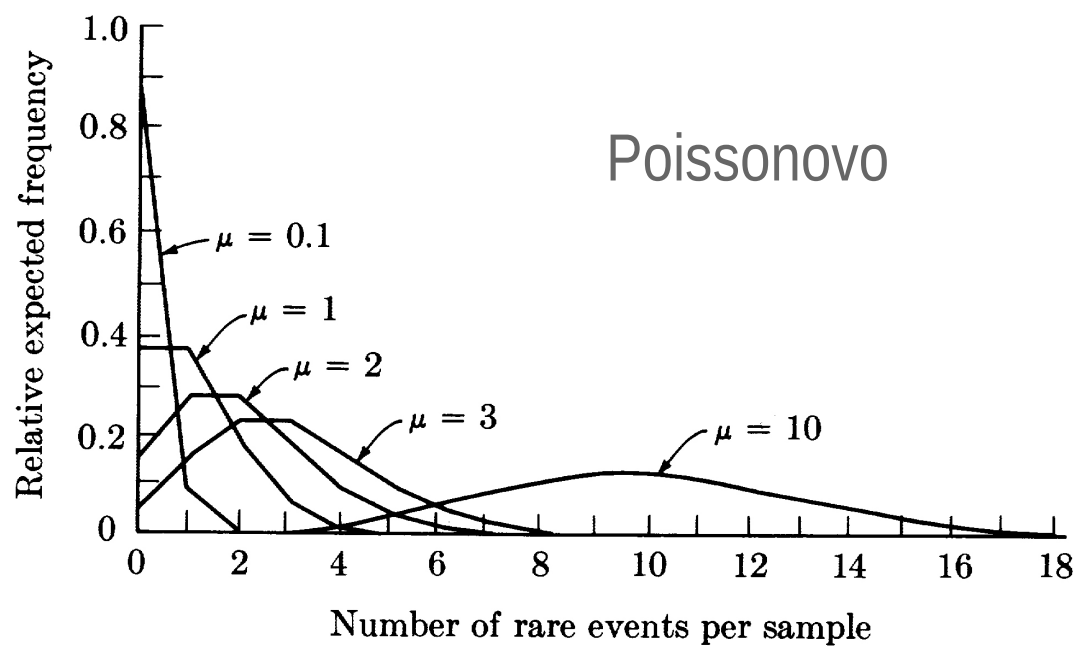
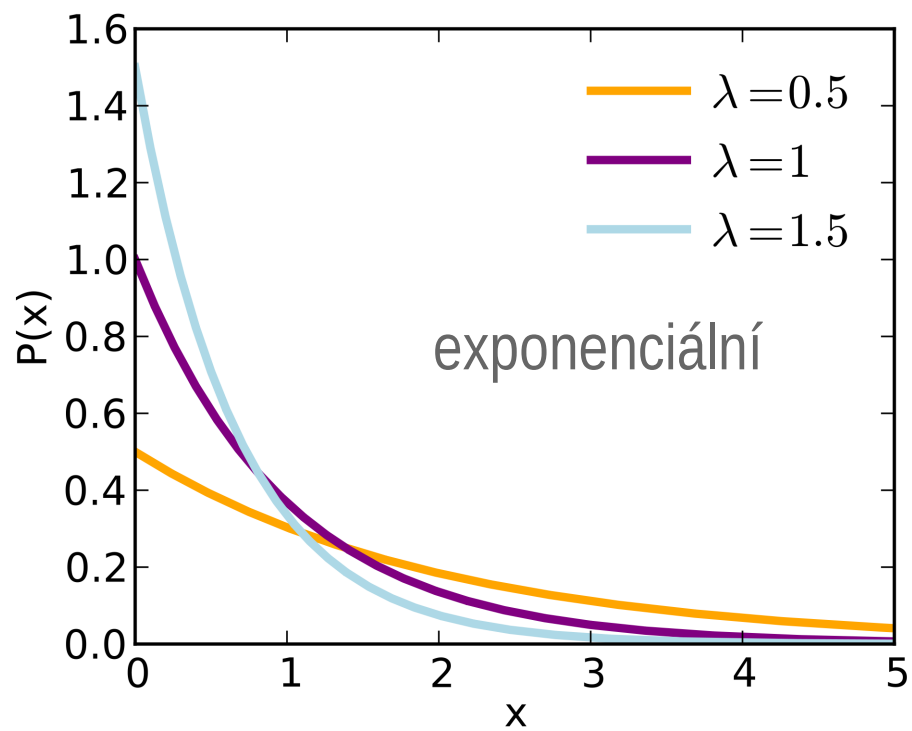
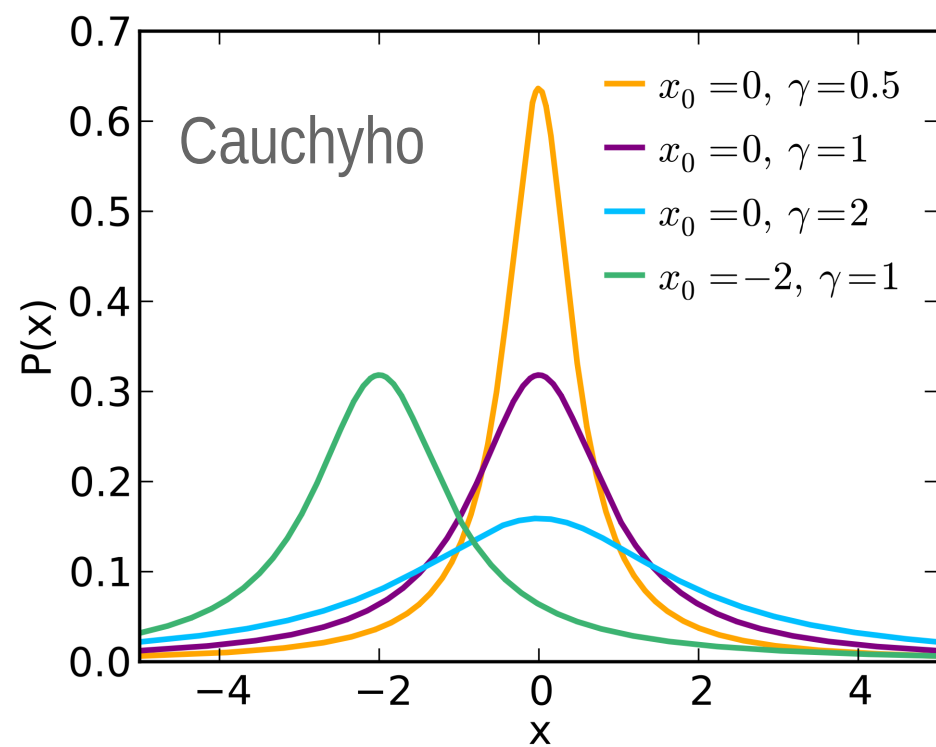
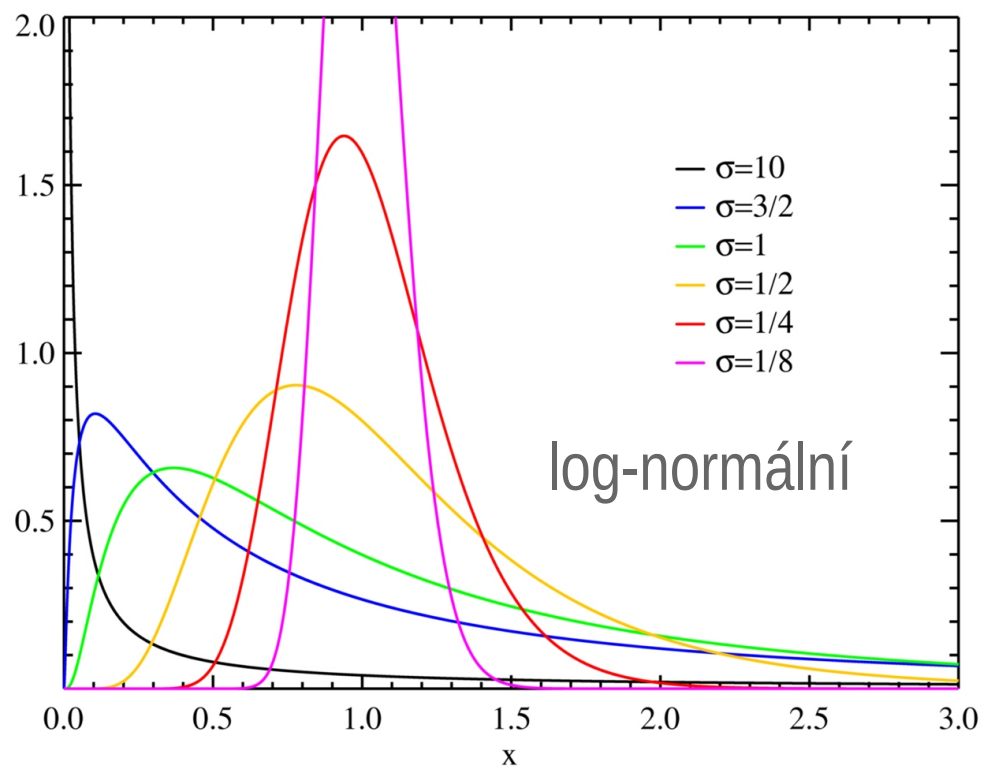
Teoretická distribuční funkce porodní hmotnosti



