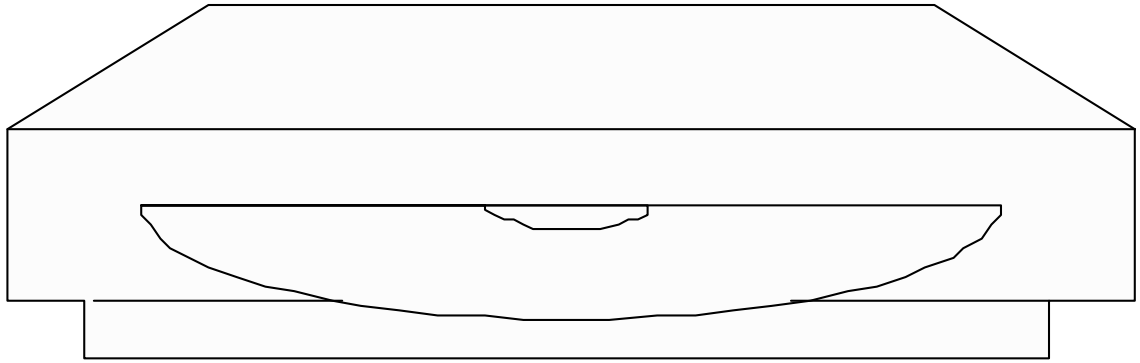
Analýza dat – lekce 03
Standardizované normální rozložení. Z – skóre.

© Petr Mareš
Fakulta sociálních studií
katedra sociologie

STANDARDIZOVANÉ NORMÁLNÍ ROZLOŽENÍ



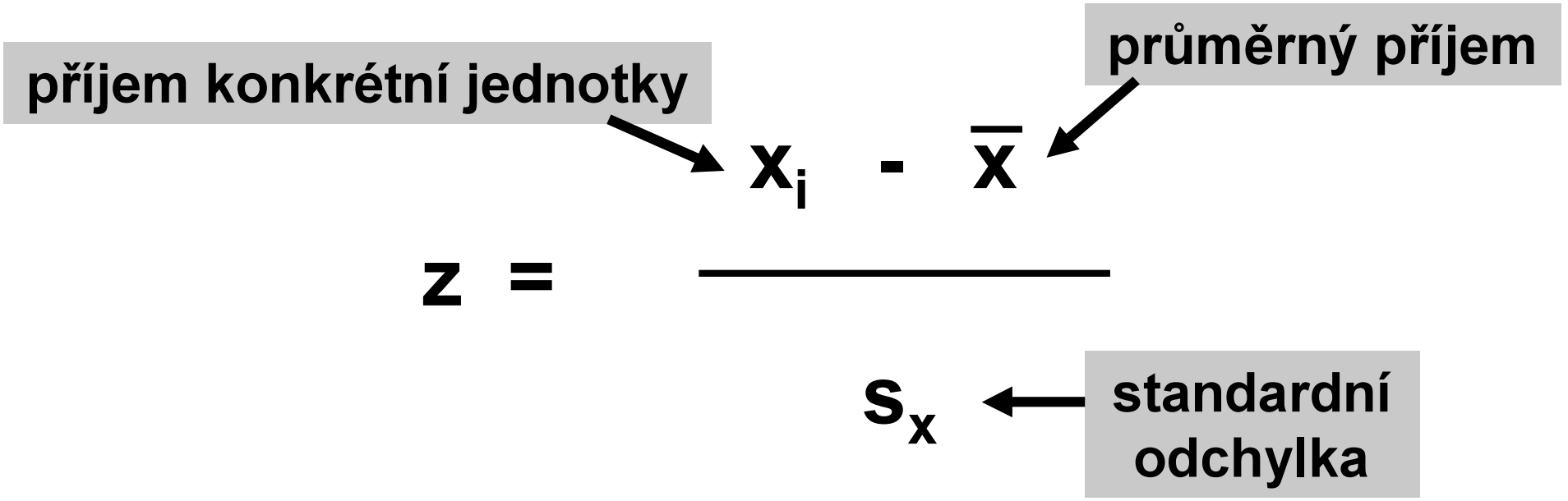
STANDARDNÍ (NORMOVANÉ) NORMÁLNÍ ROZLOŽENÍ

