8. Základy štatistiky

6MMEH1

Metody ekonomického hodnocení zdravotnických programů

doc. Ing. Peter Pažitný, MSc. PhD.

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Obsah

1. Úvod do štatistiky

- Výskumné metódy v zdravotníctve - využívajú sa najmä metódy štatistiky a epidemiológie
- **epidemiológia**
  - snaha o pochopenie vzniku a vývoja chorôb
  - sleduje distribúciu a determinanty zdravia, chorôb a úrazov
- **štatistika**
  - sleduje variácie v dátach s cieľom získať spoľahlivé výsledky a závery

Postup pri výskume

1) **Definícia problému**
   - vznik hypotézy: napr. Spôsobuje fajčenie rakovinu pľúc?
2) **Vykonanie štúdie a zber dát**
   - dizajn štúdie: deskriptívna, observačná alebo intervenčná (**epidemiológia**)
3) **Popis dát**
   - organizácia a sumarizácia dát (**deskriptívna štatistika**)
4) **Vyhodnotenie výsledkov**
   - výsledky vo vzťahu k hypotéze a ich spoľahlivosť (**prediktívna štatistika**)
5) **Rozhodnutia a závery pre prax**

Načo je dobrá štatistika?

- máme tendenciu zovšeobecňovať anekdotické prípady, často chybne
- štatistika **hovorí o priemernom alebo typickom prípade**

Načo je dobrá štatistika?

- môžeme **popísať skupinu**
  - napr. aký je výskyt diabetu v populácii?
  - štatistika neposkytuje užitočné informácie o jednotlivcoch v skupine, ale popisuje charakteristiky skupiny
  - veľa informácií sa stratí, ale môžeme ju lepšie pochopiť
- osobné príbehy sú často užitočné a zaujímavé, ale niekedy je potrebné poznať aj typické alebo priemerné skúsenosti/charakteristiky

Načo je dobrá štatistika?

- rovnako je užitočná, ak chceme **porovnať dve alebo viaceré skupiny**
  - napr. je tento liek na rakovinu lepší než druhý?
- okrem toho sa snaží odpovedať aj na otázky, či sú dve charakteristiky na sebe závislé a snažia sa robiť **predpovede** (predikcie)
  - napr. je vzťah medzi fajčením a vznikom rakoviny pľúc?

Štatistika

- deskriptívna štatistika
  - **popisuje** vlastnosti skupiny
    - zbieranie, sumarizácia, analýza a prezentácia údajov
- prediktívna štatistika
  - **robí závery a predpovede** o populácii, na základe údajov získaných zo vzorky; patrí sem
    - závery
    - testovanie hypotéz
    - analýza závislostí
    - predikcie

Základné pojmy

- **populácia**
  - celá skupina, o ktorej sa chceme niečo dozvedieť
  - nemusí byť veľká (napr. študenti tohto kurzu)
  - nemusí sa týkať len ľudí (MR, nemocnice,...)
  - môže byť definovaná explicitne alebo implicitne
- **vzorka**
  - časť populácie, o ktorej zbierame dáta, aby sme sa niečo dozvedeli o populácii
  - musí byť dobre vybratá, aby sa výsledky dali generalizovať na celú populáciu

Výber vzorky

- **náhodný**
  - každý jedinec v populácii má rovnakú šancu byť vybratý do vzorky
- **stratifikovaný (kvótny)**
  - jednotlivé podskupiny sú vo vzorke zastúpené rovnakým podielom ako v skutočnej populácii
  - zhodu vzorku s populáciou majú zaistiť kvótne znaky, ktoré sú jednoduché, ľahko identifikovateľné a z hľadiska súboru kľúčové vlastnosti (vek, pohlavie, vzdelanie, kraj, ...)
- **výhodný**
  - vyberieme vzorku tak, aby bolo vykonanie prieskumu/štúdie jednoduché
- **panely**
  - opakovaný výskum, ktorý je vykonávaný na rovnakom súbore osôb v dlhšom časovom horizonte

Opora výberu

- Je zoznam všetkých jednotiek tvoriacich cieľovú populáciu
- Zaobstaranie adekvátnej a predovšetkým aktuálnej opory výberu môže byť veľmi náročné

Výber vzorky (príklad prevalencie AD v ČR v DZR)

