PSY117 2018
Statistická analýza dat v psychologii
Přednáška 3

Transformace skórů a kvantily normálního rozložení

Shrnutí z minula

- Prvním cílem analýzy je zjistit, jaké hodnoty proměnné se v datech vyskytují, jaké jsou jejich četnosti a jak jsou četnosti rozložené.
- Rozložení pak můžeme popsat jednotlivými četnostmi a/nebo ukazateli centrální tendence a variability.
  - Četnosti, ukazatele centrální tendence a variability jsou popisné statistiky – popisují rozložení
- Rozložení zobrazujeme sloupcovými diagramy, histogramem, boxplotem

- Kódování proměnných je do značné míry arbitrární.
- Jak ovlivňují různá nakódování tvar rozložení?
- Můžeme překódováním proměnné – TRANSFORMACÍ – tvar rozložení záměrně měnit?
- Můžeme TRANSFORMACÍ usnadnit porozumění statistikám?

Transformace skórů (dat)

Pro usnadnění porozumění a možnost dalších analýz často přepočítáváme hodnoty proměnných, aby měly lepší vlastnosti

- Usnadnění interpretace – *lineární transformace*
  - např. vynásobení 10 nebo 100 pro odstranění desetinných míst
  - tvar rozložení zůstává zachován
  - možnost sjednocení různých proměnných na stejnou škálu, měřítko … Standardizace
- Změna tvaru rozložení – *nelineární transformace*
  - log/exp fce, (od)mocniny, Tukey: „ladder of powers" Hendl kap. o EDA.
  - Též „normalizace" rozložení – normalizované skóry
- Změna úrovně měření – *pořadová transformace* *(ranking)*

Lineární transformace 1

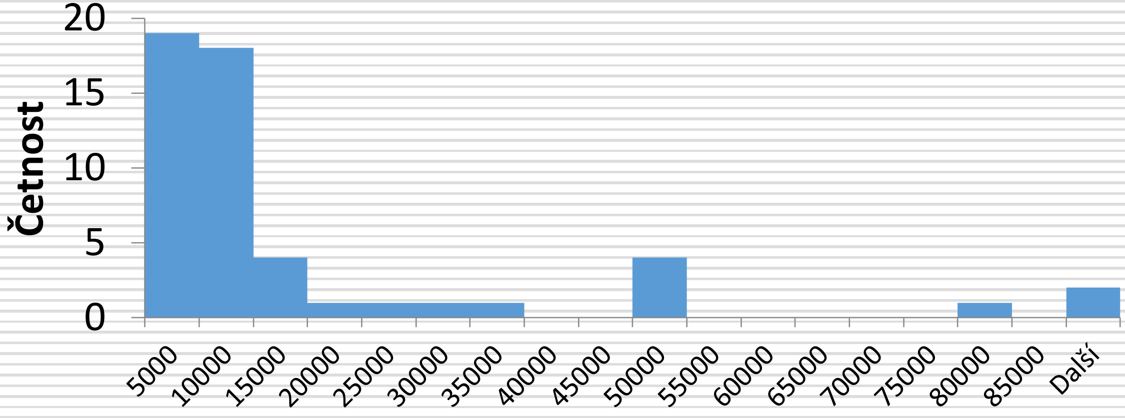
- Např. počtu psychologů z jednotek na tisíce

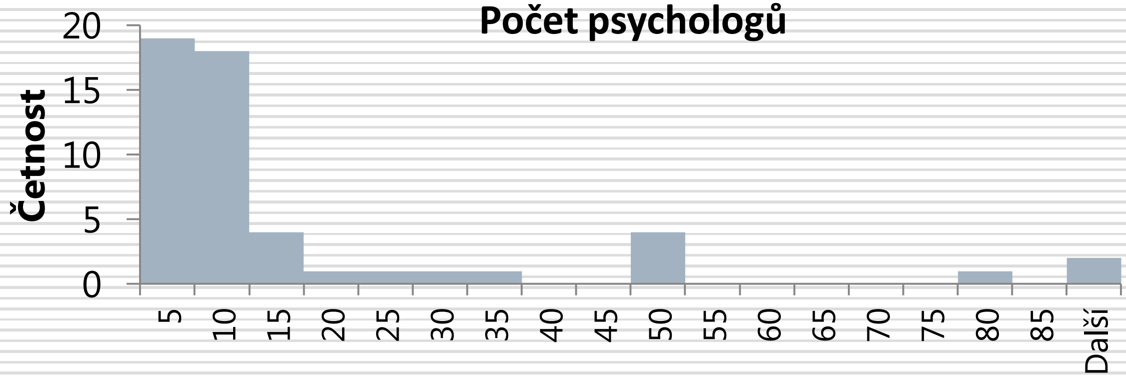
HRUBÝ SKÓR

  - Tvar rozložení zachován
  - Popisné statistiky se předpověditelně změní
  - *M, SD, Md, IQR, min, max* jsou tisíckrát menší
  - s^2 (VAR)?

	poc_psy	v tisících
	5000	5
	1200	1,2
	1000	1
	10000	10
	12000	12
	4000	4
	1500	1,5
	10000	10
	100000	100
	10000	10
	12000	12
	10000	10
	150000	150
	35000	35
	80000	80
	50000	50
	17000	17
	1385	1,385
	2000	2
	10000	10
	5000	5
	10000	10
	9999	9,999
	50000	50
	10000	10
	10000	10
	12500	12,5
	3000	3
	15000	15
	3 000	3
	6000	6
	2000	2
	5000	5
	10000	10
	10000	10
	10000	10
	5000	5
	5000	5
	5743	5,743
	8000	8
	3500	3,5
	1500	1,5
	25000	25
	10000	10
	10000	10
	50000	50
	30000	30
	50000	50
	5000	5
	7000	7
	5000	5
	1000	1
M	17602	17,60
SD	27410	27,41
VAR	751320187,5	751,3202