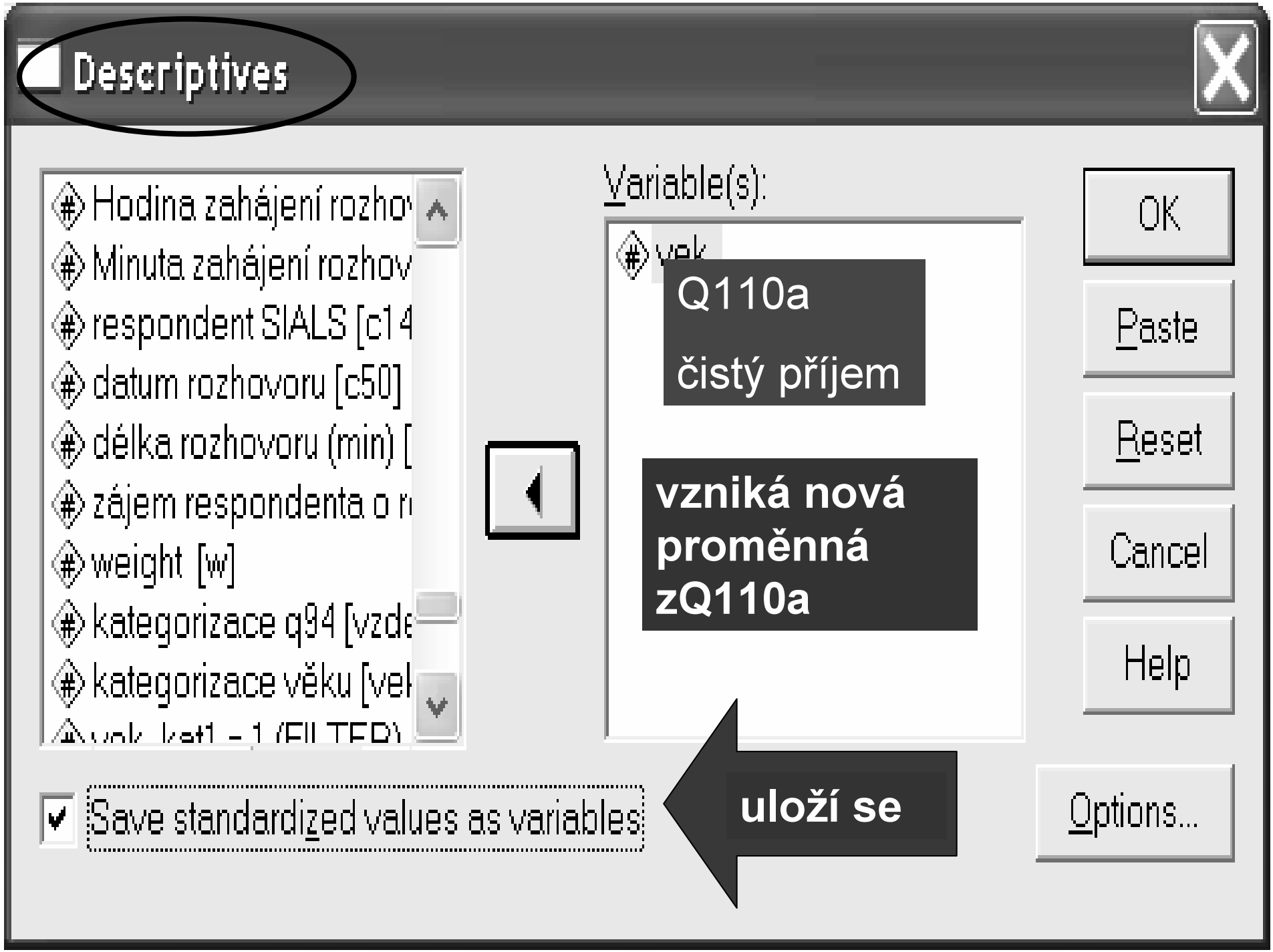
Normální rozložení získává praktický význam až standardizací (normováním).

Ve STANDARDNÍM (NORMOVANÉM) NORMÁLNÍM ROZLOŽENÍ jsou všechny hodnoty (daného znaku u všech jednotek) vyjádřeny standardním skórem.

STANDARDNÍ SKÓRE (Z-SKÓRE)

Udává *"jaký násobek standardní odchylky pod či nad průměrem se původní hodnota nachází"*.




Descriptives
Hodina zahájení rozhov
Minuta zahájení rozhov
respondent SIALS [c14
datum rozhovoru [c50]
délka rozhovoru (min) [
zájem respondenta o r
weight [w]
kategorizace q94 [vzde
kategorizace věku [vek
Variable(s):
vek
Q110a
čistý příjem
vzniká nová proměnná zQ110a
OK
Paste
Reset
Cancel
Help
Options...
Save standardized values as variables
uloží se

Příklad:

- **1. test: Průměr = 6 bodů, std. odchylka = 4,2**
- **2. test: Průměr = 10 bodů, std. odchylka = 3,**

JEDINEC DOSÁHL v:

- **1. testu 12 bodů a má standardizované skóre:**

$$z = \frac{12 - 6}{4,2} = 1,43$$

- **V 2. testu 12 bodů, má standardizované skóre:**

$$z = \frac{12 - 10}{3,6} = 0,56$$

V prvním testu se tedy umístil lépe.

POMOCÍ Z-SKÓRE STANDARDIZUJEME DATA

Do seskupovací analýzy (cluster analysis) okresů ČR vstupují proměnné o různém měřítku:

