PSY117

Statistická analýza dat v psychologii

Přednáška 8 2019

Statistické usuzování, odhady

Věci, které můžeme přímo pozorovat, jsou téměř vždy pouze vzorky.

Alfred North Whitehead

Barevná srdíčka kolegyně Michalčákové

- Jaký je podíl barevných srdíček v balení?

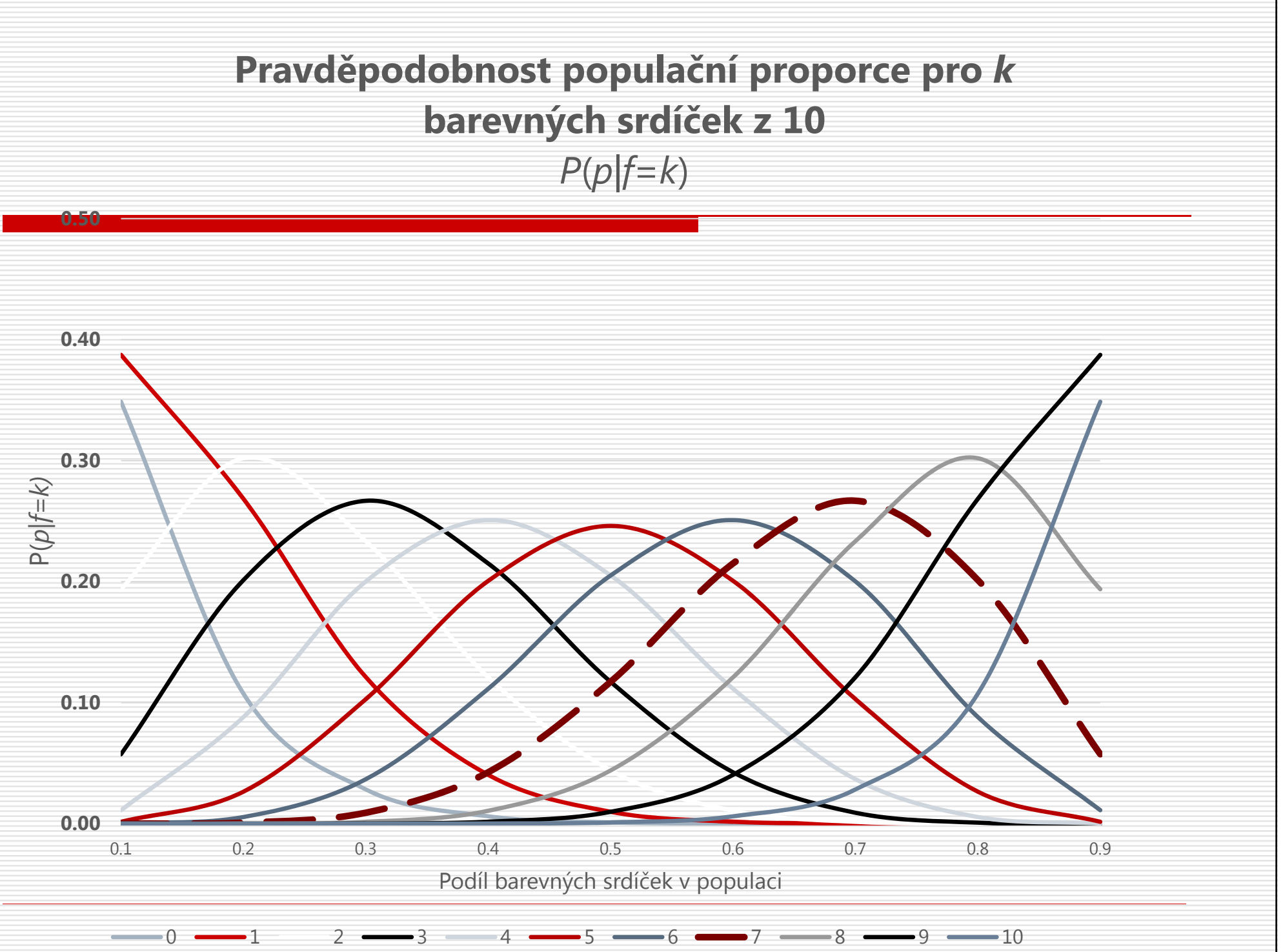
Vylosovali jsme z populace 1 vzorek 10 srdíček
Našli jsme k barevných srdíček
Pro jakou relativní četnost p je $P(p|f=k)$ nejvyšší?

	Podíl barevných srdíček v populaci								
Počet barevných z 10	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
0	0,349	0,107	0,028	0,006	0,001	0,000	0,000	0,000	0,000
1	0,387	0,268	0,121	0,040	0,010	0,002	0,000	0,000	0,000
2	0,194	0,302	0,233	0,121	0,044	0,011	0,001	0,000	0,000
3	0,057	0,201	0,267	0,215	0,117	0,042	0,009	0,001	0,000
4	0,011	0,088	0,200	0,251	0,205	0,111	0,037	0,006	0,000
5	0,001	0,026	0,103	0,201	0,246	0,201	0,103	0,026	0,001
6	0,000	0,006	0,037	0,111	0,205	0,251	0,200	0,088	0,011
7	0,000	0,001	0,009	0,042	0,117	0,215	0,267	0,201	0,057
8	0,000	0,000	0,001	0,011	0,044	0,121	0,233	0,302	0,194
9	0,000	0,000	0,000	0,002	0,010	0,040	0,121	0,268	0,387
10	0,000	0,000	0,000	0,000	0,001	0,006	0,028	0,107	0,349