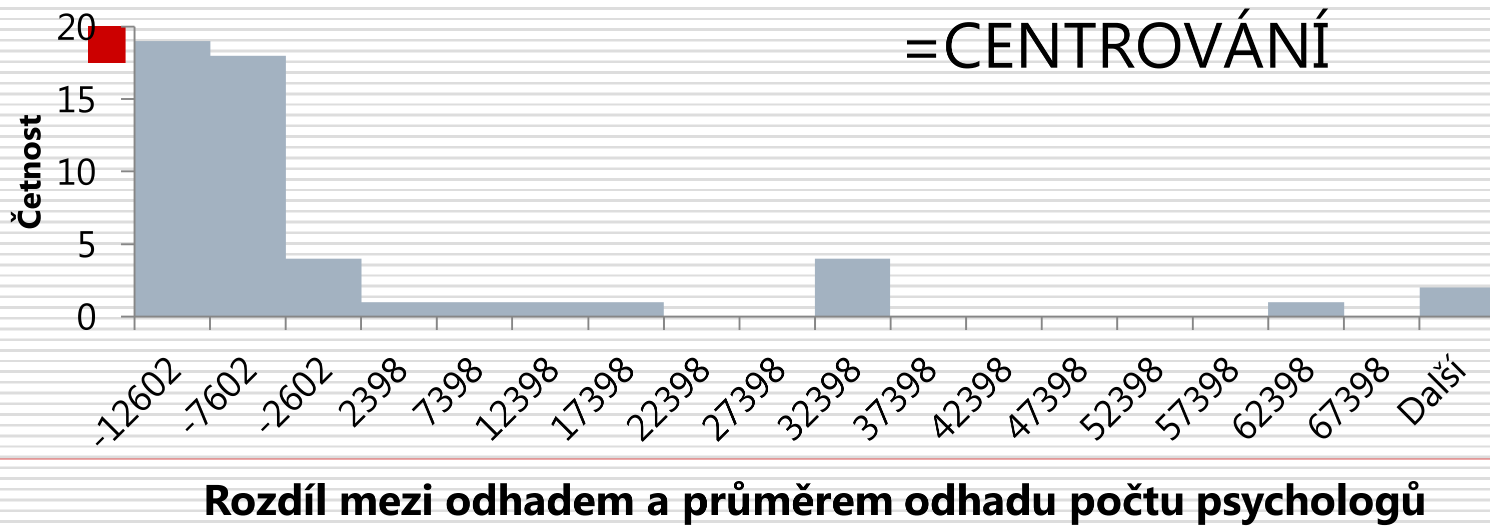
AJ: raw score


Četnost
20
15
10
5
0
5000
10000
15000
20000
25000
30000
35000
40000
45000
50000
55000
60000
65000
70000
75000
80000
85000
Další
Počet psychologů


Četnost
20
15
10
5
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
Další
Počet psychologů v tisících

Lineární transformace 2

- Deviační skóry x_i - odečtení průměru
  - Tvar rozložení zůstává zachován
  - Popisné statistiky – CT jsou o průměr menší, variabilita beze změn
  - Snadnější interpretace jednotlivých skórů
  - =CENTROVÁNÍ



	poc_psy	v tisících	dev
	5000	5	-12602
	1200	1,2	-16402
	1000	1	-16602
	10000	10	-7602
	12000	12	-5602
	4000	4	-13602
	1500	1,5	-16102
	10000	10	-7602
	100000	100	82398
	10000	10	-7602
	12000	12	-5602
	10000	10	-7602
	150000	150	132398
	35000	35	17398
	80000	80	62398
	50000	50	32398
	17000	17	-602
	1385	1,385	-16217
	2000	2	-15602
	10000	10	-7602
	5000	5	-12602
	10000	10	-7602
	9999	9,999	-7603
	50000	50	32398
	10000	10	-7602
	10000	10	-7602
	12500	12,5	-5102
	3000	3	-14602
	15000	15	-2602
	3 000	3	-14602
	6000	6	-11602
	2000	2	-15602
	5000	5	-12602
	10000	10	-7602
	10000	10	-7602
	10000	10	-7602
	5000	5	-12602
	5000	5	-12602
	5743	5,743	-11859
	8000	8	-9602
	3500	3,5	-14102
	1500	1,5	-16102
	25000	25	7398
	10000	10	-7602
	10000	10	-7602
	50000	50	32398
	30000	30	12398
	50000	50	32398
	5000	5	-12602
	7000	7	-10602
	5000	5	-12602
	1000	1	-16602
M	17602	17,60	0,00
SD	27410	27,41	27410
VAR	751320188	751,3202	751320188
Md	10000	10,000	-7602
IQR	7125	7,125	7125
min	1000	1	-16602,4423
max	150000	150	132397,558