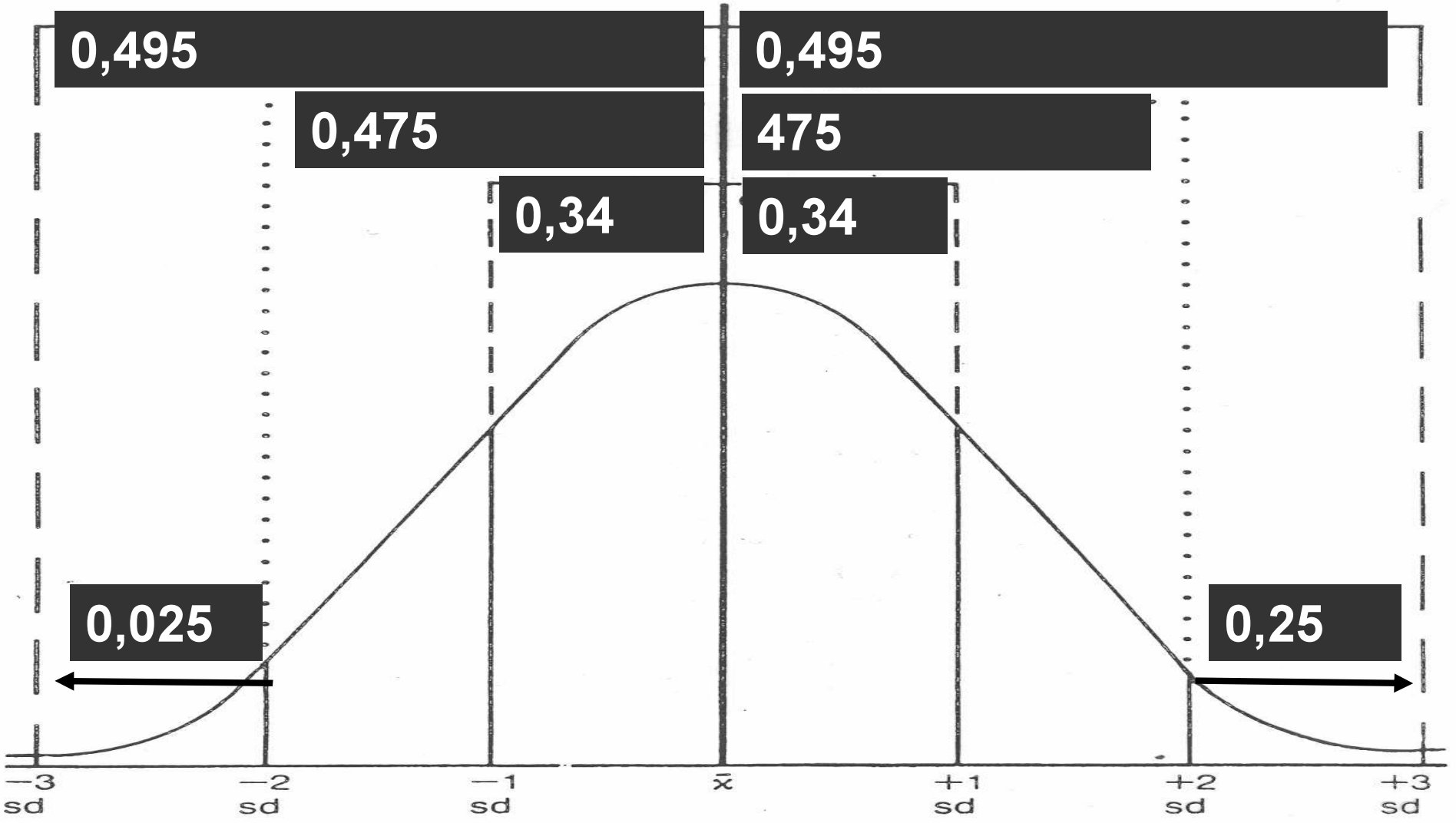
- **Míra nezaměstnanosti (má možnost nabývat hodnot od 0 do 100: nikdo není nezaměstnaný = 0 až všichni jsou nezaměstnaní = 100%).**
- **Výše přijmu (má možnost nabývat hodnot například od 0 do 1 000 000 nebo i více).**
- **Proměnná ….**

Váha proměnných o různém řádu by ve výpočtu byla nesouměřitelná (proměnné s větším řádem by měly větší váhu). Proto je zaměníme za z-skóre.

NORMÁLNÍ ROZLOŽENÍ

(jeho vlastnosti)

Teoreticky se křivka rozkládá od 0 do 1



CO OBVYKLE ZJIŠŤUJEME

OBECNĚ:

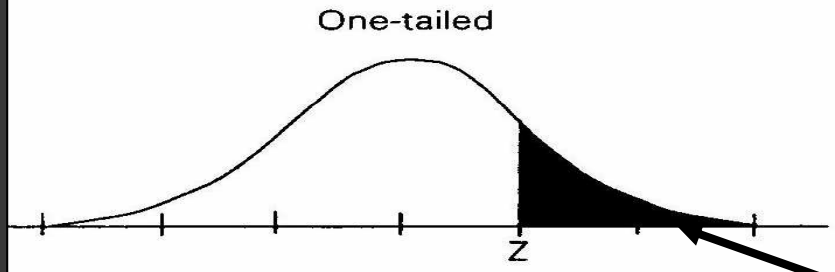
Jak je zjištěná hodnota při předpokladu „teoretického“ rozložení pravděpodobná či nepravděpodobná.

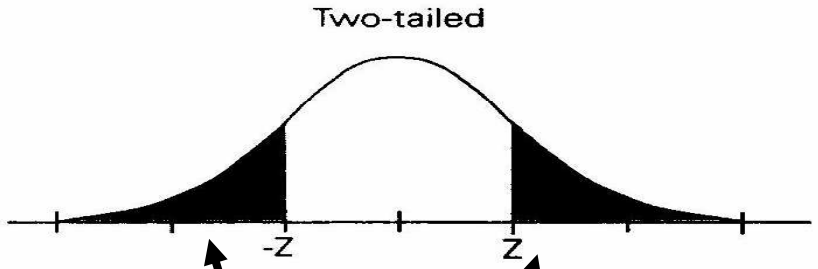
U NORMÁLNÍHO rozložení:

Jak je zjištěná hodnota při předpokladu „normálního“ rozložení pravděpodobná či nepravděpodobná.

Oblasti pod normální křivkou

kritické hodnoty: překračovány s danou pravděpodobností

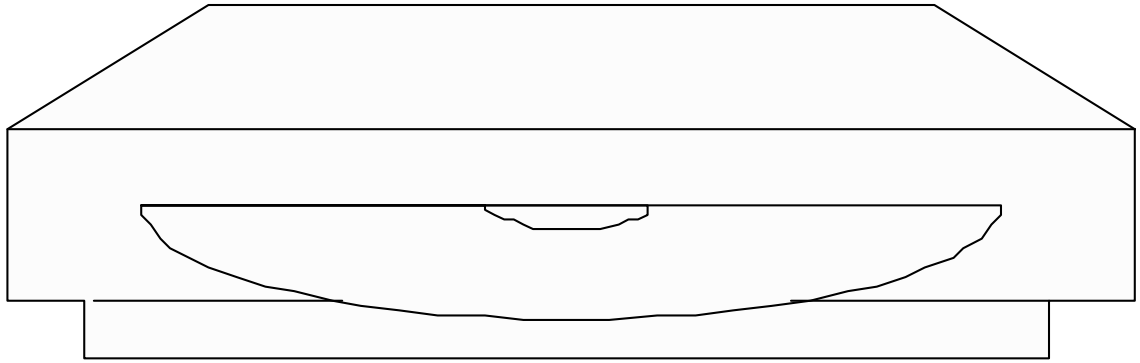




Z Scores	Probability	
	One-tailed	Two-tailed
.0	.50000	1.00000
.1	.46017	.92034
.2	.42074	.84148
.3	.38209	.76418
.4	.34458	.68916
.5	.30854	.61708
.6	.27425	.54851
.7	.24196	.48393
.8	.21186	.42371
.9	.18406	.36812
1.0	.15866	.31731
1.1	.13567	.27133
1.2	.11507	.23014
1.3	.09680	.19360
1.4	.08076	.16151
1.5	.06681	.13361
1.6	.05480	.10960
1.7	.04457	.08913
1.8	.03593	.07186
1.9	.02872	.05743

