Kapitola 3: Centrální tendence a variabilita

- Někdy chceme více než prezentovat data v tabulkách nebo grafech
- Chceme např. najít typickou hodnotu a v jaké míře se data od této typické hodnoty odchylují
- Použití:
  - k základnímu popisu distribuce proměnné
  - fundamentální k pochopení složitějších analýz

Míry centrální tendence

- Typický klient, typická mzda
- Při analýze dat typické = hledání hodnoty (čísla), která reprezentuje distribuci hodnot proměnné
- 3 nástroje: průměr (mean), medián (median), modus (mode)
- …mají 2 vlastnosti:
  - Sumarizují data
    - př. Průměrně jsme v minulém roce denně 16.4 klientových zakázek)
  - Poskytují společný referenční bod k porovnání dvou skupin dat
    - Př. Průměrná nástupní mzda absolventa BcSW je 2223 $, průměrná nástupní mzda MgrSW je 3112 $

modus

- Nejčetněji se vyskytující hodnota
  - Př. Věk klienta (N=15)
    - 28, 31, 38, 39, 42, 42, 42, 42, 43, 47, 51, 54, 55, 56, 60
- Bimodální distribuce – histogram dva vrcholy
  - Př. Počet let praxe v sociální práci (N=22)
    - 0, 0, 0, 0, 0, 1, 2, 2, 3, 4, 5, 5, 6, 7, 7, 7, 7, 7, 8, 9, 11, 14
- Nejméně předpokladů – použitelný pro každý typ proměnné
- Nepoužívá se často – nejčastější hodnota není vždy nejtypičtější hodnota

medián

- Jsou-li data nejméně ordinální
- Hodnota která dělí případy na dvě půlky
- Př. Počet navštívených terapeutických sezení (n=21)
  - 2,2,2,3,3,4,5,5,7,8,9,10,11,11,14,14,15,16,18,20,41
- Př. Počet navštívených terapeutických sezení (n=24)
  - 1,1,1,1,1,2,2,3,3,3,4,5,6,6,7,8,11,11,13,14,15,17,20
- Klient, který navštívil sezení 41* = extrémní klient
- Medián neovlivněn extrémy - extrémní hodnoty se vyruší
  - př. Bez ohledu zda byl na sezení 41*, 100*, nebo 1000*, medián tuto hodnotu vnímá jako nejvyšší a vyruší ji – medián se nemění

Medián 2

- Máme-li intervaly pak:
  - Vzorec využívající absolutní četnosti:
    - Me = lrl + i * ((0.5N – cf) / f v intervalu)
  - Kde LRL= spodní hranice intervalu pod intervalem obsahujícím medián, i=šířka intervalu, N=celkový počet případů, Cf = kumulativní frekvence pod spodním hranicí, f= frekvence
  - Vzorec využívající relativní četnosti:
    - Me = lrl + i * ((50 – rcf) / rf v intervalu)
  - Kde rcf = relativní kumulativní frekvence pod spodním hranicí, rf= relativní frekvence
  - Př. Medián (C) = 69.5 + (30/40)*10 = 77
  - Medián (X) = 69.5 + (40/40)*10=79.5

Skóre experimentální skupiny (X) (N=300)		
Skóre	Rel.četnost	Kum.rel. četnost
50-59	0	0
60-69	10	10
70-79	40	50
80-89	30	80
90-100	20	100
Skóre kontrolní skupiny (C) (N=200)		
Skóre	Rel.četnost	Kum.rel. četnost
50-59	5	5
60-69	15	20
70-79	40	60
80-89	35	95
90-100	5	100

