PSY117/454
Statistická analýza dat v psychologii
Přednáška 3

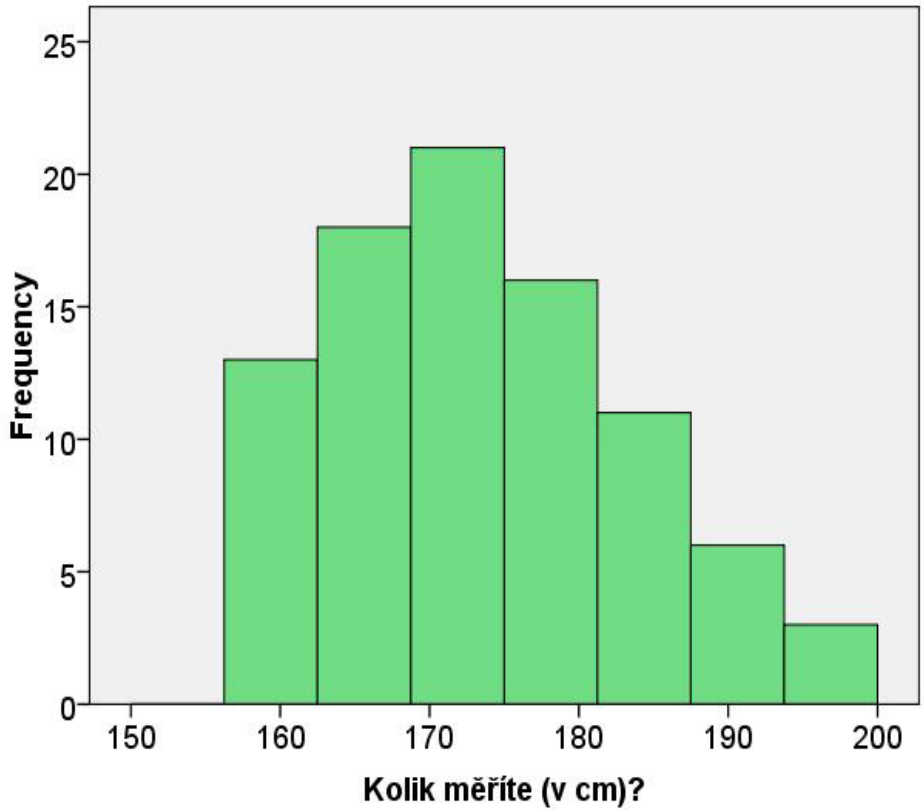
MÍRY CENTRÁLNÍ TENDENCE A VARIABILITY

He uses statistics as a drunken man uses lampposts – for support rather than illumination.

Andrew Lang

Kolik měříte?

Výška [cm]	**n**	**%**
153,75 < x <= 160,25	8	8,9
160,25 < x <= 166,75	16	17,8
166,75 < x <= 173,25	25	27,8
173,25 < x <= 179,75	14	15,6
179,75 < x <= 186,25	16	17,8
186,25 < x <= 192,75	5	5,6
192,75 < x <= 199,25	4	4,4
ChD	2	2,2



Nedalo by se rozložení hodnot proměnné popsat úsporněji než pomocí tabulky četností, histogramu?

Kde na měřené škále se data nalézají?

