

# 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 odchyľují
- Použití:
  - k základnímu popisu distribuce proměnné
  - fundamentální k pochopení složitějších analýz

# Míry centrální tendence

- Typický klient, typická mzda
- Při analýze dat typické = hledání hodnoty (číslo), 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

- Nejčastěji 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



# percentil

- Kolik procent případů se nachází pod konkrétním percentilem
- Užitečné percentily: decily, kvartily, 1st kvartil, 3rd kvartil

# medián

- =50th percentil
- 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 extrémny - 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

## ■ Máme-li intervaly pak:

- Vzorec využívající absolutní četnosti:

- $Me = lrl + i * ((0.5N - cf) / f \text{ v intervalu } )$

- 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:

- $Me = lrl + i * ((50 - rcf) / rf \text{ v intervalu } )$

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

## Ořezaný průměr

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

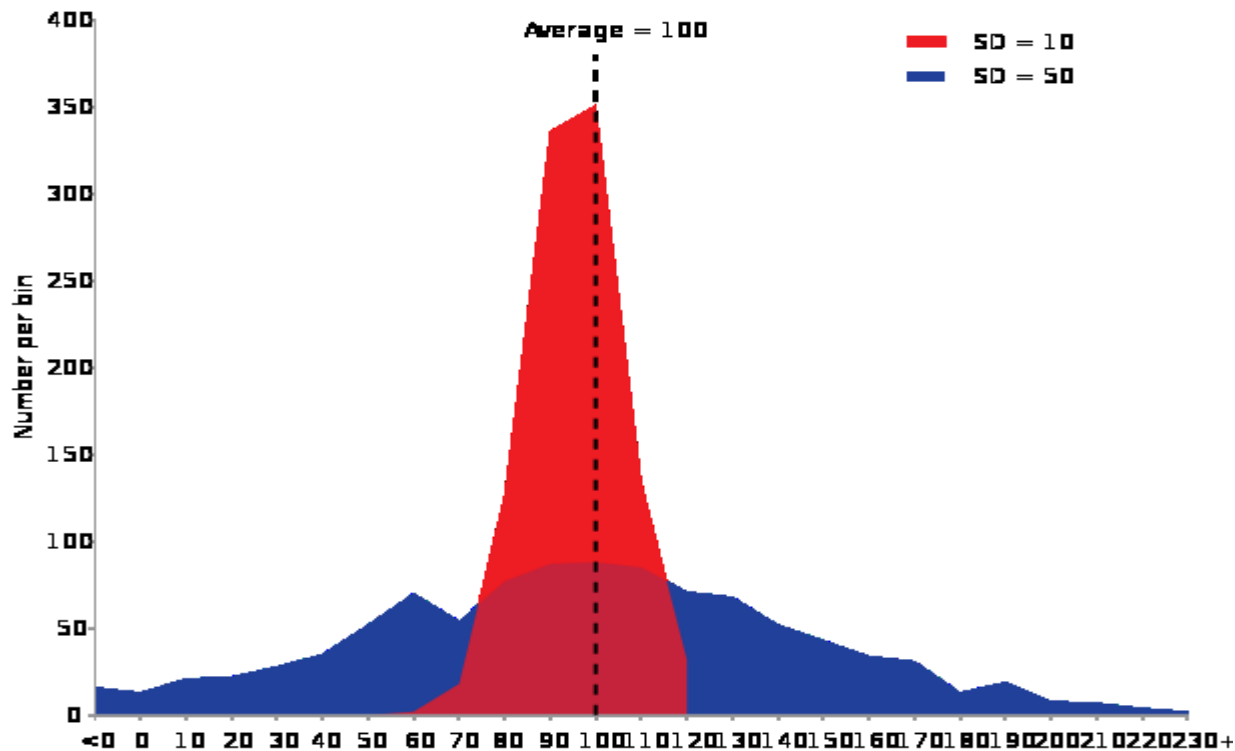




# Vážený průměr

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

# Měření rozptýlenosti



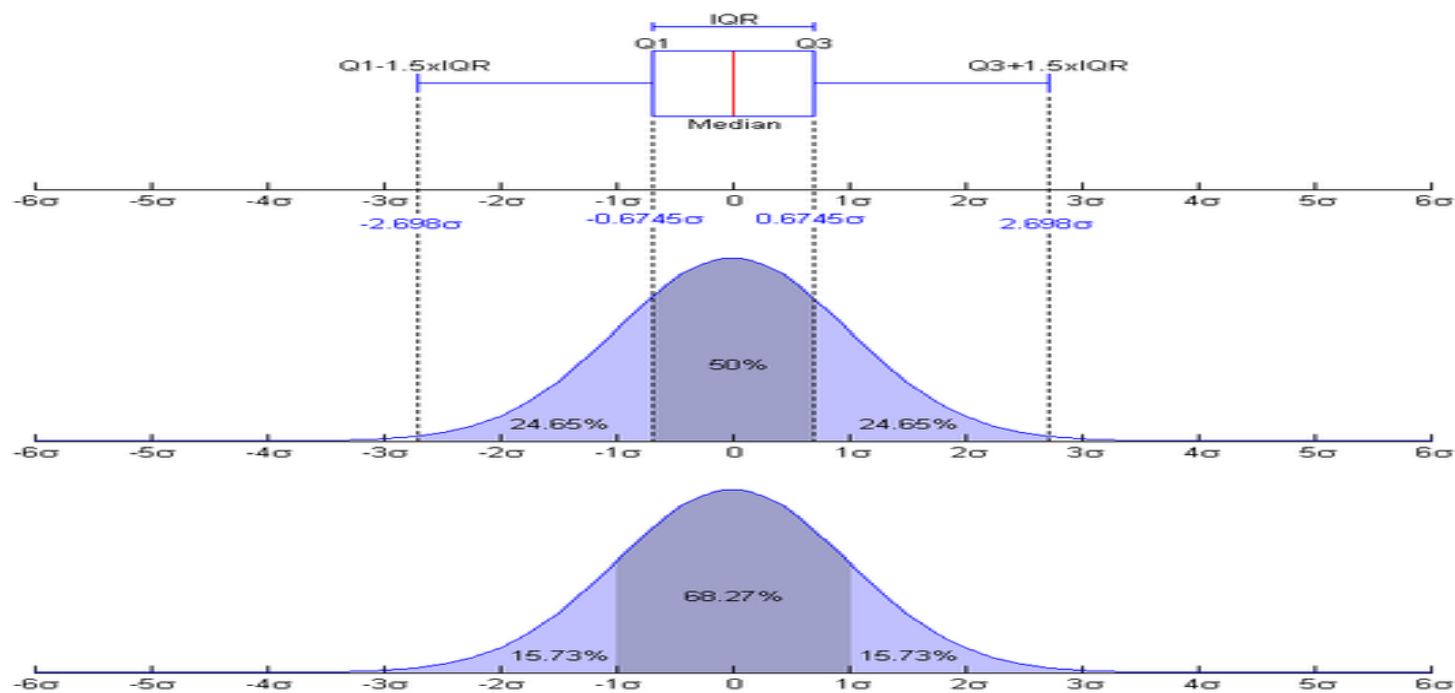
- Distribuce mají stejné průměry ale vypadají jinak - centrální tendence nestačí – je třeba změřit rozptýlenost
- = jak jsou hodnoty rozptýleny od průměru

# Variační rozpětí (range)

- $= x_{\max} - x_{\min} + 1$
- Příklad Věk ( $x_{\max}=35$ ,  $x_{\min}=30$ )
  - $R = 35 - 30 + 1 = 6$  (existuje šest potenciálních hodnot věku, které může nabýt)
- Citlivý na extrémní hodnoty

## Mezikvartilní rozpětí (interquartile range / difference)

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



# Průměrná odchylka

- $D_m = \Sigma |D_a| / N$
- = součet absolutních odchylek (odchýlení každé hodnoty od průměru) dělený počtem případů
- Příklad:  $D_m = (2 + 1 + 0 + 1 + 2) / 5 = 1.2$

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
- $\text{Var} = (X_i - X_{\text{prům}})^2 / N$
- Příklad:  $((-2)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2) / 5 = 2$