Simulace binomického rozložení

Binomické rozložení

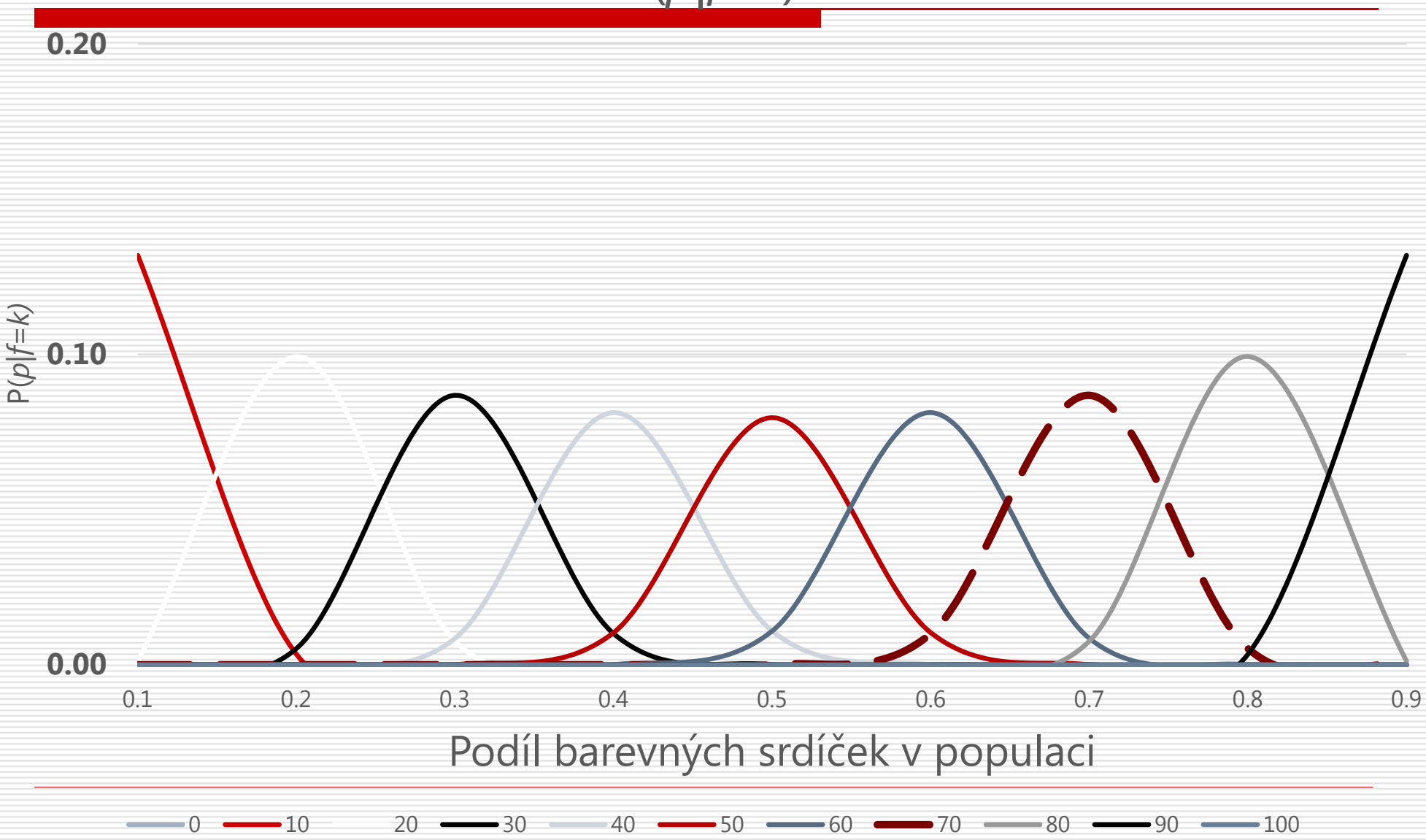
- Rozložení pravděpodobnosti k úspěchů z n pokusů při konstantní pravděpodobnosti úspěchu v jednotlivých pokusech p
- $P_{X=k} = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$
  - Například, jaká je pravděpodobnost, že si nevytáhneme žádné barevné srdíčko ($k = 0$) z 10 ($n = 10$), když je v populaci 10 % barevných srdíček ($p = 0{,}1$)?
  - $P_{X=0} = \binom{10}{0} 0{,}1^0 (0{,}9)^{10-0} = \frac{10!}{0!(10)!} 0{,}1^0 (0{,}9)^{10} = 0{,}3487$
- Pozn. k i n jsou diskrétní, $k \in \{0,1,2,\ldots,n\}$, p je spojitá


Pravděpodobnost populační proporce pro k barevných srdíček z 10
P(p|f=k)
0.50
0.40
0.30
0.20
0.10
0.00
P(p|f=k)
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Podíl barevných srdíček v populaci
0
1
2
3
4
5
6
7
8
9
10

- Nejlepší je hádat, že v populaci je takový podíl barevných srdíček, jaký je v našem vzorku
  - Nejlepší = takový dohad, který maximalizuje pravděpodobnost, že naše data vznikla náhodným výběrem z populace s daným podílem barevných srdíček (odhad maximální věrohodností, ML)
  - **Platí pro všechny slušné statistiky**

- Ale i jiné (blízké) podíly barevných srdíček jsou podobně pravděpodobné – jak moc si můžeme být jistí?