UKAZATEL CENTRÁLNÍ TENDENCE

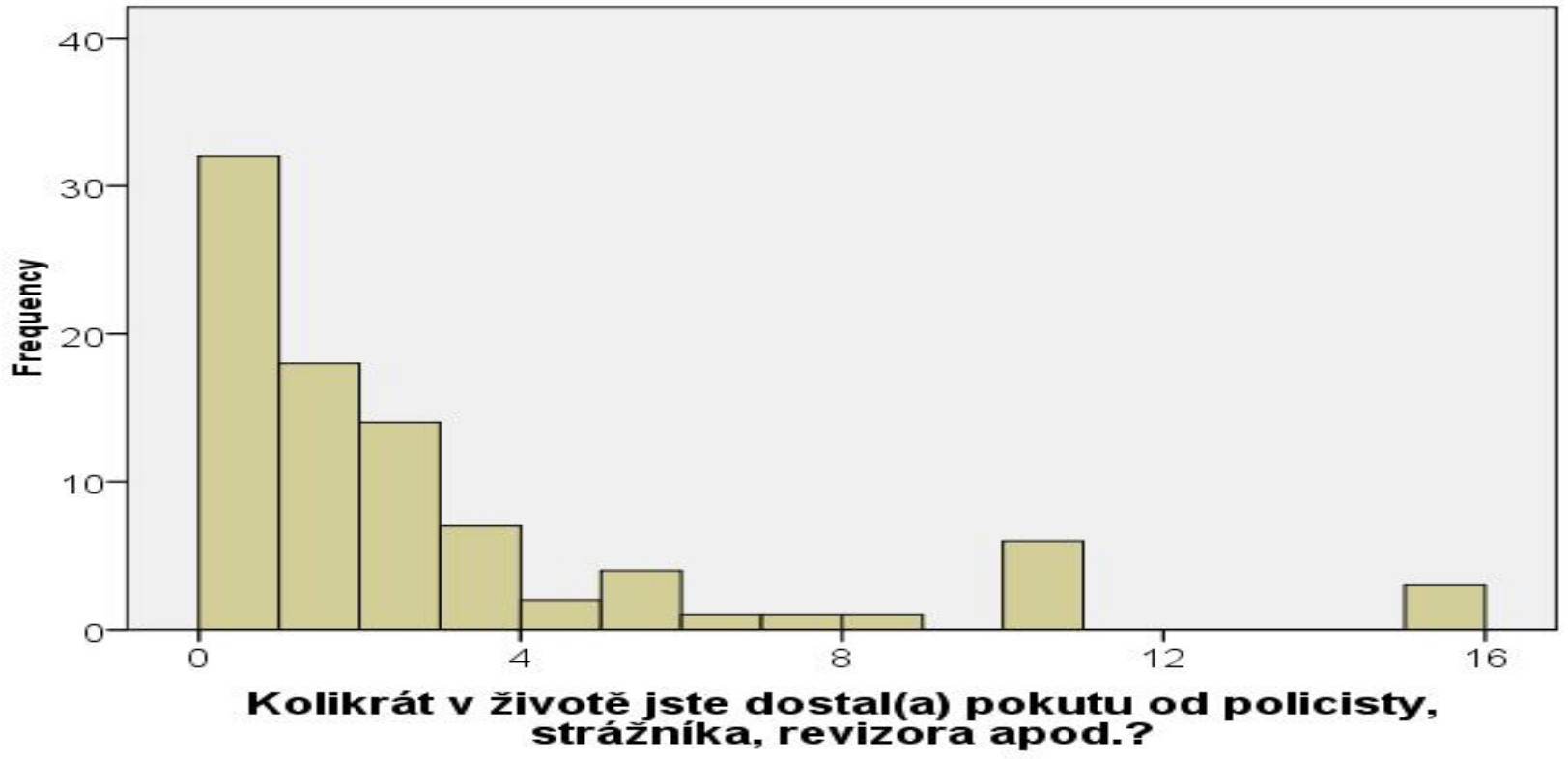
Jak moc jsou na ní rozptýlená?

UKAZATEL VARIABILITY

Centrální tendence (=střední hodnoty, umístění)

- CT je jeden údaj, jímž se snažíme popsat rozložení četností jedné proměnné
- Jeho kouzlo i zrádnost je právě v tom, že je to právě jeden údaj.
- CT udává průměrnou, typickou, reprezentativní, *očekávanou* hodnotu
  - Co se tím míníme, záleží na tom, jakou míru CT se rozhodneme použít

AJ: measures of central tendency, of location



	s 15	bez 15
Průměr	2,48	2,05
Medián	1,00	1,00
Modus	0	0

Modus, medián a průměr

Modus - kategoriální typická hodnota

$\hat{X}, Mo$

- nejčastější hodnota, h. s nejvyšší četností
- jediná možnost u nominálních dat, u vyšších úrovní často užitečnou volbou

Medián – pořadová střední hodnota

$\tilde{X}, Md$

- hodnota prvku uprostřed uspořádaného souboru, 50. percentil (P_{50})
- při sudém počtu prvků je mediánem kterékoli číslo z intervalu mezi nejbližší vyšší a nejbližší nižší hodnotou (konsensuálně střed intervalu)
- pořadová data a výše

Aritmetický průměr – deviační, ochylková, momentová střední h.