# Směrodatná odchylka (standard deviation)

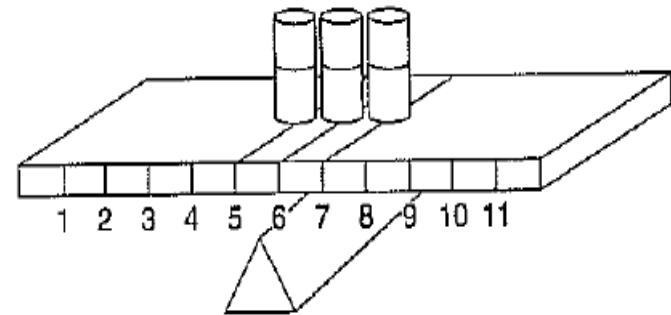
- $\sigma = \sqrt{\text{var}}$



# Př. Příklady distribucí se stejným průměrem ale různou variabilitou

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

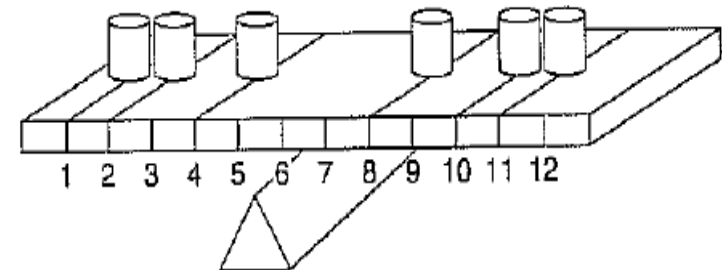
Step 1 (a) Value	Step 3 (b) Mean	Step 4 (c) Deviation from Mean	Step 5 (d) Squared Difference from Mean
5	6	-1	1
5	6	-1	1
6	6	0	0
6	6	0	0
7	6	1	1
7	6	1	1
Total			4
Step 6 =			$\frac{4}{6}$
Step 7 =			.67 (Variance)
Step 8 = $\sqrt{.67}$			.82 (Standard deviation)



**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

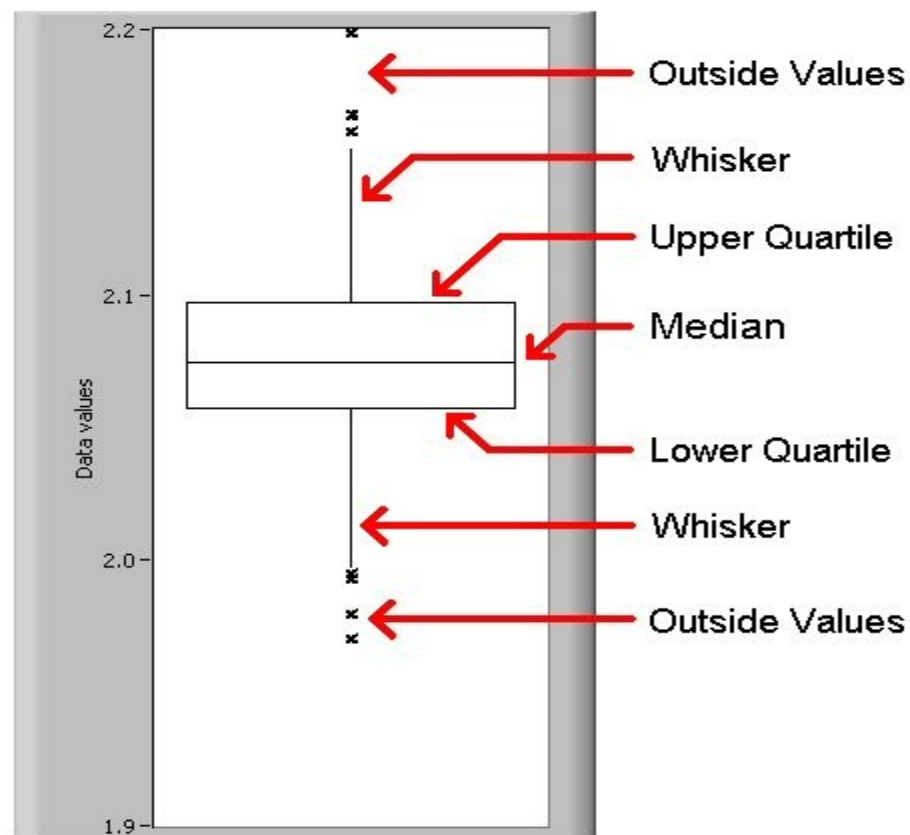
Step 1 (a) Value	Step 3 (b) Mean	Step 4 (c) Deviation from Mean	Step 5 (d) Squared Difference from Mean
1	6	-5	25
2	6	-4	16
4	6	-2	4
8	6	2	4
10	6	4	16
11	6	5	25
Total			90
Step 6 =			$\frac{90}{6}$
Step 7 =			15 (Variance)
Step 8 = $\sqrt{15}$			3.87 (Standard deviation)