Aritmetický průměr

- = součet všech hodnot dělený počtem hodnot ($X_{prům} = \Sigma X_i / N$)
- Interval/poměrová pr., nedává smysl pro nominální (vyjma dichotomické, př. Muž=1, žena=0)
- Citlivý na extrémy, vhodný pro symetrickou distribuci a velké N
- Př. Počet navštívených terapeutických sezení (n=21)
  - 2,2,2,3,3,4,5,5,7,8,9,10,11,11,14,14,15,16,18,20,41
  - $X_{prům}$ = (3*2 + 2*3 + 4 + 2*5 + 7 + 8 + 9 + 10 + 2*11 + 2*14 + 15 + 16+18+20+41) / 21 = 220 / 21 = 10,476
- Př. Zkreslenost průměru extrémem:
  - Pokud namísto 41 hodnota 20, pak $x_{prům}$ = (220 – (41-20)) / 21 = 9,476
  - 10,476 – 9,476 = 1 ! Pouze díky jednomu extrémnímu případu

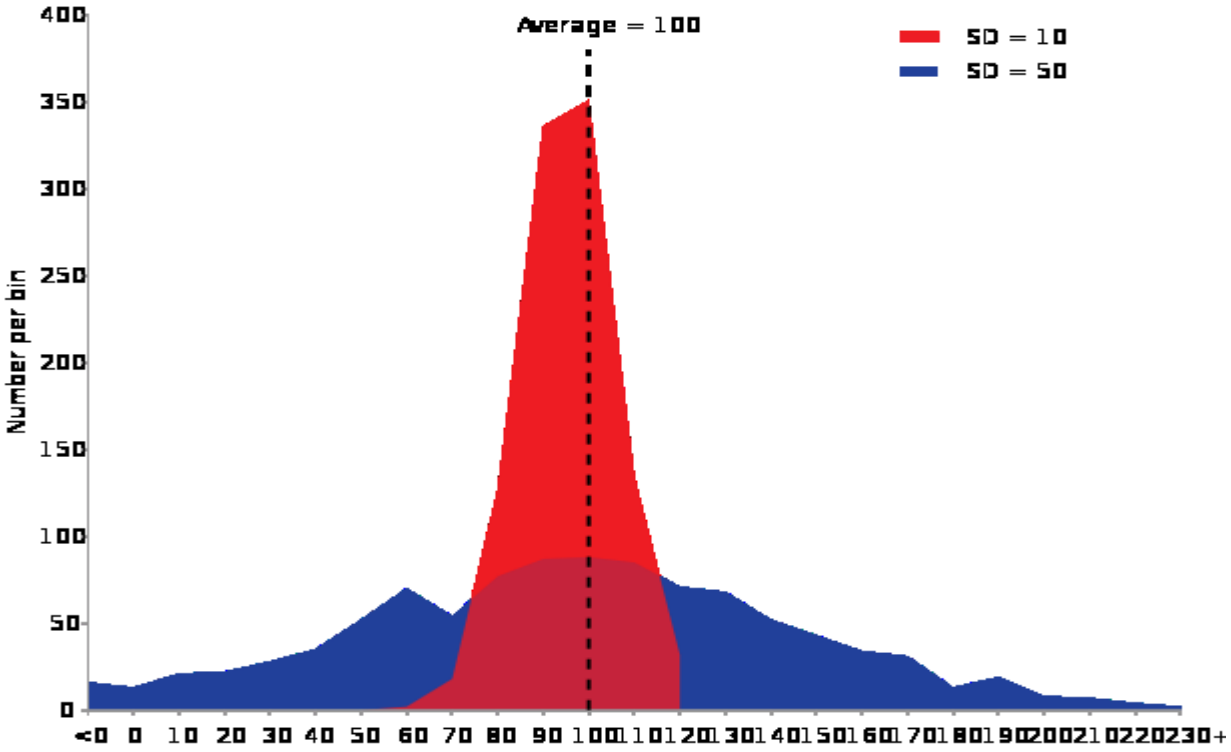
Ořezaný průměr

- = průměr osekaný o horních a dolních 5% případů – vypořádání se s extrémy
- Počítá se jen s hodnotami uvnitř 5 – 95% percentilu

Vážený průměr

- Někdy všechny hodnoty nemají stejnou váhu a je třeba je zvážit
- Př. Pracuji-li na poloviční úvazek, musím svůj výkon vynásobit 2* aby byl porovnatelný s člověkem pracujícím na plný úvazek