$\overline{X}, M, m$

- jak ho znáte ze školy
- pouze intervalová a poměrová data
- velmi citlivý na extrémní hodnoty

AJ: mode, median, mean

Míry variability (rozptýlenosti)

- Druhé číslo, jímž popisujeme rozložené hodnot proměnné
- Udává, jak moc či málo jsou data na škále rozptýlená.
  - Malá variabilita = většina hodnot v souboru je stejných nebo velmi blízkých
  - Vysoká variabilita = hodnoty jsou velmi rozmanité (n. rozložení je bimodální)

Rozpětí, rozptyl, směrodatná ochylka

Nominální – entropie – nepoužívá se

Pořadové

- (variační) rozpětí = $X_{max} - X_{min}$ (extrémně roste s velikostí vzorku)
- (inter)**kvartilové rozpětí** = $Q_3 - Q_1$, IQR

Odchylkové (deviační, momentové) ukazatele

- založené na odchylkách od průměru: $x = X - m$
- průměrná absolutní odchylka ($\Sigma|x| / n$) – nepoužívá se
- průměrná odchylka na druhou – **rozptyl**
  - populační ($\Sigma x^2 / n$) vs. výběrový ($\Sigma x^2 / (n-1)$)
  - součet odchylek na druhou = **suma čtverců**
- **směrodatná odchylka** (standardní odchylka)
  - odmocnina rozptylu - návrat k původní jednotce

AJ: measures of variability, entropy, rank-order, range, interquartile range, variance, standard deviation, sum of squares, square, square root

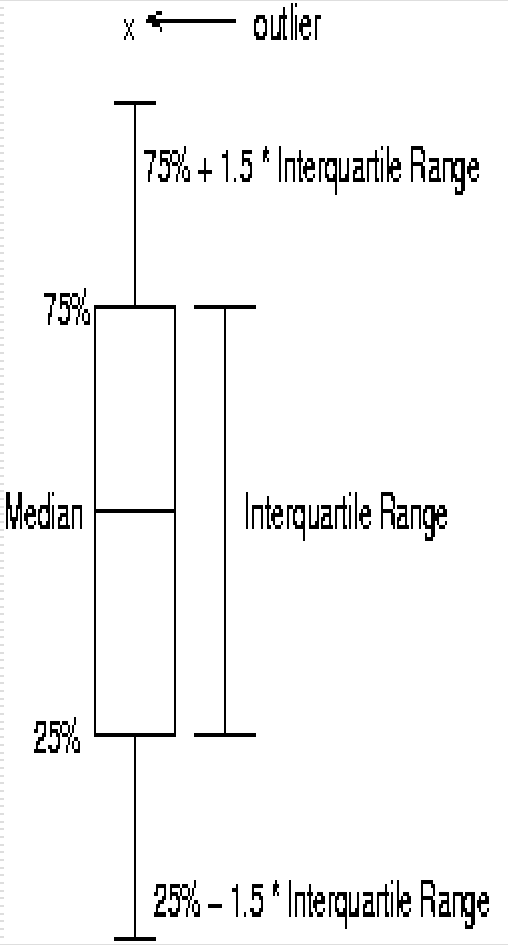
Ukazatele centrální tendence a variability - poznámky

- je třeba je umět spočítat ručně (a zopakovat si práci se sumačním symbolem Σ)
- i vážený průměr
- jak je ovlivní datové transformace přičtení konstanty a násobení konstantou
- vhodnost použití ukazatelů centrální tendence (Hendl s.95)