Pravděpodobnost populační proporce pro *k* barevných srdíček ze 100



S větším vzorkem se interval pravděpodobných hodnot podílu barevných srdíček v populaci zužuje.

$P(p|k)$

Obraťme nyní otázku:

Je-li podíl barevných srdíček p, jaké výsledky našeho výzkumu můžeme očekávat (k)?

$P(k|p)$

Pojďme na to tentokrát hrubou silou

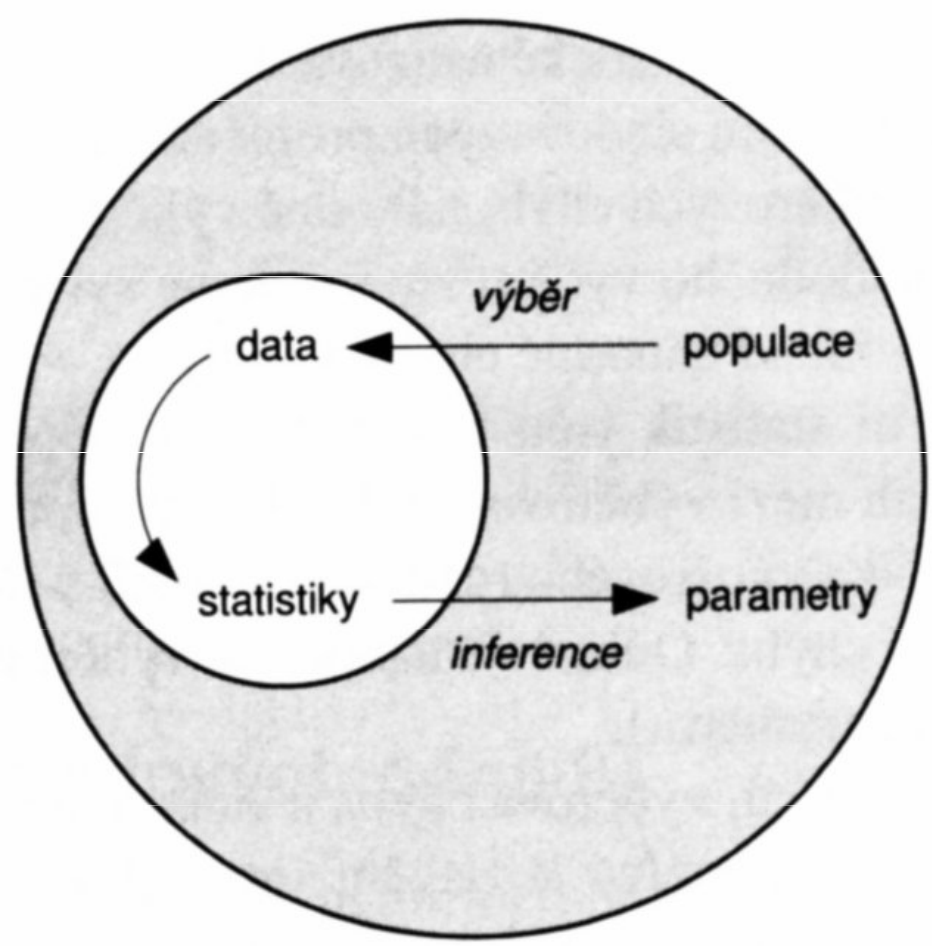
Jaké je empirické rozložení odhadů *p* pomocí *k* opakovanými výzkumy (=opakovaným losováním vzorků)?

Simulace binomického rozložení

- Můžeme také říci, že nejlepší odhad podílu barevných srdíček je **průměr** podílu barevných srdíček ve více náhodných vzorcích z populace srdíček
- Variabilita jednotlivých podílů nás upozorňuje na možnou nepřesnost odhadů

- I to platí pro většinu „slušných“ statistik – opakování experimentu a zprůměrování statistik z jednotlivých vzorků vede ke zpřesnění odhadu.

Výběr – od deskripce k indukci



- Deskripce dat, odhad parametrů
- Usuzování = inference = indukce

- Počítá se s **náhodným výběrem**
  - tj. výběr jedince splňuje podmínky náhodného pokusu
  - není-li výběr v pravém slova smyslu náhodný, uvažujeme, v čem se p-dobně liší od náhodného

AJ: statistical description, inference, population, sample, data, statistics, inference, parameters, random sample (sampling)

Statistiky a parametry

- Na vzorku (datech) počítáme **statistiky**
- Hodnotě statistiky v celé populaci říkáme **parametr**.
  - Pro parametry používáme odpovídající písmena řecké abecedy
    - např. průměr: statistika m, parametr μ (mí)
    - další: $s - \sigma$ (sigma), $r - \rho$ (ró), $d - \delta$ (delta - rozdíl)
    - někdy také parametr θ a jeho odhad $\hat{\theta}$
- Statistiky jsou **odhady** parametrů
  - tj. jsou vždy zatíženy chybou – **výběrovou chybou**
  - *chyby náhodné* – umíme spočítat, známe-li **výběrové rozložení**
  - *chyby systematické* – nevhodné statistiky, špatné měření, špatný způsob výběru vzorku (metodologie)

Jak dobré jsou tyto odhady?