Měření rozptýlenosti



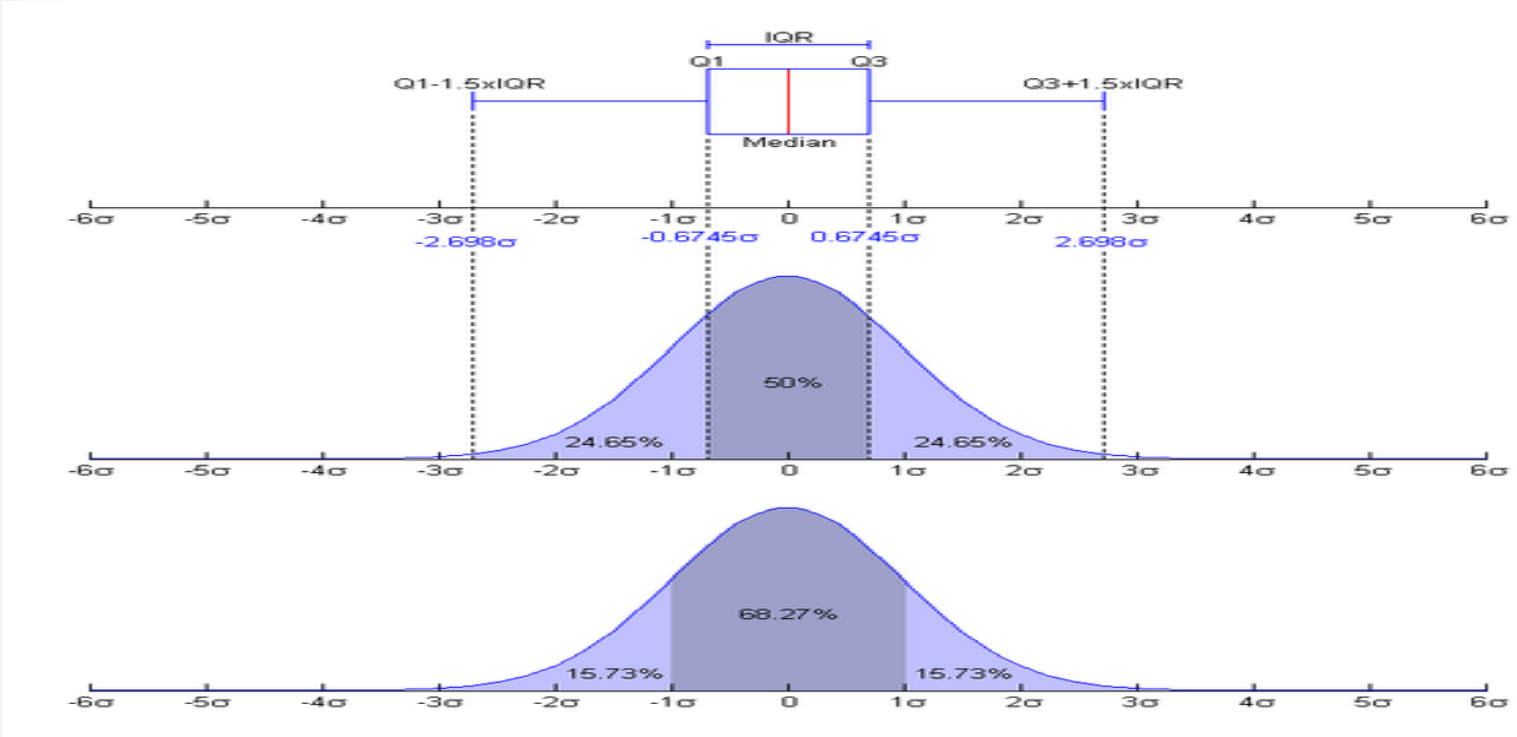
- Distribuce mají stejné průměry ale vypadají jinak - centrální tendence nestačí – je třeba změřit rozptýlenost
- = jak jsou hodnoty rozptýleny od průměru

Variační rozpětí (range)

- = x max – x min + 1
- Př. Věk (x max=35, x min=30)
  - R = 35 – 30 + 1 = 6 (existuje šest potenciálních hodnota věk může nabýt)
- Citlivý na extrémy