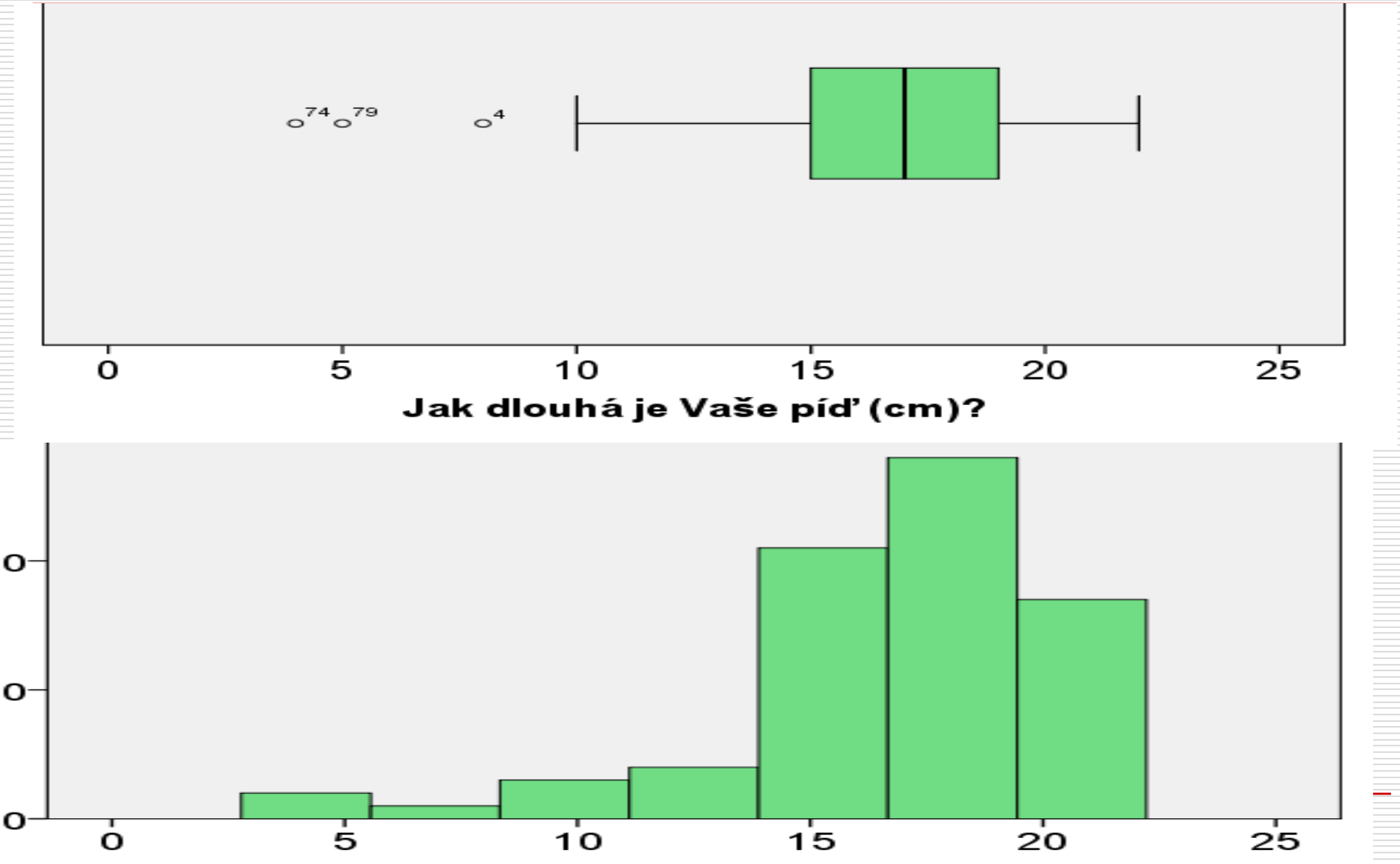
AJ: weighted mean, add, multiply

Boxplot – krabicový graf s anténami

- krabice je od Q_1 do Q_3
- v krabici se značí medián
- antény jsou X_{min} do X_{max}, **maximálně** však 1,5x délka krabice (kvartilového rozpětí)
- hodnoty vzdálenější se značí jako body – odlehlé hodnoty
- hodnoty ještě vzdálenější (více než 3x délka krabice od Q_1 nebo Q_3) jsou někdy označovány jako extrémně odlehlé hodnoty