**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




## Z skóre - základ

- = z-skor pro hodnotu  $x$  náhodné proměnné představuje jak daleko (**kolik směrodatných odchylek**) od průměru se hodnota  $x$  nachází
- = rozdíl mezi individuální hodnotou ( $X_i$ ) a průměrnou hodnotou ( $X$  prům) relativně k rozptylu distribuce ( $s$ )
  - $Z = (X_i - X \text{ prům}) / s$
- Proto  $Z = 0 =$  průměr ( $\mu$ )
- A také  $1 z = 1 \text{ SD}$ ,  $2 z = 2 \text{ SD}$  atd.
  - Důkaz:  $z = (X - X \text{ prům}) / s = (110-100)/10=s / s = 1$
  - $(120-100)/10=20/10=2$  atd.
- standardizací původních hodnot distribuce vzniká Standardizované normální rozložení  $Z \sim N(0, 1)$

# Z skóre - výklad

- Účel č.1: Zjištění relativní pozice individua k populaci
- Př. Výsledky IQ testu jsou aproximovány (blíží se) normálním rozložením o průměru  $\mu = 100$  a  $\sigma=16$ . Bob skóroval 125. Jak „chytrý“ je Bob vzhledem k ostatním?
  - $Z = (X - \mu) / \sigma = (125-100) / 16 = 1.56$
  - Bob skóroval 1.56 standardní odchylky nad průměrem
  - Jaká část populace skórovala více (nebo méně)? Viz tabulka

- 
- Účel č.2: Porovnání relativních pozic dvou individuů z rozdílných vzorků (populací?)
  
  - Př. Dvě kamarádky Rita a Miriam se účastnily jiných skupin kurzu praxe sociální práce, v 1/2 semestru složili zkoušku, Rita získala 21, Miriam 85 bodů, kdo byl lepší?
    - Srovnat maximální počet bodů v obou testech
      - Rita 21 z 25 = 84 %, Miriam 85 ze 100 = 85 %, je Miriam lepší? Co když je Miriaminých 85 % nejhorší výsledek ve skupině zatímco Ritiných 85 % nejlepší výsledek?
    - nebo srovnat jednotlivé výsledky s výsledky ostatních studentů pomocí z-skóru

- Př. Deborah pracuje jako sociální pracovnice ve studentském zdravotním centru a vede kurzy pro léčbu chronické úzkosti. Uvolnilo se jí místo ve skupině. Do skupiny se přijímá na základě testu „Škála úzkosti A“ ( $\mu = 70$ ,  $\sigma = 10$ ). Pouze studenti kteří dosáhnou min. 80 bodů na škále A mohou být přijati. Deborah se podívala do seznamu potenciálních klientů a zjistila že nejvyššího skóre 78 dosáhla Gina. Deborah však právě dostala doporučení o novém studentovi který trpí úzkostí a potřebuje pokračovat v léčbě. Doporučení také obsahovalo že student Tom dosáhl 66 bodů na jiné škále „Škále B“ ( $\mu = 50$ ,  $\sigma = 12$ ).

- Co může Deborah udělat aby srovnala oba uchazeče a vybrala potřebnějšího?
  - A) Nechat Toma otestovat „Škálou A“
  - B) Zná-li průměr a směrodatnou odchylku obou škal, může porovnat Z-skóry.