Mezikvartilní rozpětí (interquartile range / difference)

- Řeší problém extrémů
- = 75th percentil – 25th percentil



Průměrná odchylka

- $Dm = \Sigma |Da| / N$
- = součet absolutních odchylek (odchýlení každé hodnoty od průměru) dělený počtem případů
- Př. $Dm = (2 + 1 + 0 + 1 + 2) / 5 = 1.2$

Hodnota	Průměr	odchylka
1	3	-2
2	3	-1
3	3	0
4	3	1
5	3	2

Rozptyl (variance)

- Var = součet čtverců individuálních absolutních odchylek od průměru dělený počtem případů
- Var = $(Xi - X\ prům)^2 / N$
- Př. $((-2)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2) / 5 = 2$

Směrodatná odchylka (standard deviation)

- $\sigma = \sqrt{var}$

Př. Příklady distribucí se stejným průměrem ale různou variabilitou

TABLE 3.5 Determining the Standard Deviation of Years of Employment for Agency A

Step 1 (*a*) Value		Step 3 (*b*) Mean		Step 4 (*c*) Deviation from Mean	Step 5 (*d*) Squared Difference from Mean
5	–	6	=	–1	1
5	–	6	=	–1	1
6	–	6	=	0	0
6	–	6	=	0	0
7	–	6	=	1	1
7	–	6	=	1	1
Total				0	Step 6 = 4

Step 7 = $\frac{4}{6}$

= .67 (Variance)

Step 8 = $\sqrt{.67}$

= .82 (Standard deviation)

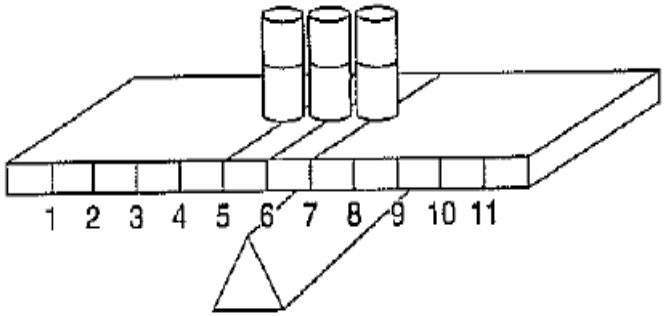


FIGURE 3.2 Variability of Years of Employment for Agency A (from Table 3.5)

TABLE 3.6 Determining the Standard Deviation of Years of Employment for Agency B

Step 1 (*a*) Value		Step 3 (*b*) Mean		Step 4 (*c*) Deviation from Mean	Step 5 (*d*) Squared Difference from Mean
1	–	6	=	–5	25
2	–	6	=	–4	16
4	–	6	=	–2	4
8	–	6	=	2	4
10	–	6	=	4	16
11	–	6	=	5	25
Total				0	Step 6 = 90

Step 7 = $\frac{90}{6}$

= 15 (Variance)

Step 8 = $\sqrt{15}$

= 3.87 (Standard deviation)

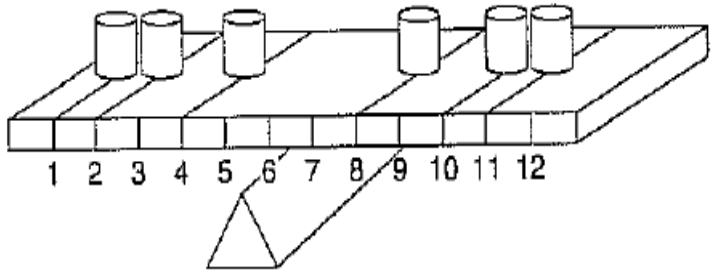


FIGURE 3.3 Variability of Years of Employment for Agency B (from Table 3.6)

Krabicový diagram

- Umožňuje identifikovat distribuci proměnné
- Shora:
  - Extrémní hodnoty
  - Nad horním kvartilem
  - Horní kvartil (75 percentil)
  - Medián
  - Dolní kvartil (25 percentil)
  - Hodnoty pod dolním kvartilem
  - Extrémní hodnoty