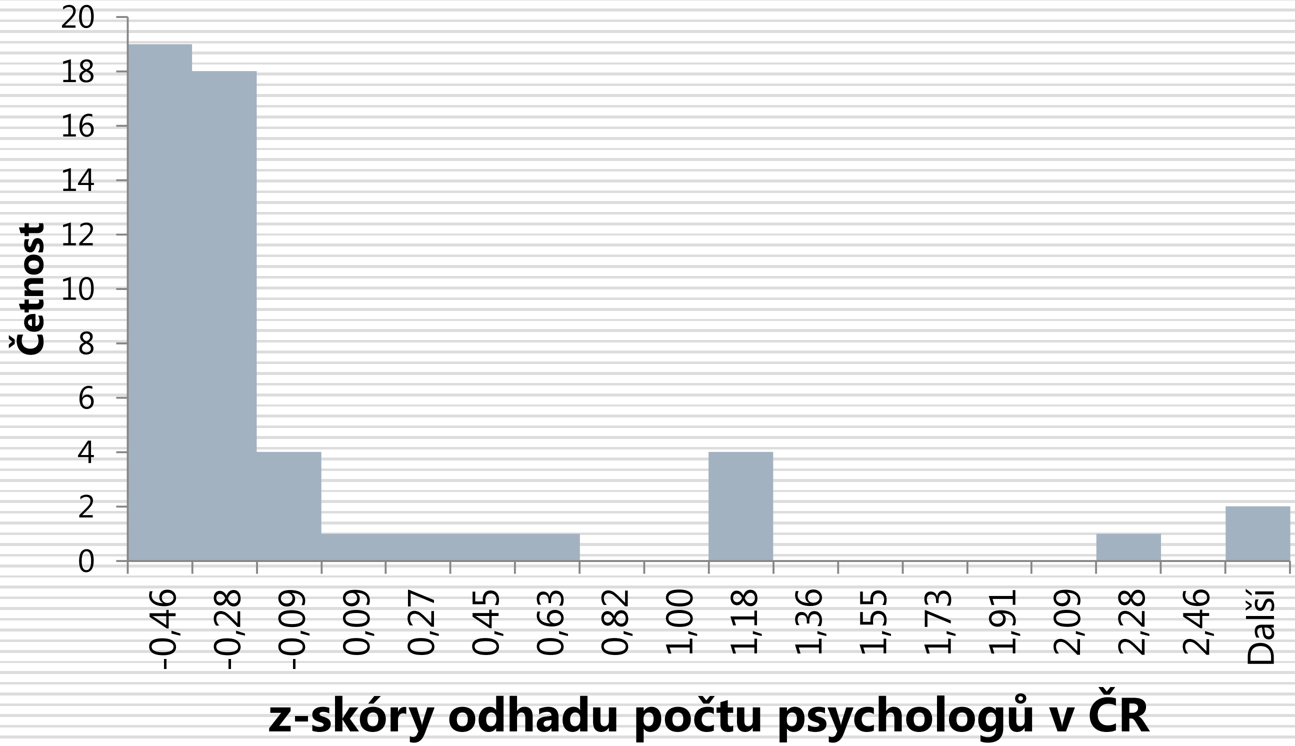
Lineární transformace - standardizace *z*-skóry, standardizované skóry

- Nejobvyklejší lineární transformace - **standardizace**
  - transformace sady skórů, aby $m = 0$, $s = 1$
  - **jednotkou měření se stává *s*,** možnost srovnávání různých škál (ale pozor rozdíly v rozložení zůstávají!)

$$z_i = (X_i - m) / s$$

  - s. je zajímavá zvláště u normálně rozložených dat, protože známe řadu jeho percentilů zpaměti
    - u přibližně normálně rozložených dat o lidech je většina (přes 90%) lidí mezi -3 a 3

AJ: data transformations, standard(ized) scores, *z*-scores, normalized scores



	poc_psy	v tisících	dev	z
	5000	5	-12602	-0,46
	1200	1,2	-16402	-0,60
	1000	1	-16602	-0,61
	10000	10	-7602	-0,28
	12000	12	-5602	-0,20
	4000	4	-13602	-0,50
	1500	1,5	-16102	-0,59
	10000	10	-7602	-0,28
	100000	100	82398	3,01
	10000	10	-7602	-0,28
	12000	12	-5602	-0,20
	10000	10	-7602	-0,28
	150000	150	132398	4,83
	35000	35	17398	0,63
	80000	80	62398	2,28
	50000	50	32398	1,18
	17000	17	-602	-0,02
	1385	1,385	-16217	-0,59
	2000	2	-15602	-0,57
	10000	10	-7602	-0,28
	5000	5	-12602	-0,46
	10000	10	-7602	-0,28
	9999	9,999	-7603	-0,28
	50000	50	32398	1,18
	10000	10	-7602	-0,28
	10000	10	-7602	-0,28
	12500	12,5	-5102	-0,19
	3000	3	-14602	-0,53
	15000	15	-2602	-0,09
	3 000	3	-14602	-0,53
	6000	6	-11602	-0,42
	2000	2	-15602	-0,57
	5000	5	-12602	-0,46
	10000	10	-7602	-0,28
	10000	10	-7602	-0,28
	10000	10	-7602	-0,28
	5000	5	-12602	-0,46
	5000	5	-12602	-0,46
	5743	5,743	-11859	-0,43
	8000	8	-9602	-0,35
	3500	3,5	-14102	-0,51
	1500	1,5	-16102	-0,59
	25000	25	7398	0,27
	10000	10	-7602	-0,28
	10000	10	-7602	-0,28
	50000	50	32398	1,18
	30000	30	12398	0,45
	50000	50	32398	1,18
	5000	5	-12602	-0,46
	7000	7	-10602	-0,39
	5000	5	-12602	-0,46
	1000	1	-16602	-0,61
M	17602	17,60	0,00	0,00
SD	27410	27,41	27410	1
VAR	751320188	751,3202	751320188	1
Md	10000	10,000	-7602	0
IQR	7125	7,125	7125	0,25994
min	1000	1	-16602,4423	-0,6057
max	150000	150	132397,558	4,830226