Z Scores	Probability	
	One-tailed	Two-tailed
1.96	.02500	.05000
2.0	.02275	.04550
2.1	.01786	.03573
2.2	.01390	.02781
2.3	.01072	.02145
2.4	.00820	.01640
2.5	.00621	.01242
2.6	.00466	.00932
2.7	.00347	.00693
2.8	.00256	.00511
2.9	.00187	.00373
3.0	.00135	.00270
3.1	.00097	.00194
3.2	.00069	.00137
3.3	.00048	.00097
3.4	.00034	.00067
3.5	.00023	.00047
3.6	.00016	.00032
3.7	.00011	.00022
3.8	.00007	.00014

INFERENČNÍ STATISTIKA



STATISTICKÁ INFERENCE

Výběrový soubor ⟹ základní soubor.

Smysluplné je jen:

- **Jde-li o VÝBĚR (při vyčerpávajícím šetření to nemá smysl).**
- **Jde-li o NÁHODNÝ VÝBĚR (jednotky mají stejnou pravděpodobnost, že budou vybrány.**
- **Jde-li o NEZÁVISLÝ VÝBĚR (výběr žádné jednotky nezvyšuje ani nesnižuje pravděpodobnost výběru jiných jednotek).**

Příklady závislého výběru:

- **Opisují-li studenti v testu, jejich výsledky nejsou nezávislé).**
- **Párovaná data.**

PARAMETR je NEZNÁMÁ veličina (nemáme-li možnost vyčerpávajícího šetření) vlastnost základního souboru

μ = průměr v základním souboru

σ = standardní odchylka v základním souboru

σ^2 = variance v základním souboru

STATISTIKA je ZNÁMÁ vlastnost výběrového souboru

x = průměr ve výběrovém souboru

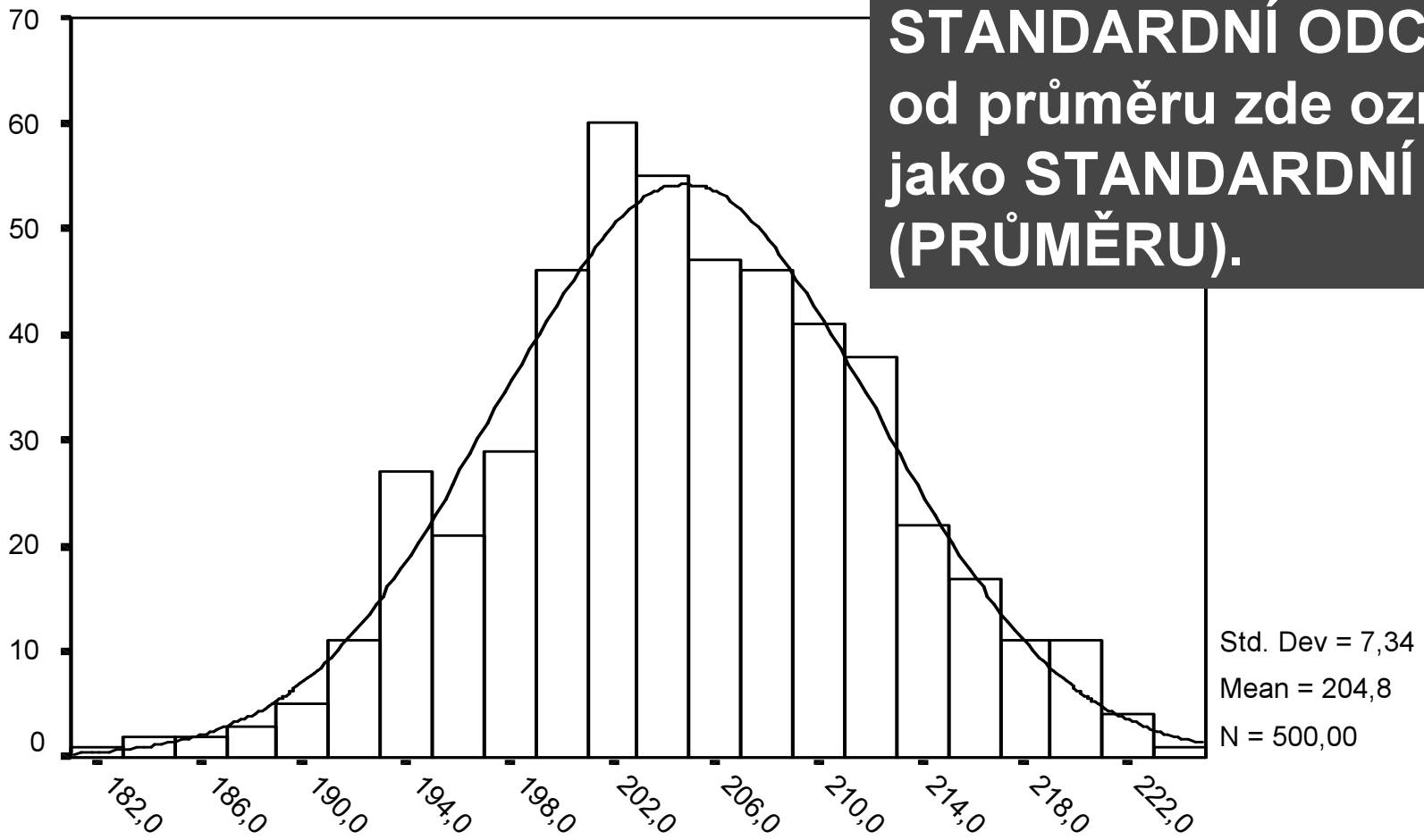
s = standardní odchylka ve výběrovém souboru

$s2$ = variance ve výběrovém souboru

Standardní odchylku rozdělení výběrových průměrů nazýváme STANDARDNÍ CHYBOU

(SE.M. = Standard Error of Mean)

Je to směrodatná odchylka distribuce výběrových souborů - *sampling distribution*. Říká nám, jak variují VÝBĚROVÉ PRŮMĚRY z téže populace kolem PARAMETRU.