Boxplot - příklad



Souhrn

- Kategoriální deskriptivy
  - modus, (entropie)
- Pořadové deskriptivy
  - medián, kvartily, percentily (a jiné *kvant**ily***)
  - kvartilové rozpětí
  - grafické znázornění rozložení pomocí pořadových deskriptiv - **BOXPLOT**
- Odchylkové, momentové deskriptivy
  - aritmetický průměr
  - rozptyl, směrodatná odchylka (k=2)
  - zešikmení (k=3) $= \Sigma(X-M)^k / N$
  - špičatost (strmost) (k=4)

AJ: skewness, kurtosis

Volba popisných statistik

- Zvažujeme
  - úroveň měření
  - tvar rozložení – symetrie, normalita
  - cíl studie – pouze popis X usuzování, porovnávání
- Tedy…
  - Je-li cílem především deskripce dat, pak použijeme **POŘADOVÉ** ukazatele. Připojíme-li i odchylkové, nic nezkazíme.
    - ***N, min, Q_1, Md, Q_3, max***
    - **boxplot**
    - pro individuální skóry **percentily**
  - Je-li cílem další usuzování, porovnávání apod., používáme **ODCHYLKOVÉ** ukazatele … pokud to úroveň měření dovoluje
    - ***N, m, s*** (*N, M, SD*)
    - popis rozložení
    - pro individuální skóry **z-skóry**