Skóry odvozené ze *z*-skórů

Používané primárně v psychodiagnostických metodách

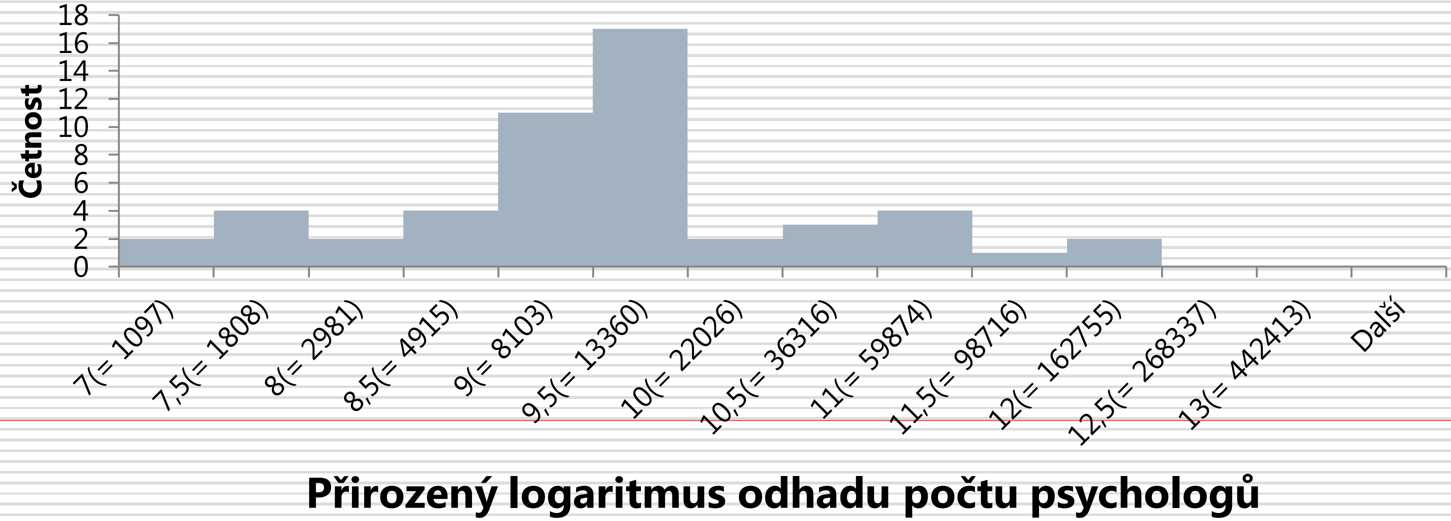
- ***IQ* skóry** (*m*=100, *s*=15)
- ***T* skóry** (*m*=50, *s*= 10)

- **staniny**, staninové skóry (standard *nine*) (*m*=5, *s*= 2) + kategorizace zaokrouhlením na *celá* čísla … od 1 do 9
- **steny**, stenové skóry (standard *ten*) (*m*=5,5, *s*= 2) + kategorizace zaokrouhlením na *celá* čísla … od 1 do 10

Standardní skóry mají pořád stejné rozložení jako hrubé skóry!

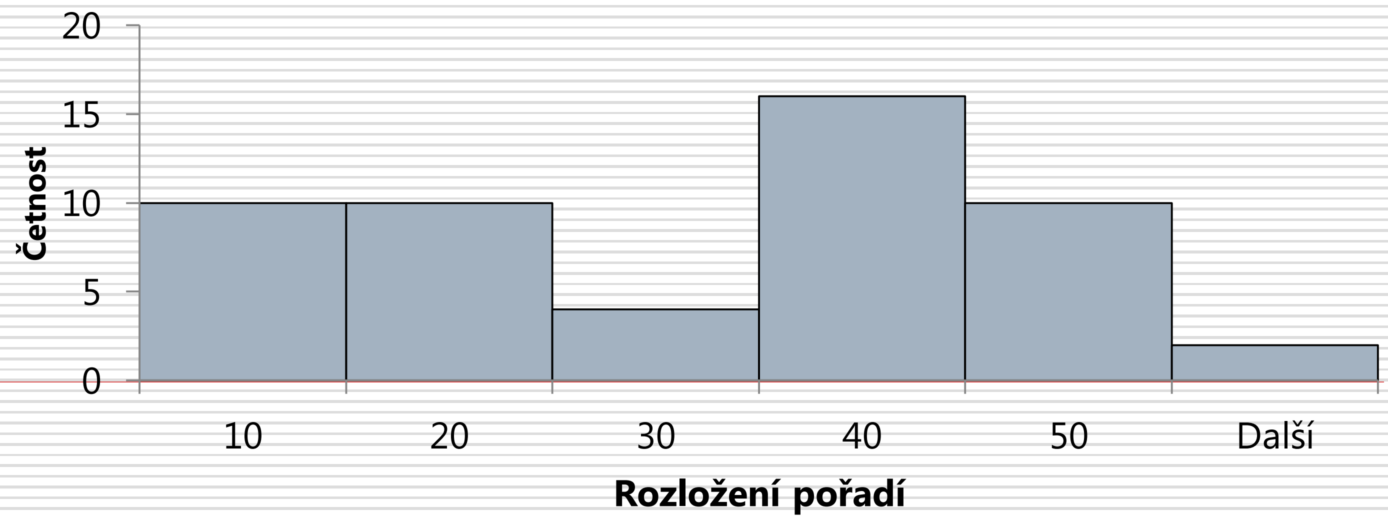
Nelineární transformace 1

- Změna rozložení (obv. směrem k normálnímu)
  - Pro smysluplnější využití momentových statistik
  - Pro možnost využití analytických technik, které nějakou podobu rozložení vyžadují
- Popisné statistiky se mění složitěji
- Př. logaritmus počtu psychologů



Nelineární transformace 2

- Transformace na pořadí – ranking
  - Eliminace odlehlých hodnot, odhlédnutí od velikosti rozdílů mezi lidmi
  - Obvykle vzestupně (nejnižší hodnota má pořadí 1)
  - Stejné hodnoty dostávají průměrné pořadí (=RANK.AVG)
  - Výsledné rozložení je (přibližně) uniformní