STANDARDNÍ ODCHYLKU od průměru zde označujeme jako STANDARDNÍ CHYBU (PRŮMĚRU).

Histogram nepředstavuje rozložení hodnot nějaké proměnné ve výběrovém souboru, ale ROZLOŽENÍ PRŮMĚRŮ JEJÍCH ROZLOŽENÍ V 500 VÝBĚRECH. Distribuce je cca normální (čím více výběrů bychom provedli, tím více by se normální distribuci blížila).

Standardní chybu průměru můžeme vypočítat, známe-li velikost výběrového souboru a standardní odchylku v populaci.

(obvykle ji sice neznáme, ale lze ji odhadnout).

VÝBĚROVÁ CHYBA

Protože je rozložení průměrů všech možných výběrů NORMÁLNÍ, pak lze určit kde se zvolenou pravděpodobností leží parametr.

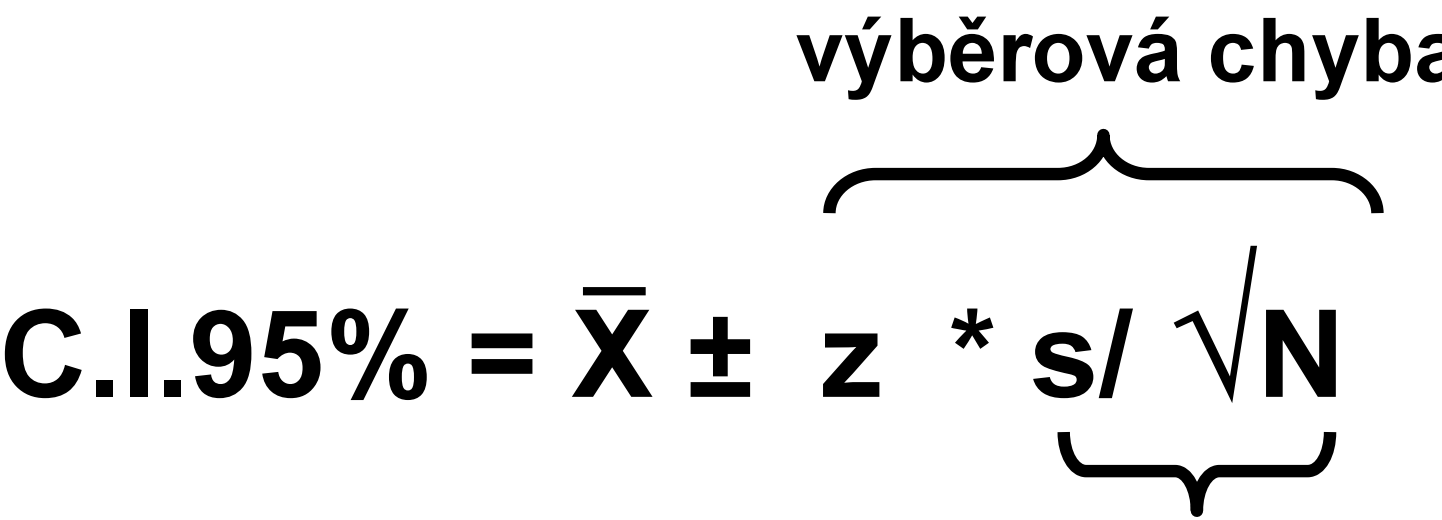
Zvolíme-li např. pravděpodobnost 95% (5% riziko chyby), měl by PARAMETR ležet v intervalu ± 1,96 směrodatné chyby (což je VÝBĚROVÁ CHYBA pro tuto pravděpodobnost) od průměru průměrů ze všech možných výběrů.

(SPSS nám standardní chybu vypočítá)

INTERVAL SPOLEHLIVOSTI

Nevíme tedy, kde parametr leží přesně, víme však alespoň to v jakém intervalu parametr leží při zvolené pravděpodobnosti.