- Uvažujme výskum zameraný na prevalenciu Alzheimerovej choroby v ČR v domovoch so zvláštnym režimom (DZR)
- „Populácia“ v tomto prípade sú všetky DZR v ČR 310
- „Opora výberu“ je menný zoznam všetkých 310 DZR, vrátane adries a kontaktov, podľa krajov, vlastníka a veľkosti zariadenia
- „Veľkosť vzorky“ – počet DZR (minimálna veľkosť je pri určených predpokladoch je 76 zariadení DZR)
- „Vzorka“ –podmnožina z celej populácie, na ktorej vykonávam výskum
- „Náhodný výber“- zariadenia by som vyberal úplne náhodne z opory výberu
- „Kvótny výber“ – zabezpečím, aby z každého kraja bolo aspoň 5 DZR, pričom z týchto 5 musí byť aspoň jedno súkromné
- „Výhodný výber“ – Zozbieram data iba z okolitých krajov (Jihočeský, Vysočina, Středočeský)

Dáta

- Analýzou dobrých dát získame užitočné informácie

- Štatisticky správne vykonanou analýzou zlých dát získame nesprávne a škodlivé informácie
  (po anglicky často nazývané aj GIGO analýzy „Garbage In – Garbage Out)

Dáta v štatistike

- **premenná (variable)**
  - čokoľvek, čo môže nadobúdať rôzne hodnoty, ktoré sa dajú "škatuľkovať“
- **konštanta (constant)**
  - niečo, čo nadobúda len jednu hodnotu

Dáta v štatistike

- **kvantitatívne premenné**
  - označujú **množstvo** alebo kvantitu
  - napr. výška, hodnota krvného tlaku
- **kvalitatívne premenné (kategorické)**
  - označujú **kategóriu** alebo charakteristiku
  - napr. pohlavie, typ poskytovateľa, kraj
  - rôzne hodnoty sa ***nedajú navzájom porovnávať*** – čo je viac, sú len odlišné
  - špeciálny typ: **binárna premenná** (dichotómna)
    - nadobúda len 2 hodnoty
    - napr. muž/žena, je/nie je prítomná

Kódovanie premenných

Kvalitatívne premenné (kategorické):

- **nominálne premenné**
  - kategórie bez poradia
  - hodnoty bez súvislosti s váhou alebo hodnotou premennej
  - napr. kraj 1, kraj 2, ..., kraj 14
  - napr. Pohlavie – muž/žena
- **ordinálne premenné**
  - hodnoty reprezentujú poradie, ale rozdiely medzi hodnotami nezodpovedajú skutočným rozdielom
  - napr. poradie DZR podľa veľkosti a kraja – výsledkom je poradie (1, 2, 3, ...) za každý kraj, nie je však zrejmá veľkosť rozdielu
  - napr. Školské známky, hviezdičky

Kódovanie premenných

Kvantitatívne premenné (kardinálne):

- **intervalové premenné**
  - hodnotami sú čísla, pri ktorých poznáme jednotku merania, a vzdialenosť medzi jednotlivými možnými hodnotami merania (použitej škály) je rovnaká, nemajú prirodzenú absolútnu nulu
  - ak má niečo teplotu 0, tak to má stále teplotu
  - napr. teplota
- **pomerové premenné**
  - Rovnaké vlastnosti ako intervalová + prirodzená nula
  - **„nula“–** skutočná absencia meranej vlastnosti
  - ak má niečo dĺžku „nula“, tak nemá žiadnu dĺžku
  - napr. dĺžka

Kódovanie premenných

Nominálna

Ordinálna

Intervalová

Pomerová

2. Deskriptívna štatistika - Sumarizácia a popis dát

- kontinuálne dáta
  - stredné hodnoty: priemer, modus, medián
  - variabilita: variačné rozpätie, rozptyl, štandardná odchýlka
- počty
  - výskyt, pomer

Meranie centrálnych tendencií

- priemer (mean)
  - aritmetický priemer
  - $x = \sum x_i / n$
- medián (median)
  - prostredná hodnota (50-ty percentil)
  - ak sú 2 hodnoty v strede, je to ich priemer
- modus (mode)
  - najčastejšia hodnota

Hodnotenie variability

- centrálne tendencie poskytujú síce užitočné, ale obmedzené informácie

  „ak máte hlavu v mrazničke a nohy v ohni, vaša priemerná teplota je fajn“

1) variačné rozpätie
2) rozptyl
3) štandardná odchýlka

Hodnotenie variability

1. variačné rozpätie (range)