Výběrové rozložení a sm. chyba

- Spočítáme-li tutéž statistiku na mnoha nezávislých náhodných vzorcích
  - získáme mnoho různých odhadů parametru
  - tyto odhady mají nějaké rozložení - **výběrové rozložení (statistiky)**

http://onlinestatbook.com/stat_sim/sampling_dist/index.html

- **Výběrové rozložení** statistik obvykle můžeme popsat
  - průměrem – ten se u dobrých statistik blíží hodnotě **parametru**
  - směrodatnou odchylkou – říkáme jí **směrodatná chyba** ((odhadu) parametru) nebo také střední chyba a obecněji i výběrová chyba **(S.E.)**
  - Čím je velikost vzorku/ů větší, tím je směrodatná chyba menší

AJ: sampling distribution, standard error (of the mean)

Jak zjistíme výběrové rozložení statistiky?

Možnost opakovaně vybírat z populace a empiricky tak sestrojit výběrové rozložení je obvykle nedostupná, drahá, či neetická….

Parametrická teorie

- Výběrová rozložení běžných statistik jsou matematicky odvoditelná a známá…
  - …za nějakých předpokladů (typicky o parametrech popisovaných rozložení)

Neparametrická teorie – bootstrapping

- Výběrové rozložení se dá nasimulovat mnoha opakovanými výběry z našeho vzorku

Výběrové rozložení (odhadu) **průměru** dle teorie

Odhad průměru má přibližně **normální rozložení**,

- jehož průměr je μ se směrodatnou chybou ..... $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$
- Platí to i tehdy, když rozložení proměnné není normální.
  - a to „díky" **centrálnímu limitnímu teorému**
- Jenomže my obvykle neznáme σ...

Neznáme-li σ, musíme použít s

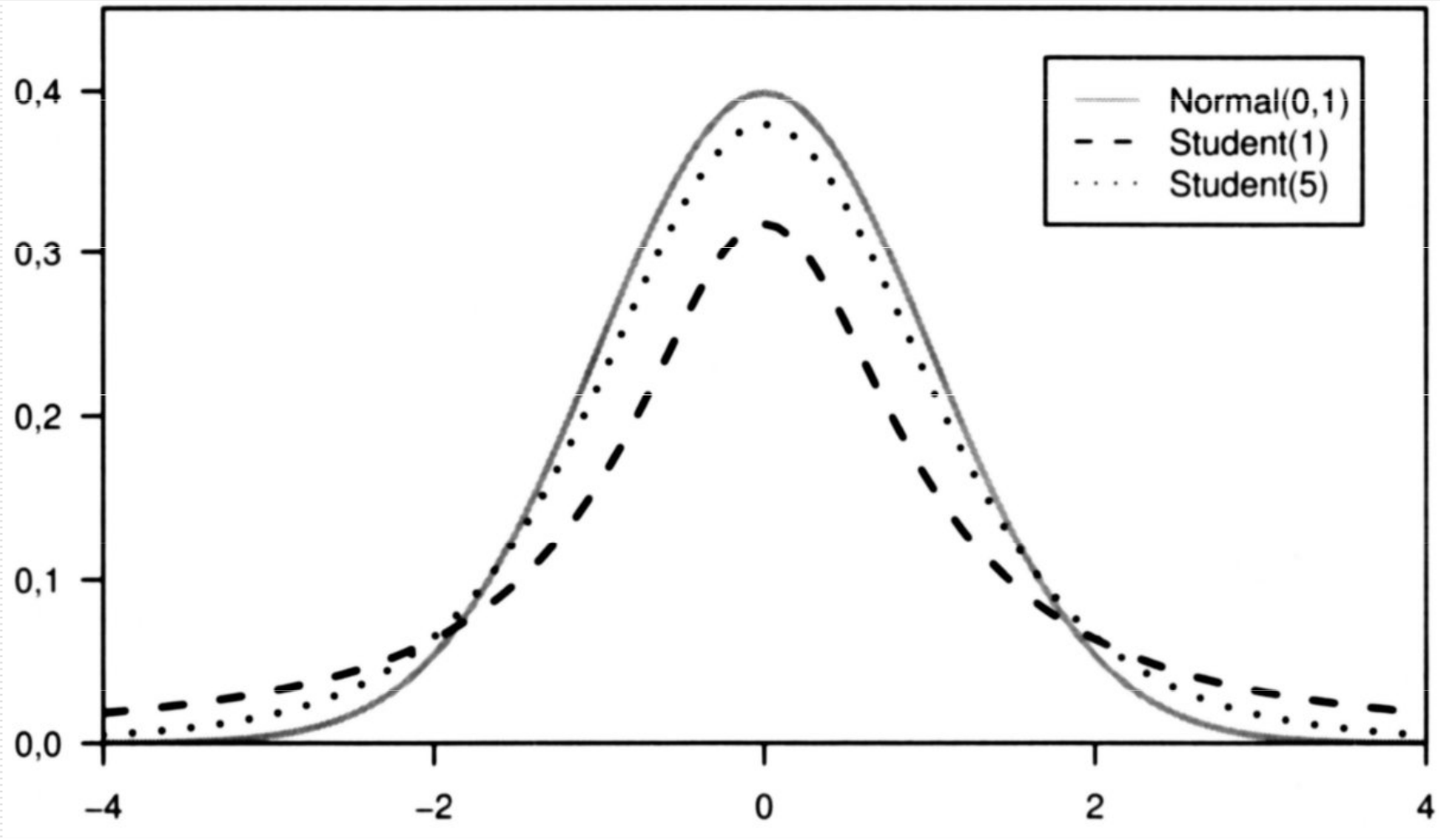
- průměr zůstává μ, směrodatná chyba je nyní ..... $s_{\bar{x}} = \frac{s}{\sqrt{N}}$
- výběrové rozložení není normální, jde o