INTERVAL SPOLEHLIVOSTI



INTERVAL SPOLEHLIVOSTI

(Confidence Interval **pro průměr na HV = 95%)**

výběrová chyba

$$C.I.95\% = \bar{X} \pm 1{,}96 * s/\sqrt{N}$$

standardní/směrodatná chyba

Při HV = 99% bychom jako z použili ± 2,96

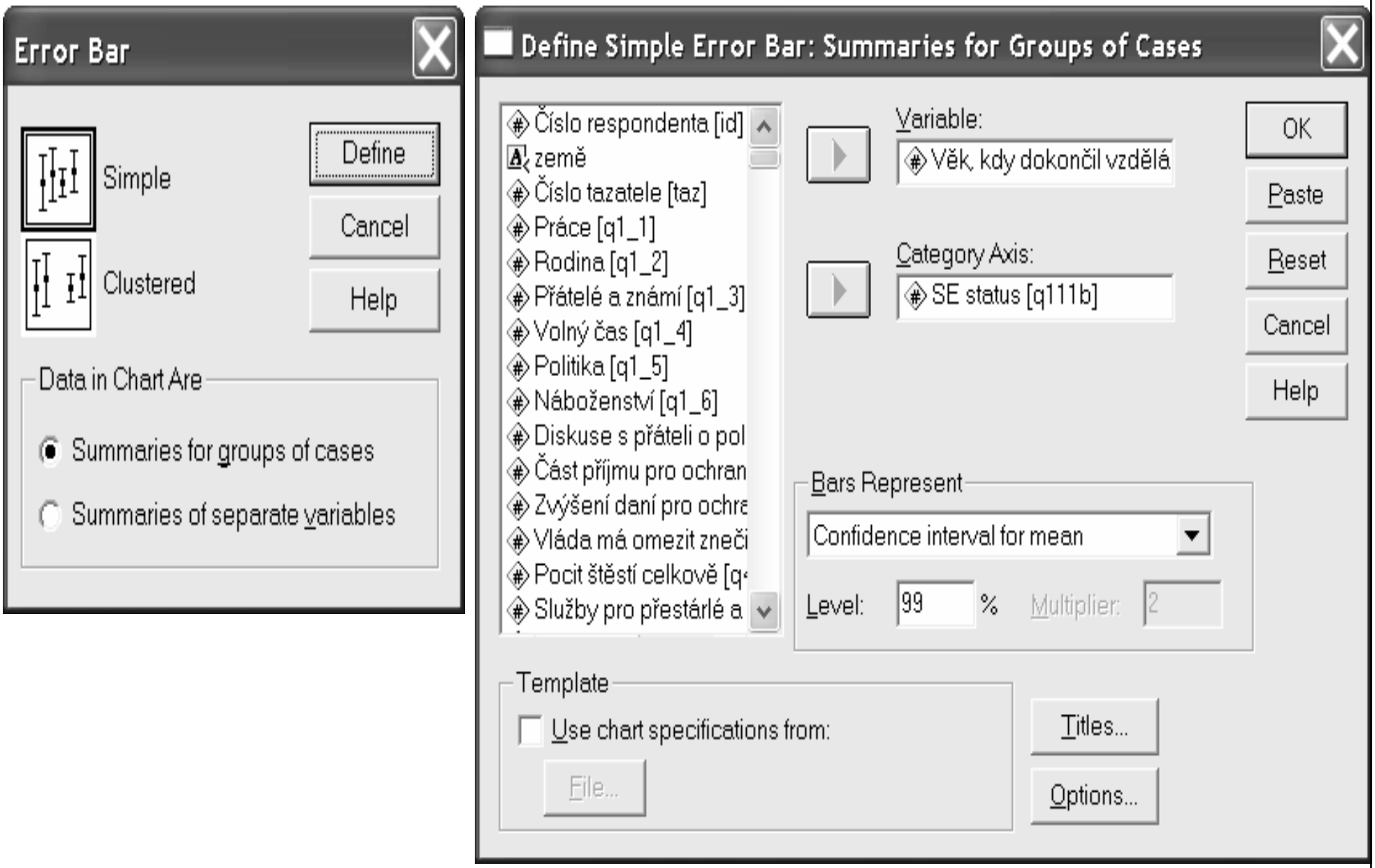
- Obvyklý je 95% interval spolehlivosti (ze 100 výběru bude 95 správných). Vybereme nejmenší interval pod normálním rozložením $\bar{X}$, jemuž odpovídá 95% pravděpodobnost. (dvě krajní oblasti s pravděpodobností 2,5% na každé straně ponecháme stranou).
- Z tabulky kumulativních pravděpodobností normovaného normálního rozložení lze zjistit kritickou hodnotu z = 1,96 pro tuto pravděpodobnost. Interval je určen $\pm$ 1,96 směrodatné odchylky výběrového průměru (směrodatné chyby).

Věk, kdy dokončil vzdělání

Descriptives

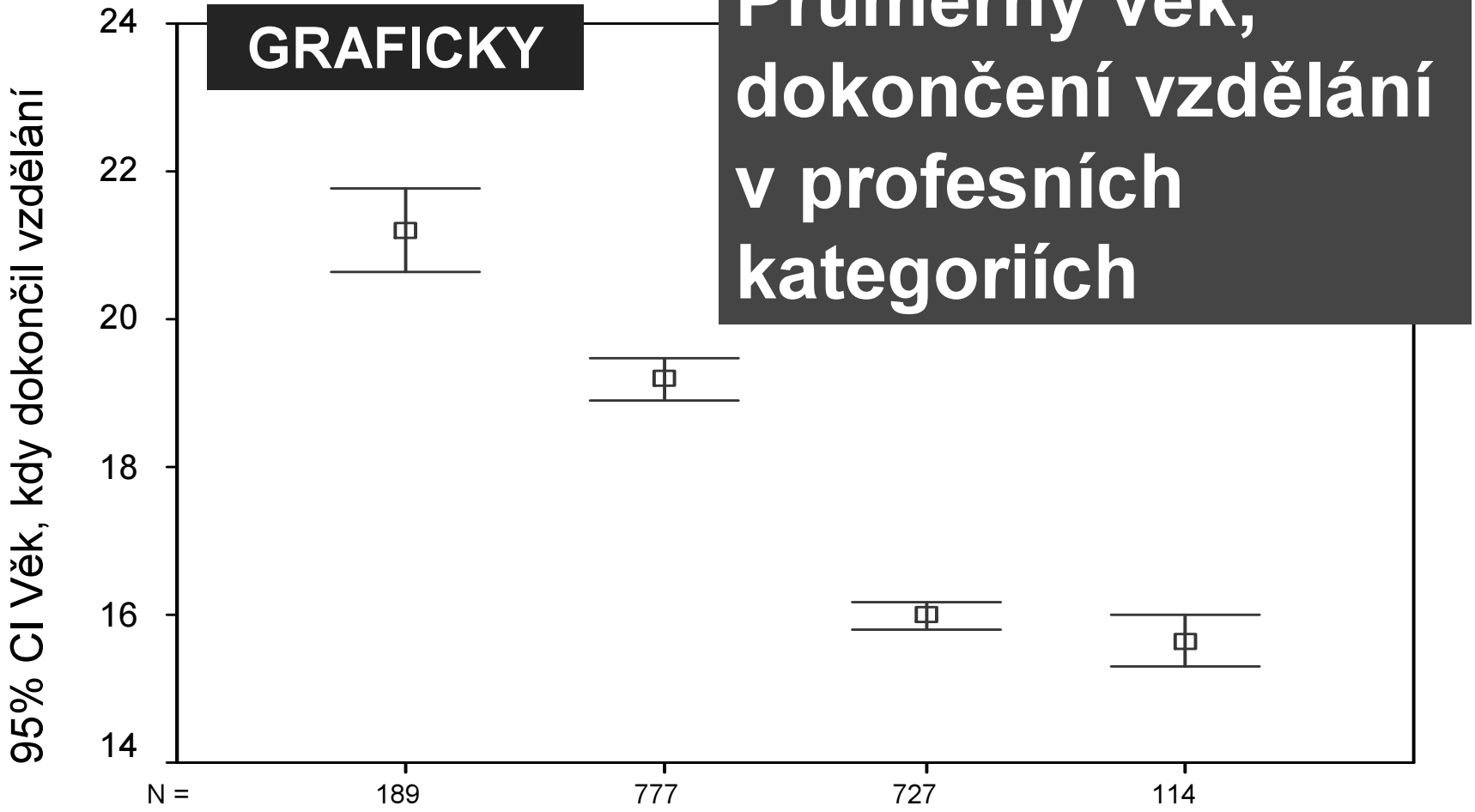
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Min.	Max.
					Lower Bound	Upper Bound		
vyšší, vyšší střední třída	189	21,21	3,980	,290	20,64	21,78	14	40
střední nemanuální pracovníci	777	19,19	4,015	,144	18,91	19,47	10	42
kvalifikovaní a polokvalifikovaní dělníci	727	15,99	2,382	,088	15,82	16,16	12	44
pomocní dělníci, nezaměstnaní	114	15,64	1,876	,176	15,29	15,98	14	24
Total	1807	17,89	3,837	,090	17,71	18,07	10	44