■ Řešení:

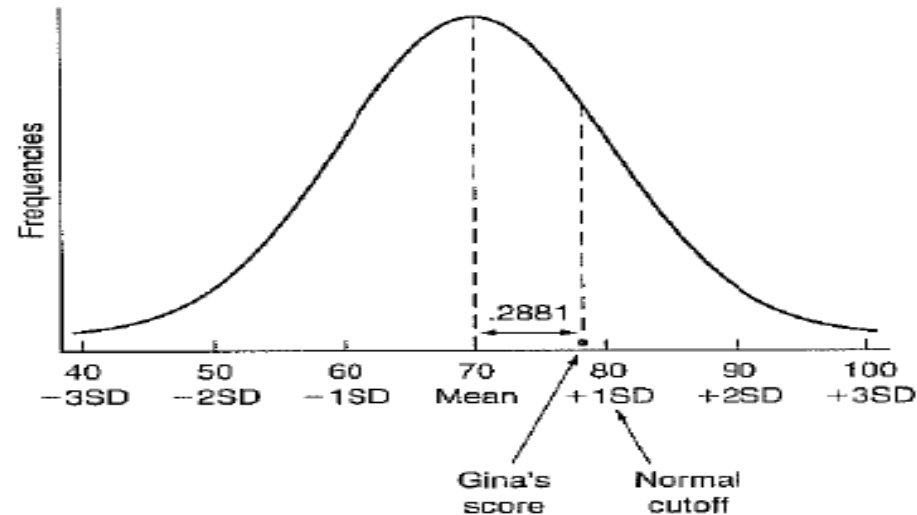
■  $Z_{\text{Gina}} = (78 - 70) / 10 = 0.8$

■ Tabulka Z 0.8 =  
28.81 + 50 =  
78.81 = 79th  
percentil

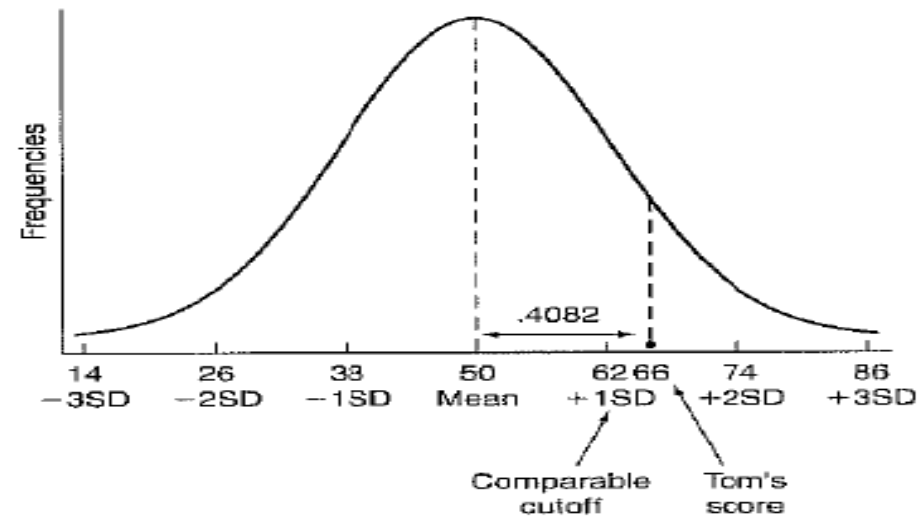
■  $Z_{\text{Tom}} = (66 - 50) / 12 = 1.33$

■ Tabulka Z 1.33 =  
40.82 + 50 =  
90.82 = 91st  
percentil

■ Tom byl vybrán jako  
potřebnější na  
základě relativně  
vyšší úrovně úzkosti



**FIGURE 4.10** Distribution of Scores on Anxiety Scale A (Mean = 70; Standard Deviation = 10)



- Účel č. 3: Odvození syrového skóre z percentilu (z-skóru)
- Sociální pracovnice Lauren chce vytvořit skupinu pro léčbu studentů s vysokou úrovní úzkosti, na základě výsledků z testů na „Škále B“ ( $\mu = 50$ ,  $\sigma = 12$ ), přičemž chce přijmout jen horních 10 procent nejvážnějších případů.
- Řešení:
  - Lauren musí najít mezní bod (cut-off point) pro syrové skóre, který by nejlépe odpovídal 90th percentilu. Studenti nad toto skóre budou přijati, ostatní ne.
  - $X = \mu + z * \sigma$
  - Postup: najít z-skor pro kumulativní pravděpodobnost 90 – 50 = 40
  - Jaké Z odpovídá hodnotě 40?:  $Z = 1.28$
  - $1.28$  (Z-skóre) =  $(x - 50) / 12$   
 $(12 * 1.28) + 50 = x$   
 $65.36 = x$

Odpověď: Pro vstup do skupiny je třeba získat 66 bodů.