  Studentovo *t* -rozložení
  - jako normální s těžšími konci (*t* je pro t-rozložení totéž, co *z* pro normální rozložení)
  - má různé tvary pro různá *n* : stupně volnosti – ν (ný)
    - zde $\nu = N-1$; čím vyšší *N*, tím se *t*-rozložení blíží normálnímu

AJ: central limit theorem, Student's t-distribution, degrees of freedom (d.f.)

Poznámka: odhad průměru = statistika *m*.

Studentovo *t* -rozložení



Výběrové rozložení (odhadu) **průměru**

… má tedy rozptyl *N*-krát menší než je rozptyl proměnné v populaci…

- $s^2_m = s^2/N$,

… a známe tvar jeho rozložení (N nebo *t*)

- Můžeme si tedy klást otázky typu *„Jak často se bude při velikosti vzorku N lišit námi spočítaný průměr od jeho populačního parametru o více než C?“*

Na vzorku 100 studentů psychologie jsme zjistili, že jejich průměrná hodnota potřeby struktury je 4,0 (s=0,9). Rozpětí škály je 1-5.

Jaká je *P*, že se mýlíme o více než 0,5 bodu, když tvrdíme, že v populaci studentů psychologie je M=4,0?

- Výběrové rozložení průměrů je t-rozložení s 99 stupni volnosti, průměrem μ a směrodatnou chybou $S.E.(M) = s_m = \frac{0,9}{\sqrt{100}} = 0,09$
- Mýlit se o více než 0,5 bodu znamená mýlit se o $\frac{0,5}{0,09} = 5,6$-násobek směrodatné chyby.
- Pravděpodobnost, že bychom se mýlili o více než 5,6-násobek S.E., je mizivá
  - Přesněji 2*(1-T.DIST(5,6; 99; 1))= 1,92E-07

Poznámky

- O kolik se tedy mýlíme? Nevíme přesně, neznáme přeci μ. Ale většina odhadů se od μ neliší o více než 1-2 S.E.

Výběrová rozložení dalších statistik

Nyní je tedy třeba ke každé popisné statistice znát ještě další vlastnost – její teoretické **výběrové rozložení**

- relativní četnost – přibližně normální - Hendl 162
- rozptyl – po transformaci χ^2-rozložení (chí kvadrát) - Hendl 159
- Pearsonova *r* – po Fisherově transformaci normální – Hendl 252

Teoretická výběrová rozložení různých statistik jsou různá

- Statistika je obvykle transformována do podoby, která má jedno z běžných teoretických rozložení: normální, chí-kvadrát rozložení (Pearsonovo), *t*-rozložení (Studentovo), *F*-rozložení (Fisherovo, Snedecorovo)
- Netřeba je znát z hlavy, programy je používají za vás, ale stojí za to vědět, že existují přehledy – např. Receptář Oseckých nebo Sheskin ISBN 1584884401
- Pro interpretační potřeby si obvykle vystačíme s představou výběrového rozložení průměru
- Pozor, centrální limitní teorém se týká pouze výběrového rozložení průměru!

AJ: chi-square distribution, F-distribution

Bootstrapping – způsob zjištění výběrového rozložení (jakékoli) statistiky hrubou silou

1. Máme náhodný vzorek z populace o velikosti N
2. Z našeho vzorku náhodně vylosujeme nový vzorek o velikosti N – výběr s vracením/opakováním – bootstrap, resample.
3. Na bootstrapu spočítáme kýženou statistiku a zaznamenáme si ji
4. Opakujeme body 2 a 3 mnohokrát, třebas 1000x
5. Získáme 1000 statistik, jejichž rozložení je výběrovým rozložením statistiky. Buď spočítáme jeho směrodatnou odchylku – S.E., nebo pracujeme přímo s jeho kvantily.

Česky je resampling převzorkování, bootstrapping je bootstrapping.