GRAPHS ⟶ ERROR BAR



Průměrný věk, dokončení vzdělání v profesních kategoriích

GRAFICKY



SE status

Cases weighted by W

	průměr	stand. chyba	95% interval spolehlivosti		99% interval spolehlivosti širší interval	
Vyšší, vyšší střední	**21,21**	**0,290**	**20,64**	**21,78**	**20,35**	**22,07**
Střední nemanuální	**19,19**	**0,144**	**18,91**	**19,47**	**18,76**	**19,62**
Kvalifikovaní dělníci	**15,99**	**0,088**	**15,82**	**16,16**	**15,73**	**16,25**
Pomocní dělníci	**15,64**	**0,176**	**15,29**	**15,98**	**15,12**	**16,16**
CELKEM	**17,89**	**0,090**	**17,71**	**18,07**	**17,62**	**18,16**

INTERVAL SPOLEHLIVOSTI

(Confidence Interval **pro % výskytu** na HV = 95%)

$$C.I.95\% = p \pm 1{,}96 * \sqrt{p*(1-p) / N}$$

- **p = pozorovaný podíl, kolem něhož je interval spolehlivosti konstruován**
- **N = velikost výběrového souboru**

Příklad: Ve výběrovém souboru 1100 osob ze základní populace by volilo určitou politickou stranu 30% voličů:

$$C.I.95\% = p \pm 1{,}96 \cdot \sqrt{p*q / N}$$

$$C.I.95\% = 30 \pm 1{,}96 \cdot \sqrt{30*70} / 1100$$

$$C.I.95\% = 2{,}7 \sim 3$$

V základním souboru by ji s 95% pravděpodobností volilo:

ne méně než 27% a ne více jak 33% voličů.

VŠIMNĚME SI

Pokud by volilo určitou politickou stranu jen 5% voličů:

$$C.I.95\% = 5 \pm 1{,}96 \cdot \sqrt{5*95 / 1100} = 4{,}3$$

Výběrová chyba je nejen větší, ale má i větší význam.

Stejně velká výběrová chyba má různý dopad dle velikosti inferované hodnoty.