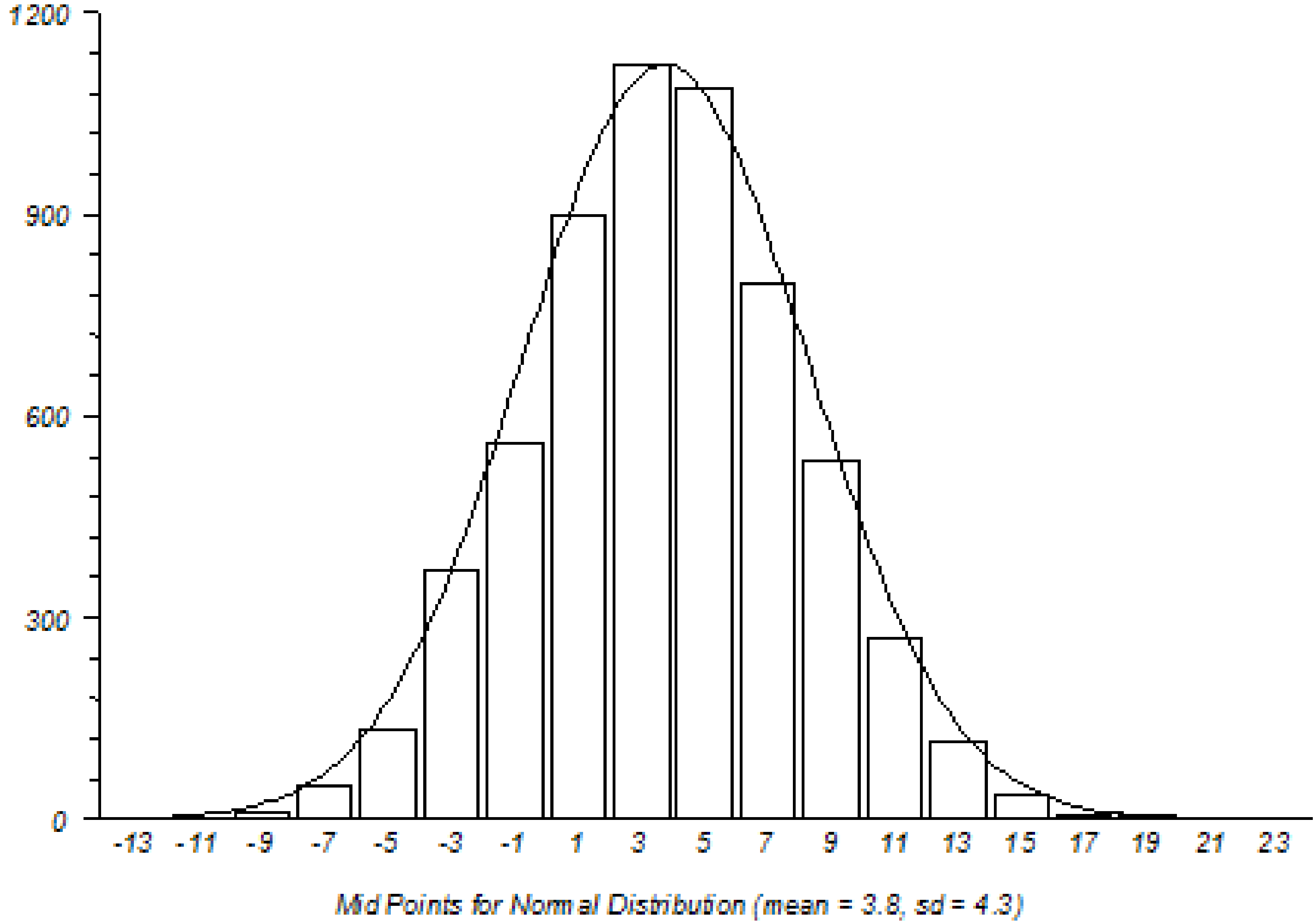
Transformace na percentily

- Zvláštní (standardizovaná) podoba transformace na pořadí
- Používá se při tvorbě norem psychodiagnostických metod a ve výběrových testech

Psychodiagnostická kalkulačka

- Převody různých skórů online.
- Nástroj vyvíjí Hynek Cígler a Martin Šmíra
- http://kalkulacka.testforum.cz/transformace-skoru

Histogram for Normal Distribution (mean = 3.8, sd = 4.3)