Příklad postupu v R

https://www.tutorialspoint.com/execute_r_online.php

```
# Výběr N=100 vzorku z populace s M=0 a SD=1
vzorek <- rnorm(100)

# Vytvoření prostoru, kam se budou ukládat průměry bootstrapů
prumery_b <- rep(NA, 1000)

# Bootstrapování
for (i in 1:1000) {prumery_b[i] <- mean(sample(vzorek, 100, replace=TRUE))}

# Histogram průměrů bootstrapových vzorků, tj. výběrové rozložení
hist(prumery_b)

# SD průměrů bootstrapových vzorků, tj. odhad SE
sd(prumery_b)
```

Estimační kvality statistik I

Kvality statistiky jako prostředku odhadu „skutečné“ hodnoty v populaci

TABLE 5.1 The Expected Values of the Range, s^2, and s as a Function of Sample Size of n Observations from a Random Sample from a Normal Distribution in which $\sigma = 10$

If $\sigma = 10$ *n*	*Expected Value of the Range*	*Expected Value of* s^2	*Expected Value of s*	*Expected Value of Range/s*
2	11	100	8.0	1.4
5	23	100	9.4	2.4
10	31	100	9.73	3.2
20	37	100	9.87	3.7
50	45	100	9.95	4.5
100	50	100	9.97	5.0
200	55	100	9.987	5.5
500	61	100	9.993	6.1
1,000	65	100	9.997	6.5

upraveno dle Glass, Hopkins

Estimační kvality statistik II

- Nezkreslenost
  - tj. že systematicky nenad(pod)hodnocuje
  - např. *s* podhodnocuje
- Konzistence
  - s velikostí vzorku roste přesnost odhadu
- Relativní účinnost
  - jak rychle roste přesnost s velikostí vzorku
  - zde vítězí *M* nad *Md* a strhává s sebou i další momentové statistiky
    - *jejich výhodou je i snadné počítání s nimi*

Alternativně Kvalita bodového odhadu viz Hendl 175

AJ: unbiasedness, consistency, relative efficiency

Bodové vs. intervalové odhady

α je p-nost chyby a proto je hladina spolehlivosti 1-α, tj. 95% spolehlivost znamená 5% chybovost: (1-0,05)

Parametr se můžeme snažit odhadnout…

- **bodovým odhadem** – tj. odhadujeme přímo hodnotu parametru, např. průměr.
- **intervalovým odhadem** – tj. odhadnutím intervalu, který parametr s určitou p-ností zahrnuje
  - výsledkem intervalového odhadu je **interval spolehlivosti**
  - interval spolehlivosti tvoříme z bodového odhadu a znalosti jeho výběrového rozložení, tj. (bod±odchylka)
  - intervalový odhad lepší - více informací $_{(1-\alpha)}CI = \overline{X} \pm_{1-\alpha/2} z\sigma_{\overline{X}}$
  - té p-nosti se v tomto kontextu říká **hladina spolehlivosti** (1-α)
    - typicky se používá 95% a 99% hladina spolehlivosti
    - pak říkáme, že hledaný parametr je s 95% p-ností v intervalu spolehlivosti

Zkuste si sami: http://onlinestatbook.com/stat_sim/conf_interval/index.html

AJ: point estimate, interval estimate, confidence interval (CI), level of confidence, consistency, unbiasedness, relative efficiency, resistence

$$_{(1-\alpha)}CI = \bar{X} \pm_{1-\alpha/2} z\sigma_{\bar{X}}$$

$$_{1-\alpha}\boldsymbol{CI} = \boldsymbol{M} \pm {}_{1-\alpha/2}\boldsymbol{zs_M}$$

$$_{1-\alpha}\boldsymbol{CI} = (\boldsymbol{M} - {}_{1-\alpha/2}\boldsymbol{zs_M};\ \boldsymbol{M} + {}_{1-\alpha/2}\boldsymbol{zs_M})$$

$$_{1-\alpha}\boldsymbol{CI} = (\boldsymbol{M} + {}_{\alpha/2}\boldsymbol{zs_M};\ \boldsymbol{M} + {}_{1-\alpha/2}\boldsymbol{zs_M})$$