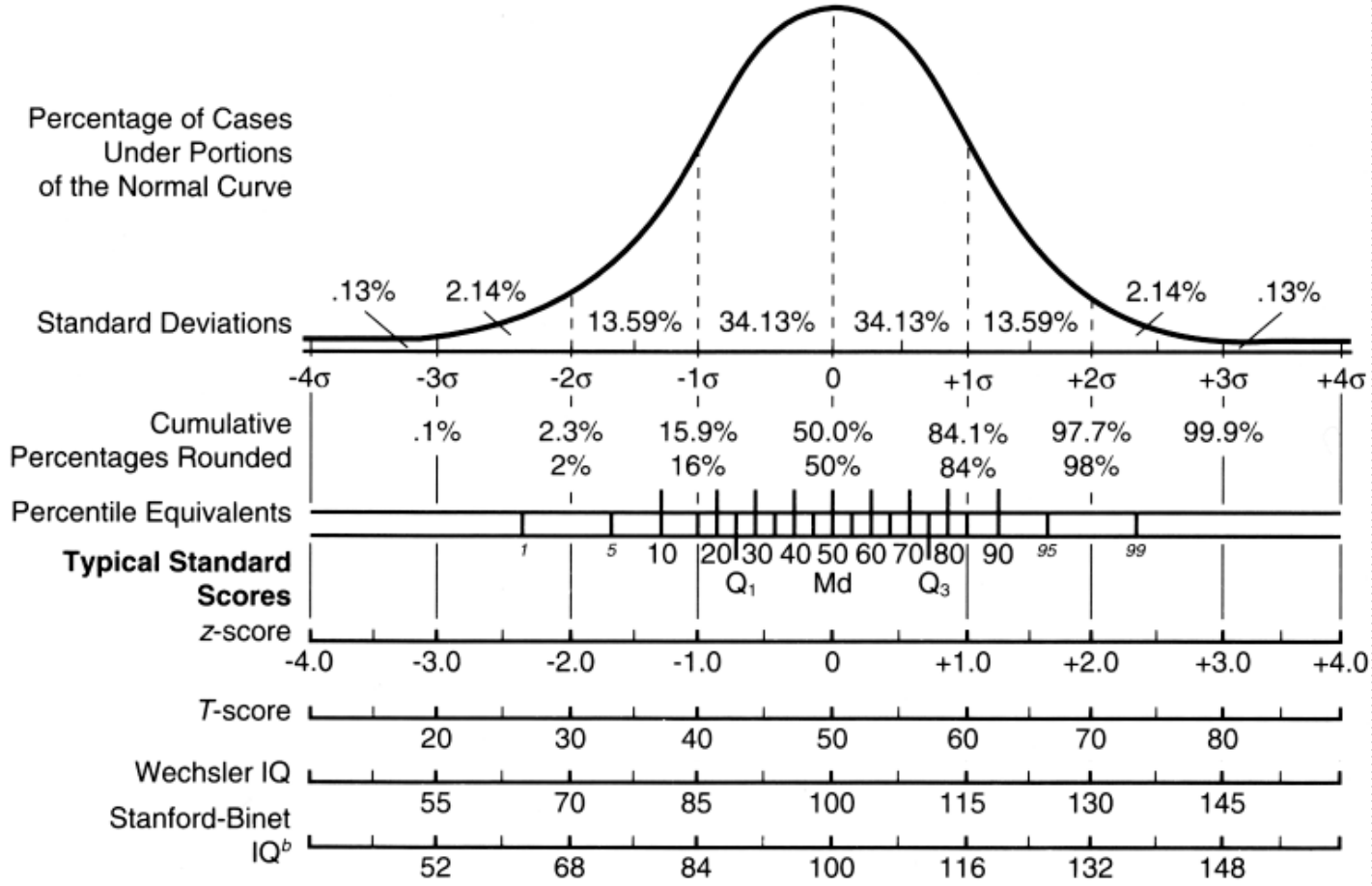
Prezentace deskriptiv ve studiích

- **Vždy**! Bez ohledu na to, jak složité statistiky následují.
- Popis rozložení
  - Obvykle se neuvádějí tabulky četností a jejich grafické podoby, pokud ovšem není cílem studie právě statistická deskripce (např. manuál k testu inteligence).
  - Tvar rozložení obvykle podle potřeby zmiňujeme verbálně („přibližně normální, zleva zešikmené…"). Většinou se řeší pouze normalita a odchylky od ní.
- Obvykle pouze pro proměnné, s nimiž pracujeme (interpretujeme…)
  - Minimální triáda: *N, m, s* (či jejich pořadové ekvivalenty Q_1, *Md*, Q_3, *IQR*)
  - Vhodná pětice: *N*, X_{min}, X_{max}, *m*, *s*
  - Pro puntičkáře/v případě potřeby: N, X_{min}, X_{max}, *m*, *s*, zešikmení, špičatost
- Obvykle na 2 významné číslice / 2 desetinná místa
- V českém textu česky, v anglickém anglicky!
  - Pozor na konvence spojené s jazykem: značky, desetinné tečky, chybějící nuly
- Podoba tabulek je podchycena i normami, např. publikační manuál APA

z-skóry, standardizované skóry

- Transformace dat
  - změna rozložení (např. log, (od)mocniny, Hendl 111)
  - usnadnění interpretace
- Standardizace
  - transformace hodnot tak, aby $m = 0$, $s = 1$
  - **jednotkou měření se stává *s*,** možnost srovnávání různých škál
  - $z_i = (x_i - m) / s$
  - u přibližně normálně rozložených dat o lidech je většina (přes 90%) lidí mezi -3 a 3
  - ze *z*-skórů pak např. T-skóry (m=50,s=10), IQ-skóry (100, 15) apod.
  - **Zásadní pro porozumění normám psychologických testů!**