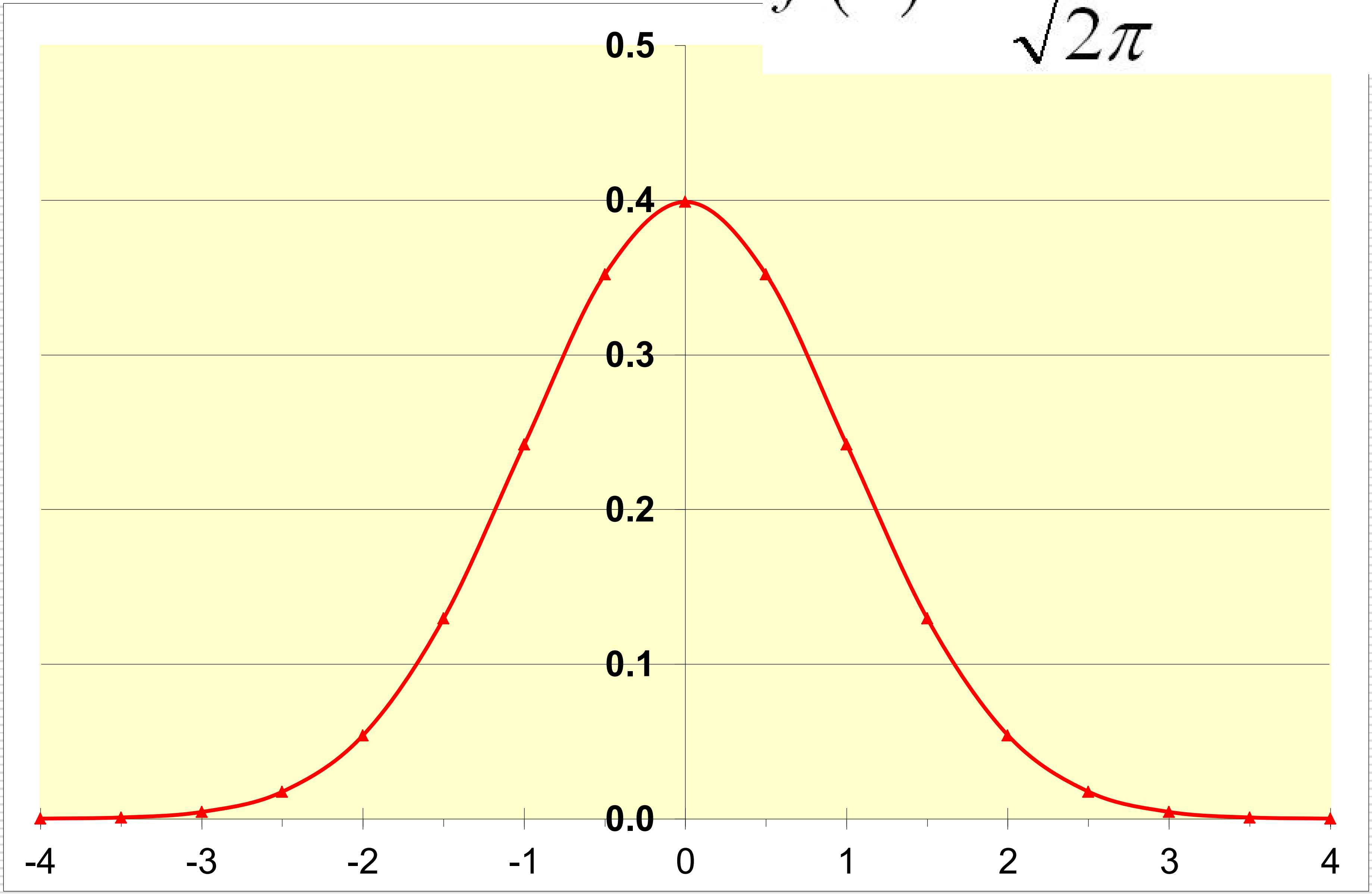


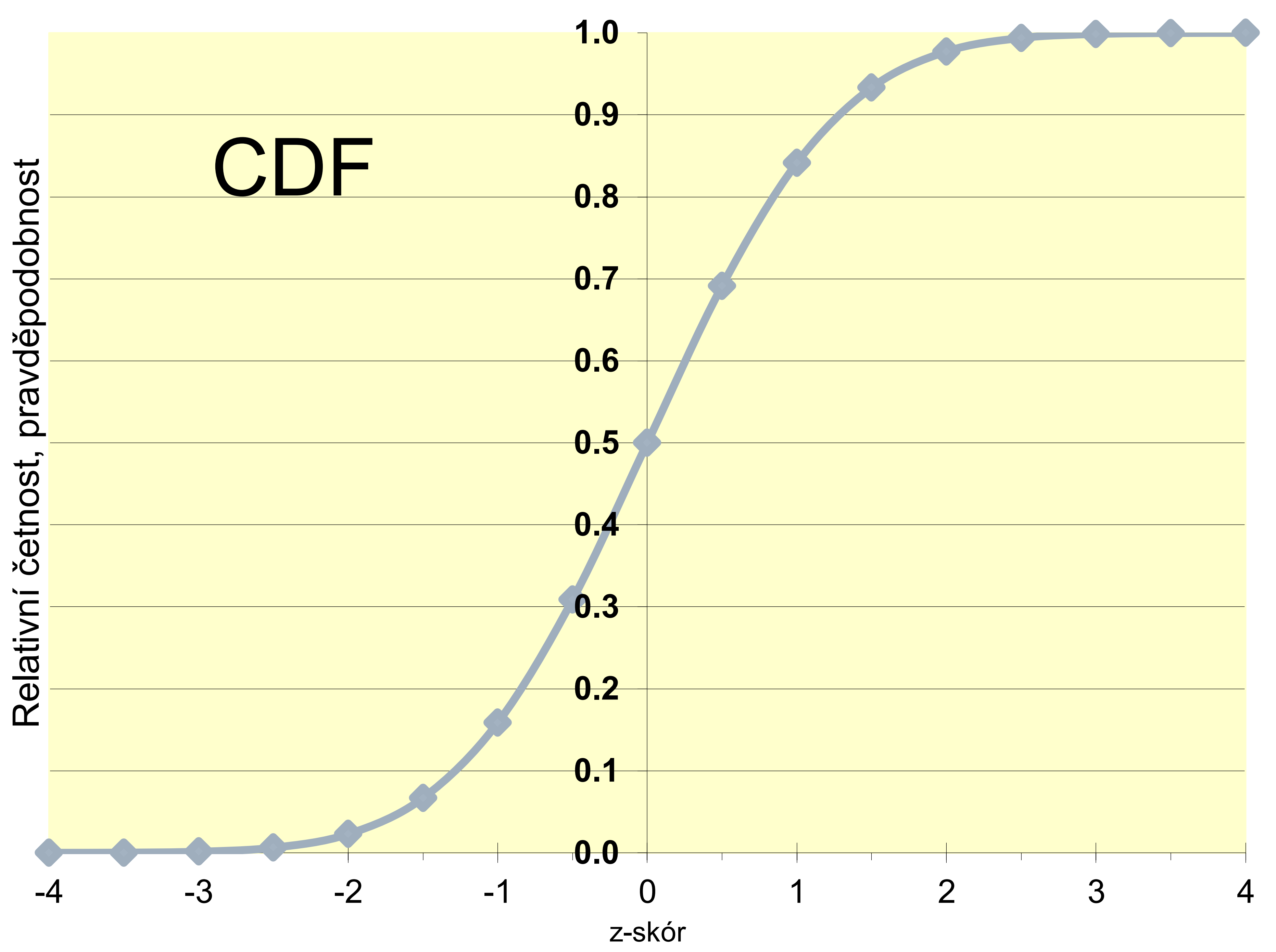
Normální rozložení

Gaussovo, bell-curve

- Rozložení…
  - …náhodných chyb
  - …jevů v přírodě ovlivněných mnoha nezávislými faktory, jejichž efekty se sčítají
- Dlouhá historie – od 17. stol.
  - DeMoivre – sázení
  - Laplace a Gauss – chyby v astronomických pozorováních
  - Quetelet – lidi, *l'homme moyen*, BMI

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(\mu-x)^2}{2\sigma^2}}$$




CDF
Relativní četnost, pravděpodobnost
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4
-3
-2
-1
0
1
2
3
4
z-skór

K čemu/proč normální rozložení?