Výsledek čteme přímo z grafu distribuční funkce: $p = 0.80$

PROČ UVAŽOVAT TEORETICKÉ ROZDĚLENÍ

- Převedení do funkce umožňuje veličinu (princip) přesně popsat
 - interval hodnot (výčet, $(0, \infty)$, $(-\infty, \infty)$, ...)
 - střední hodnota $E(X)$
 - rozptyl $\text{Var}(X)$
 - šikmost, špičatost, ...
 - „tvar“ - normální, rovnoměrné, lognormální, gamma, exponenciální, Poissonovo, ...
 - směs různých rozdělení, asymptotické,
- hledáme, který **tvar** nejlépe sedí a pak odhadujeme jeho **parametry** (z dat)
- s matematickou funkcí se dají konstruovat testy a modely



TEORETICKÝ MODEL A DATA

- pro řadu rozdělení je **aritmetický průměr** dat vhodným odhadem **střední hodnoty**
- podobně je **výběrový rozptyl** vhodným odhadem teoretické **variance**
- u hodu kostkou je **relativní četnost** odhadem **pravděpodobnosti** výskytu dané hodnoty
- výběrový korelační koeficient je nejlepším odhadem teoretické kovariance u normálního rozdělení
- ...

A CO PŘÍŠTĚ?

- To, čím jsme popisovali konkrétní data, můžeme použít jako první/jediný dostupný odhad skrytých parametrů.
- Je to ale zatíženo nějakou nejistotou. S tím se musíme nějak vypořádat.
- To budeme řešit příště.
- Dostaneme se tím ke konstrukci jednoduchých statistických testů.

DĚKUJI ZA POZORNOST

www.biostatisticka.cz