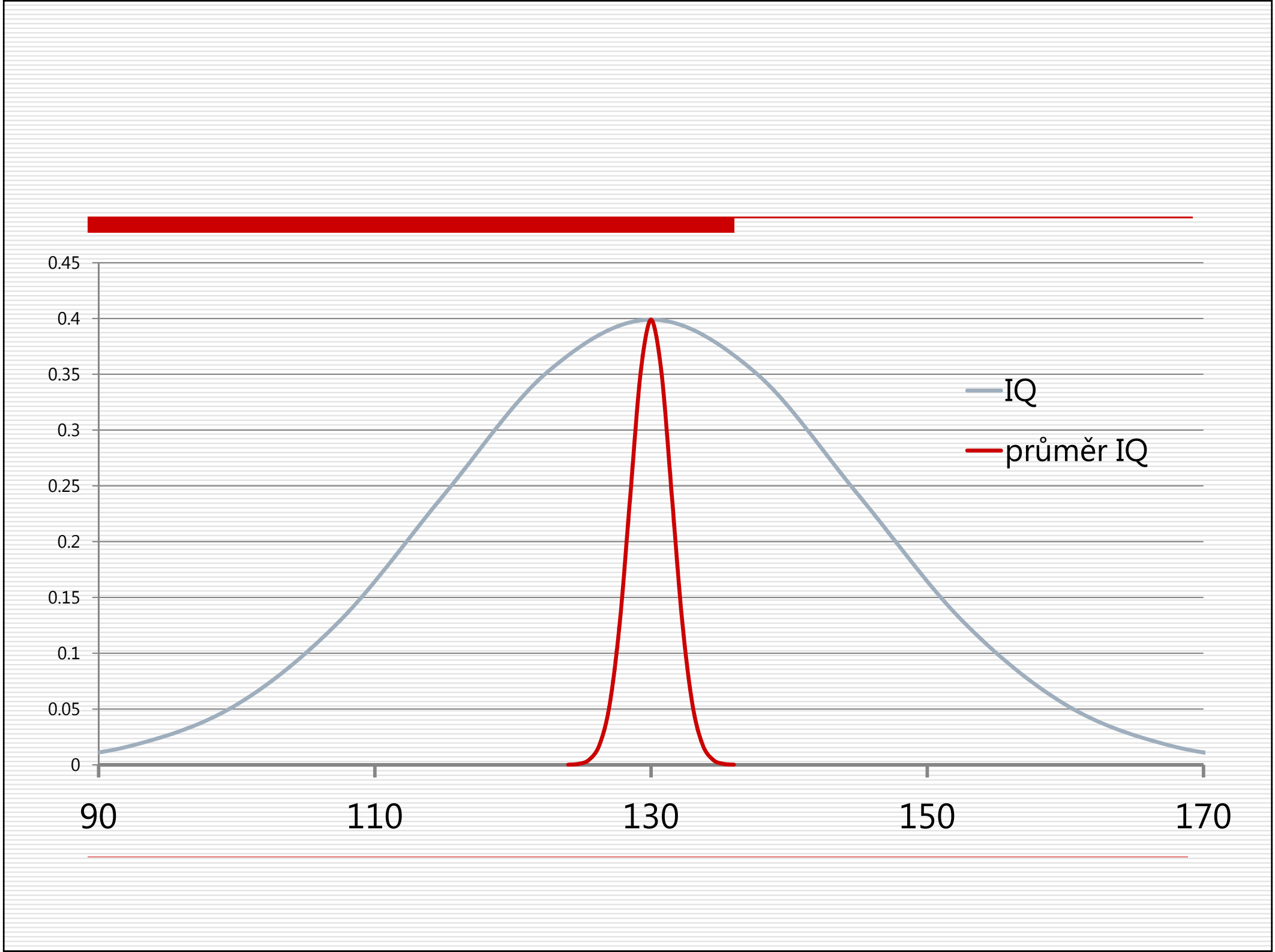
šířka intervalu: od $P_{\alpha/2}$ do $P_{1-\alpha/2}$ výběrového rozložení

umístění intervalu: kolem (výběrové) statistiky

Příklad konstrukce intervalu spolehlivosti pro průměr 1

Na vzorku dětí (*N*=100) s různobarevnýma očima jsme spočítali průměrné IQ 130, přičemž víme, že σ =15.

- **bodový odhad** průměrného IQ v populaci dětí s různobarevnýma očima (tj. parametru, μ) je 130
- **intervalový odhad**
  - Známe-li σ, výběrové rozložení průměru má **normální rozložení…**
  - …se středem v μ. μ neznáme, a tak použijeme bodový odhad $m = 130$
  - … se směrodatnou chybou odhadu průměru $s_m = \sigma/\sqrt{N} = 15/\sqrt{100} = 1{,}5.$
  - Zvolíme-li hladinu spolehlivosti $1\text{-}\alpha = 95\%$,
  - pak v tabulkách/Excelu zjistíme, že 95% normálního rozl. je mezi hodnotami $z = -1{,}96$ a $1{,}96$,tj. $_{1-\alpha/2}z = {}_{0{,}975}z = 1{,}96$, Excel: =NORMSINV(0,975)
  - interval spolehlivosti: $(m - 1{,}96s_m;\ m + 1{,}96s_m) = (127{,}1\ ;\ 132{,}9)$,
  - **tj. s 95% pravděpodobností 127,1 ≤ μ ≤ 132,9**


0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
90
110
130
150
170
IQ
průměr IQ

Příklad konstrukce intervalu spolehlivosti pro průměr 2

Na vzorku dětí (*N*=100) s různobarevnýma očima jsme spočítali průměrné IQ 130 a *s* =15.

- **bodový odhad** průměrného IQ v populaci dětí s různobarevnýma očima (tj. parametru, μ) je 130
- **intervalový odhad**
  - střed intervalu spolehlivosti bude na bodovém odhadu, tj. $m = 130$
  - víme, že výběrové rozložení průměru má *t*–rozložení se stupni volnosti $\nu = N-1 = 99$
  - zvolíme-li hladinu spolehlivosti $1-\alpha = 95\%$,
  - pak v tabulkách (Excelu) zjistíme, že 95% *t*-rozložení je mezi hodnotami t=-1,98 a 1,98 (tj. ${}_{1-\alpha/2}t\,(\nu) = {}_{0,975}t\,(99) = 1,98$ excel: TINV(0,05;99))
  - směrodatná chyba odhadu průměru $s_m = s/\sqrt{n} = 15/\sqrt{100} = 1,5$
  - interval spolehlivosti: $(m - 1,98s_m;\ m + 1,98s_m) = (127,0\ ;\ 133,0)$,
  - **tj. s 95% pravděpodobností $127,0 \leq \mu \leq 133,0$**

pozor na tento rozdíl: ve středu intervalu je *m*, někde v intervalu je v 95% případů μ