- Mnoho proměnných je takto rozloženo
  - Můžeme pak hádat, kolik jakých hodnot v populaci je
- Mnoho statistických postupů s normálním rozložením pracuje, předpokládá ho
  - v různých modifikacích a rolích

Vlastnosti normálního rozložení
https://en.wikipedia.org/wiki/Normal_distribution

- Symetrické, unimodální
- Průměr=medián=modus
- V hodnotách M+-SD se mění prohnutí
- Je-li SD = 1, pak plocha pod křivkou je 1
- Zešikmení (skewness) je 0
- Strmost (kurtosis) je 3
  - často se prezentuje hodnota K-3
- *Forma, od níž odrážíme popis pozorovaných rozložení*

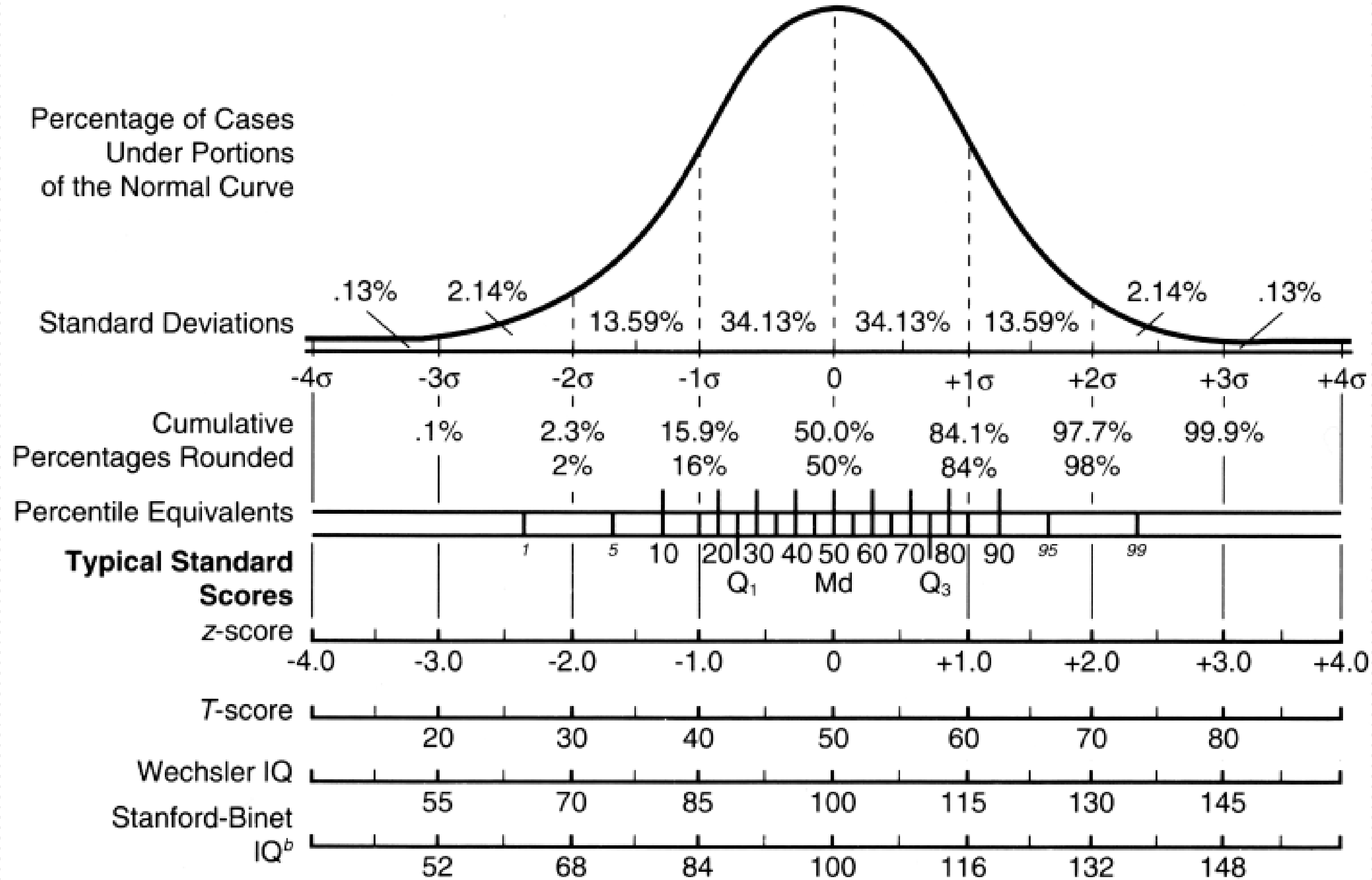
$$Skewness = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{SD(x)}\right)^3$$

$$Kurtosis = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{SD(x)}\right)^4$$

Mnohost normálních rozložení

- Jeden tvar, ale různé umístění na různých škálách (*M*) a různé roztažení (*SD*)
  - http://www.intmath.com/counting-probability/normal-distribution-graph-interactive.php
- Standardní normální rozložení: *N*(0,1)
  - tj. převedení normálně rozložené proměnné na z-skóry

Kvantily standardního normálního rozložení $N(0;1)$
alias oblasti pod křivkou normálního rozložení



upraveno dle Glass, Hopkins, s. 88

Kvantily přesněji v MS Excel

- NORM.S.DIST(*z*;1) – udává percentil odpovídající zadanému z-skóru, tj. kolik lidí má z-skór roven z nebo menší
  - Procento lidí v rozmezí z-skórů = NORM.S.DIST(vyšší z;1) – NORM.S.DIST(nižší z;1)
- NORM.S.INV(*p*) – udává z-skór odpovídající zadanému percentilu

Bez toho S. poskytují funkce tutéž informaci pro normální rozložení s jiným M a SD

Starší a napříč tabulkovými kalkulátory kompatibilní funkce jsou NORMSDIST(z) a NORMSINV(p).

Kvantily přesněji postaru

Table A-1 The Standard Normal Distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985

Jak usoudíme, že naše pozorované rozložení je (přibližně) normální?