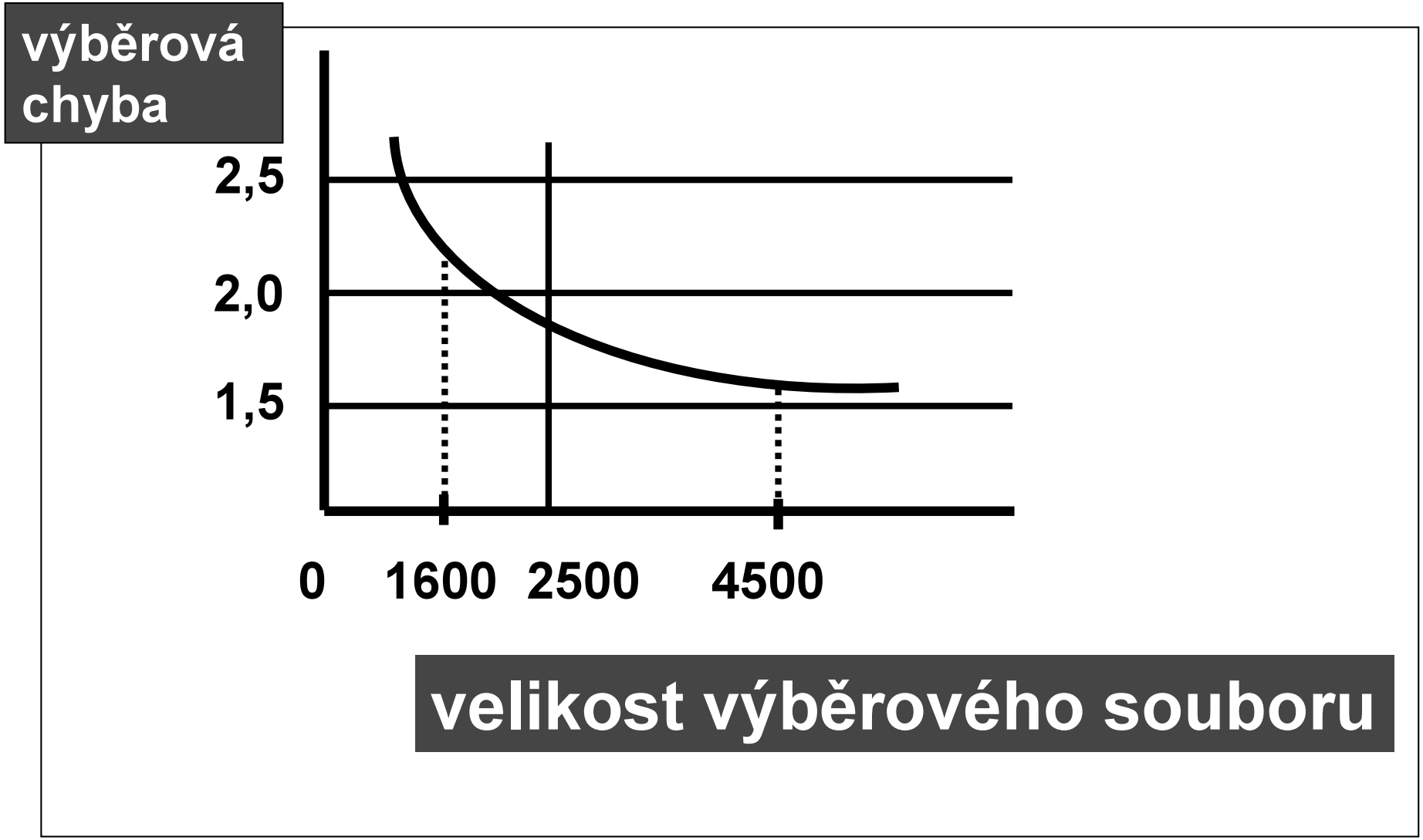
3% ± 3	<0;6>
5% ±3	<2;8>
60%	<57;63>

CO OVLIVŇUJE VELIKOST STANDARDNÍ CHYBY

V případě, že by byla v populaci stejná proporce dvou vlastností (muž, žena, ...), pak ve výběru o 100 jedincích by byl interval spolehlivosti ±10%, ale ve výběru o 400 jedincích ±5% a ve výběru 1000 jedinců ±3%

(konkrétně při požadované hladině významnosti 95%).

VÝBĚROVÁ CHYBA s velikostí výběru klesá, ale po dosažení jeho určité velikosti je její pokles s dalším zvětšováním výběru nepodstatný (zvětšování výběru není ekonomické).



výběrová chyba v %	Interval spolehlivosti	velikost výběru (sample size)	výběrová chyba v %	interval spolehlivosti	velikost výběru (sample size)
1,0	±1,0	10 000	6,0	± 6,0	277
1,5	± 1,5	4 500	6,5	± 6,5	237
2,0	± 2,0	2 500	7,0	± 7,0	204
2,5	± 2,5	1 600	7,5	± 7,5	178
3,0	± 3,0	1 100	8,0	± 8,0	156
3,5	± 3,5	816	8,5	± 8,5	138
4,0	± 4,0	625	9,0	± 9,0	123
4,5	± 4,5	494	9,5	± 9,5	110
5,0	± 5,0	400	10,0	± 10,0	100
5,5	± 5,5	330			

PŘÍKLAD:

- **Kdyby bylo ve výběrovém souboru (při jeho velikosti 100 jednotek) 90% osob podporujících vstup ČR do EU a 10% odpůrců tohoto vstupu (nebo naopak), pak by byl interval spolehlivosti ±6%.**
- **Kdyby byl podíl podpory vstupu do EU a odporu proti němu ve stejně velkém výběrovém souboru vyrovnaný (50% a 50%), interval spolehlivosti by byl ±10%**

(požadovaná hladina významnosti 95%).

počet	1% nebo 99%	5% nebo 95%	10% nebo 90%	15% nebo 85%	20% nebo 80%	25% nebo 75%	30% nebo 70%	35% nebo 65%	40% nebo 60%	45% nebo 55%	50%
25	4,0	8,7	12,0	14,3	16,0	17,3	18,3	19,1	19,6	19,8	20,0
50	2,8	6,2	8,5	10,1	11,4	12,3	13,0	13,5	13,9	14,1	14,2
75	2,3	5,0	6,9	8,2	9,2	10,0	10,5	11,0	11,3	11,4	11,5
100	2,0	4,4	6,0	7,1	8,0	8,7	9,2	9,5	9,8	9,9	10,0
150	1,6	3,6	4,9	5,9	6,6	7,1	7,5	7,8	8,0	8,1	8,2
200	1,4	3,1	4,3	5,1	5,7	6,1	6,5	6,8	7,0	7,0	7,1
250	1,2	2,7	3,8	4,5	5,0	5,5	5,8	6,0	6,2	6,2	6,3
300	1,1	2,5	3,5	4,1	4,6	5,0	5,3	5,5	5,7	5,8	5,8
400	0,99	2,2	3,0	3,6	4,0	4,3	4,6	4,8	4,9	5,0	5,0
500	0,89	2,0	2,7	3,2	3,6	3,9	4,1	4,3	4,4	4,5	4,5
600	0,81	1,8	2,5	2,9	3,3	3,6	3,8	3,9	4,0	4,1	4,1
800	0,69	1,5	2,1	2,5	2,8	3,0	3,2	3,3	3,4	3,5	3,5
1000	0,63	1,4	1,9	2,3	2,6	2,8	2,9	3,1	3,1	3,2	3,2
2000	0,44	0,96	1,3	1,6	1,8	1,9	2,0	2,1	2,2	2,2	2,2