Interpretace intervalu spolehlivosti

- … je prostá, avšak zrádná
- 95% interval spolehlivosti znamená, že sestrojujeme-li tento interval dle výše uvedených instrukcí, **v 95% případů sestrojení intervalu tento interval zahrnuje odhadovaný parametr**, tj. v 95% případů je závěr, že μ je mezi čísly *a* a *b*, správný.
- V tomto smyslu to také znamená, že máme subjektivní 95% jistotu, že parametr je v námi určeném intervalu.
- V konkrétním případě, kdy jsme spočetli konkrétní interval spolehlivosti ($127 \leq \mu \leq 133$), to neznamená, že v 95% případech je μ v intervalu od 127 do 133.
  - To proto, že μ je konstanta; při opakovaných výzkumech se nemění. Díky omylnému výběru v každém výzkumu vychází poněkud jiný interval sestrojený podle jiného výběrového průměru. Jinými slovy, trefujeme se obručí na kolík a ne kolíkem do obruče.
- O čem tohle slovíčkaření je? O rozdílu mezi četnostním a subjektivním (Bayesovským) pojetím pravděpodobnosti.

...Výběrové rozložení **mediánu**

- Simulace: www.stat.tamu.edu/~jhardin/applets/signed/SampDist2.html
- V případě normálního rozložení je taky normální a směrodatná chyba je cca 1,25 směrodatné chyby průměru
- Pořadový způsob nabízí Campbell a Gardner[1]
  - Přibližný interval (pro N>100) se stanovuje opravdu pořadovým způsobem, tj. počítáme pořadí, které určuje horní a dolní mez intervalu
  - Pro 95% interval spolehlivosti pak je *r* pořadí určující horní mez a *s* pořadí určující dolní mez

$$r = \frac{n}{2} - z_{1-\alpha/2}\frac{\sqrt{n}}{2} \qquad s = 1 + \frac{n}{2} + z_{1-\alpha/2}\frac{\sqrt{n}}{2}$$

- Bootstrap
  - Obecná metoda, nejen pro mediány, téměř bez předpokladů (neparametrická)
  - Algoritmus:
    - 1. Proveďte výběr s navracením ze svého výběru (o velikosti N)
    - 2. Spočítejte medián a uložte
    - 3. Opakujte kroky 1 a 2 tisíckrát
  - 95% interval je ohraničen 25. a 975. nejvyšším spočítaným mediánem.

[1]Campbell, M.J., Gardner, M.J. (2000). Medians and their differences. In Altman et al., *Statistics with confidence (36 – 44)*. BMJ Books.

…Výběrové rozložení **relativní četnosti** p

- Pro dostatečně velkou populaci ($np>10$; $n(1-p)>10$)…
- …je přibližně normální s průměrem p a směrodatnou chybou $\sqrt{p(1-p)/n}$
- $(1-\alpha)$% interval spolehlivosti má tedy podobu:

$$(p - z_{1-\alpha/2}\sqrt{p(1-p)/n};\ p + z_{1-\alpha/2}\sqrt{p(1-p)/n})$$

…proto na malých vzorcích může být těžké usuzovat na rozložení proměnné

…Výběrové rozložení **rozptylu** s^2

- Rozložení poměru $(s^2/\sigma^2)(n-1)$ má podobu chí-kvadrát rozložení s $\nu = n-1$ stupni volnosti

$$\frac{s^2}{\sigma^2}(n-1) \sim \chi^2(\nu)$$

- $(1-\alpha)$% interval spolehlivosti pro σ^2 má tedy podobu:

$$\left(s^2 \frac{n-1}{\chi^2_{1-\alpha/2}(\nu)}; s^2 \frac{n-1}{\chi^2_{\alpha/2}(\nu)} \right)$$

- V Excelu =CHISQ.INV(1-α;*df*)=$\chi^2_{1-\alpha}(df)$ [=CHIINV(α;*df*)]

…Výběrové rozložení Pearsonovy **korelace** *r*

- Výběrové rozložení korelace neznáme.
- Známe výběrové rozložení korelace po Fisherově transformaci: $Z = 0{,}5\ ln((1+r)/(1-r)) = \text{arctgh}(r) = \text{FISHER}(r)$
- Výběrové rozložení Z je přibližně normální s průměrem Z a směrodatnou chybou $s_Z = 1/\sqrt{(n-3)}$
- (1–α)% CI pro Z: $\left(Z - z_{1-\alpha/2}s_Z ; Z + z_{1-\alpha/2}s_Z\right)$
- Nutno transformovat zpět do metriky korelačního koeficientu: $r=(e^{2Z}-1)/(e^{2Z}+1)=\text{FISHERINV}(Z)$

$$\left(FISHERINV\left(Z - z_{1-\alpha/2}s_Z\right); FISHERINV\left(Z + z_{1-\alpha/2}s_Z\right)\right)$$

Shrnutí

- Na vzorcích počítáme **statistiky**, které jsou odhadem populačních **parametrů**.
- K posouzení přesnosti odhadu musíme znát **výběrové rozložení** statistiky, kterou k odhadu používáme, zejména jeho variabilitu – **směrodatnou chybu**.
- Výběrové rozložení známe buď z teorie, nebo ho získáme **bootstrapováním**
- Směrodatná chyba klesá především s velikostí vzorku a s variabilitou jevu v populaci.
- Přesnost odhadu parametru sdělujeme prostřednictvím **intervalu spolehlivosti**.