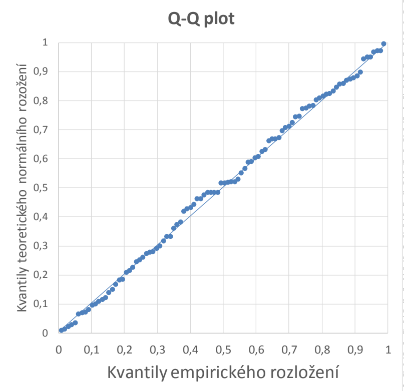
1. TVAR
   - symetrická zvonovitost – histogram, Q-Q plot
   - zešikmení – přibližně 0 (ne víc než +-1)
   - strmost – přibližně 0 (ne víc než +-1)

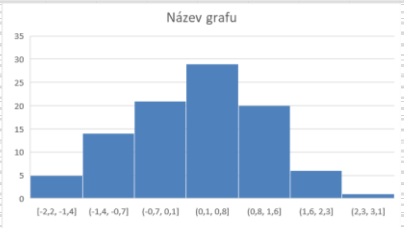
2. SPOJITOST
   - musí být smysluplné předpokládat, že i když máme v datech diskrétní hodnoty, měřená veličina je spojitá

Q-Q plot

- Vynesení kvantilů pozorovaného rozložené proti kvantilům normálního rozložení se stejným M a SD.

M=	0	0,174688
SD=	1	0,979585
Náhodné hodnoty vylosované z normálního rozložení	empirický	teoretický kvantil
0,673452757	0,68	0,69468
-0,254517006	0,329	0,330639
0,223402008	0,536	0,519831
-0,122761689	0,371	0,380698
1,416489461	0,917	0,897544
1,026622631	0,793	0,807765
0,133794478	0,474	0,483351
0,598862191	0,659	0,667498
-0,414559664	0,268	0,273744
0,720424236	0,701	0,711274
0,132454997	0,453	0,482806
0,583042836	0,639	0,661611
-1,279209056	0,072	0,068878
-0,343407769	0,309	0,29844
0,823331673	0,731	0,746066
1,344120811	0,907	0,883723
-0,710502126	0,195	0,183094
0,486911437	0,618	0,625034
-2,190650784	0,01	0,007875
0,442038208	0,608	0,607543
1,272130862	0,876	0,86871
0,934984616	0,762	0,781167
2,046030251	0,979	0,971955
-0,368135391	0,298	0,289743
1,115791575	0,835	0,831653
0,504007548	0,628	0,631633
-0,401308396	0,278	0,278266
0,394833552	0,587	0,588907
0,080686106	0,422	0,461776
1,079465736	0,814	0,822161
1,00762038	0,783	0,802418
-0,562740622	0,226	0,225786
-0,025498409	0,381	0,419037
0,133441721	0,463	0,483207
1,785956729	0,938	0,949999
-0,454901535	0,257	0,260206
-0,891385612	0,154	0,138233
0,391660863	0,577	0,587646
0,214783568	0,505	0,516324
1,787121061	0,948	0,950122
0,00692787	0,402	0,432011
-1,596020511	0,051	0,035333
0,133948804	0,484	0,483413
0,213693733	0,494	0,515881
0,221943413	0,525	0,519237
-0,500071993	0,237	0,245467
2,69781337	0,989	0,994998
1,170514769	0,845	0,845323
0,299929521	0,556	0,550867
-1,085546216	0,113	0,099134
0,914122474	0,752	0,774829
1,056571999	0,804	0,81601
0,243269068	0,546	0,527907
0,030188265	0,412	0,441364
-1,308878168	0,061	0,064951
1,220983337	0,865	0,857262
-1,70815316	0,041	0,027298
1,320358531	0,896	0,878908
0,757611381	0,711	0,724102
0,598803814	0,649	0,667476
-1,262951022	0,082	0,071106
-0,596598396	0,216	0,215535
0,941757153	0,773	0,783203
0,709049397	0,69	0,707294
-0,396954468	0,288	0,279759
0,428259	0,597	0,602127
0,114739125	0,443	0,475601
-0,179767269	0,35	0,358735
1,979437525	0,958	0,967289
-0,620749525	0,206	0,208391
-1,001660487	0,134	0,114902
-0,252897967	0,34	0,331238
-0,291267161	0,319	0,317156
1,713195716	0,927	0,941859
0,818586655	0,721	0,744512
1,085079534	0,824	0,82365
-1,104037616	0,103	0,095882
-0,714605656	0,185	0,181985
1,294290412	0,886	0,873467
-1,798818244	0,03	0,021971
2,019794131	0,969	0,970188
-0,974516126	0,144	0,120367
0,08260445	0,432	0,462553
0,610607827	0,67	0,671842
1,213575196	0,855	0,85555
0,216392553	0,515	0,516979
-1,20094937	0,092	0,080114
-1,040239357	0,123	0,107442
-0,841458149	0,164	0,149792
0,336558733	0,567	0,565624
0,900567846	0,742	0,770656
-0,006664109	0,391	0,426563
-0,770630746	0,175	0,167267
-2,033070716	0,02	0,012105
-0,145039222	0,36	0,372064
-0,483434333	0,247	0,250843





Jak usoudíme, že naše pozorované rozložení je (přibližně) normální?

1. TVAR
   - symetrická zvonovitost – histogram, Q-Q plot
   - zešikmení – přibližně 0 (ne víc než +-1)
   - strmost – přibližně 0 (ne víc než +-1)

2. SPOJITOST
   - musí být smysluplné předpokládat, že i když máme v datech diskrétní hodnoty, měřená veličina je spojitá

Statistické zkratky a značky

- různé systémy, je třeba dobře popisovat
- N, n = velikost vzorku, podvzorku(skupiny)
- X_i = skór i-té osoby u proměnné X
- x_i = deviační skór, odchylka od průměru
- M, m, $\bar{X}$ = průměr
- SD, s = směrodatná odchylka
- VAR, s^2 = rozptyl
- Sumační operátor, např. zkráceně

AJ: statistical notation, sample size, subsample, score, deviation score

Yo Momma
is so Mean,
she has no
Standard Deviation