Oblasti pod křivkou normálního rozložení, percentily



upraveno dle Glass, Hopkins, s. 88

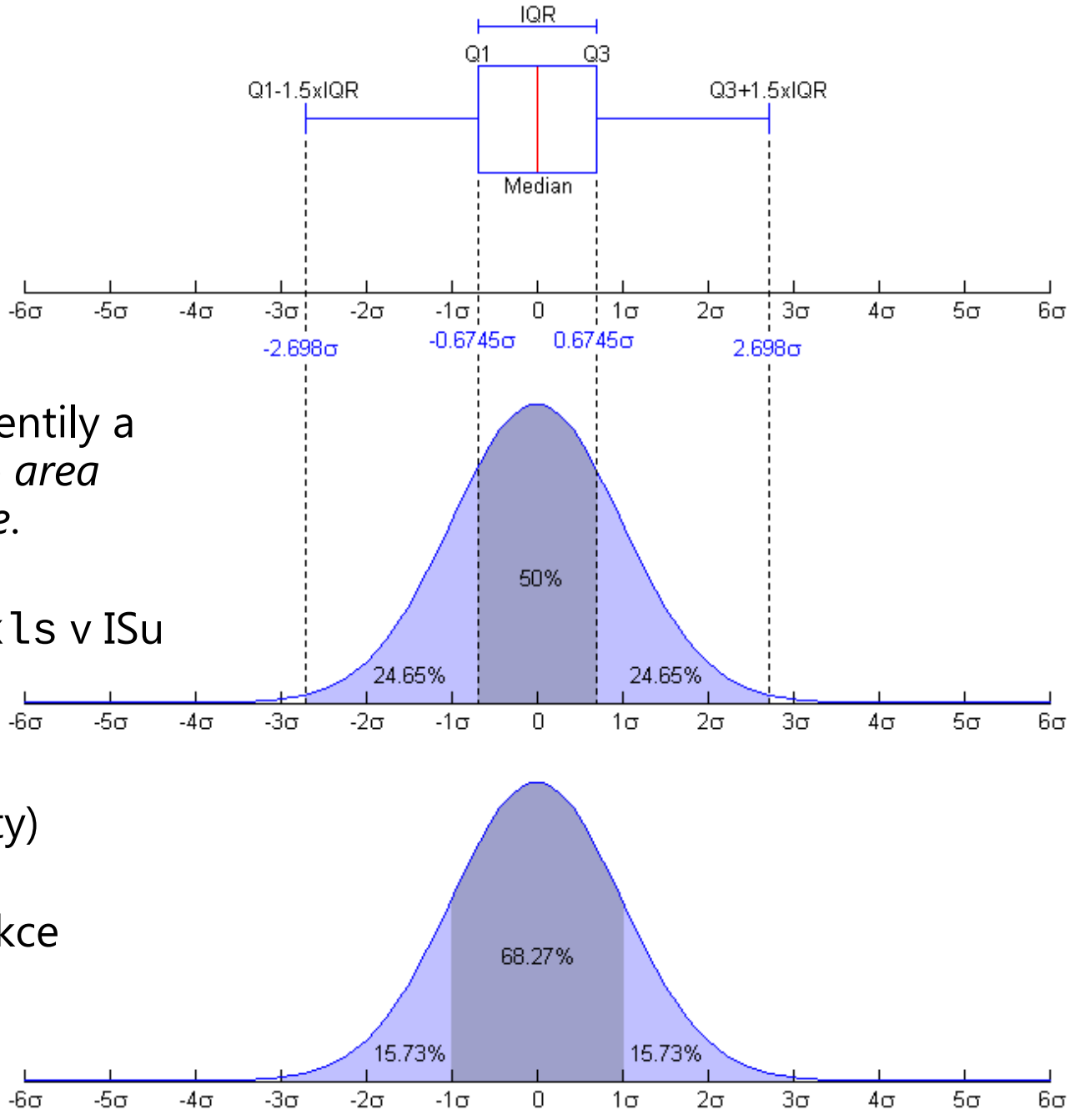
Ze z-skórů na percentily a zase zpět aneb *area under the curve*.

- `Normální rozložení.xls` v ISu

2 zobrazení

- hustota (density)
  jako histogram
- distribuční funkce
  jako graf cum %

Obrázek je z WIkipedie.



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
- s^2 = rozptyl

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

Obecné principy k zapamatování

- Míru shody mezi modelem a daty obvykle konceptualizujeme jako sumu rozdílů mezi modelem a daty umocněných na druhou.
  - nejjednodušší model je průměr a odchylky od něj tvoří rozptyl
- Abstrakce od jednotek měření standardizací
  - často převádíme statistiky na takové škály, které známe