- ***rozdiel medzi najväčšou a najmenšou hodnotou***

Range = Max - Min

Hodnotenie variability

2. rozptyl (variance)

- „štatistický rozdiel množstva disperzie v distribúcii výsledkov okolo priemeru“
- ťažko interpretovateľná, ale: čím väčší rozptyl, tým väčšia variabilita dát
- nepoužíva sa samostatne, ale je vstupným parametrom pre viaceré štatistické výpočty

$$s^2 = \frac{\sum (X - \bar{x})^2}{n - 1}$$

Hodnotenie variability

3. štandardná odchýlka (SD, standard deviation)

- odchýlka = rozdiel medzi individuálnym výsledkom a priemerom distribúcie
- štandardná = typická, priemerná
- ***typický rozdiel medzi individuálnym skóre a priemerom distribúcie***

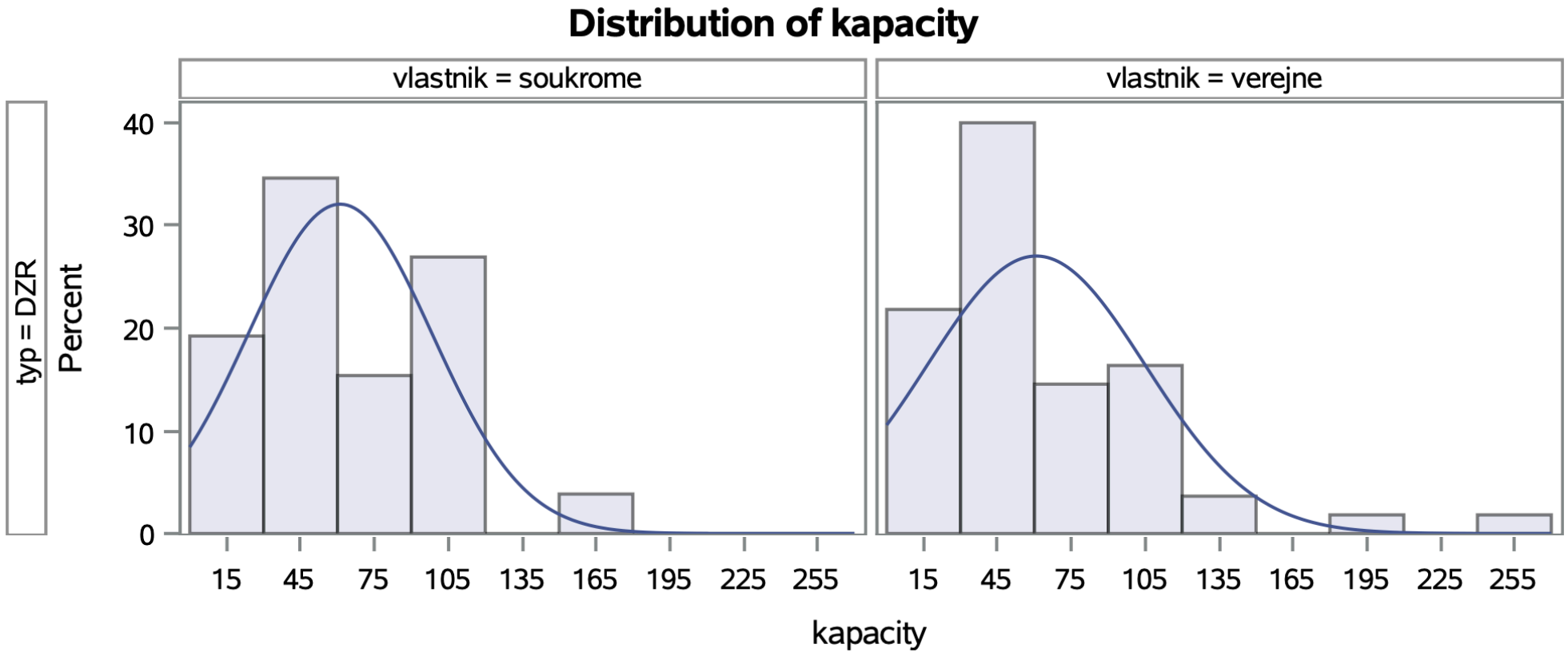
- veľmi užitočný parameter – kombinácia priemeru a SD poskytuje pomerne dobrý obraz o distribúcii výsledkov vo výbere

$$s = SD \quad \sqrt{\frac{\sum (X - x)^2}{N - 1}}$$

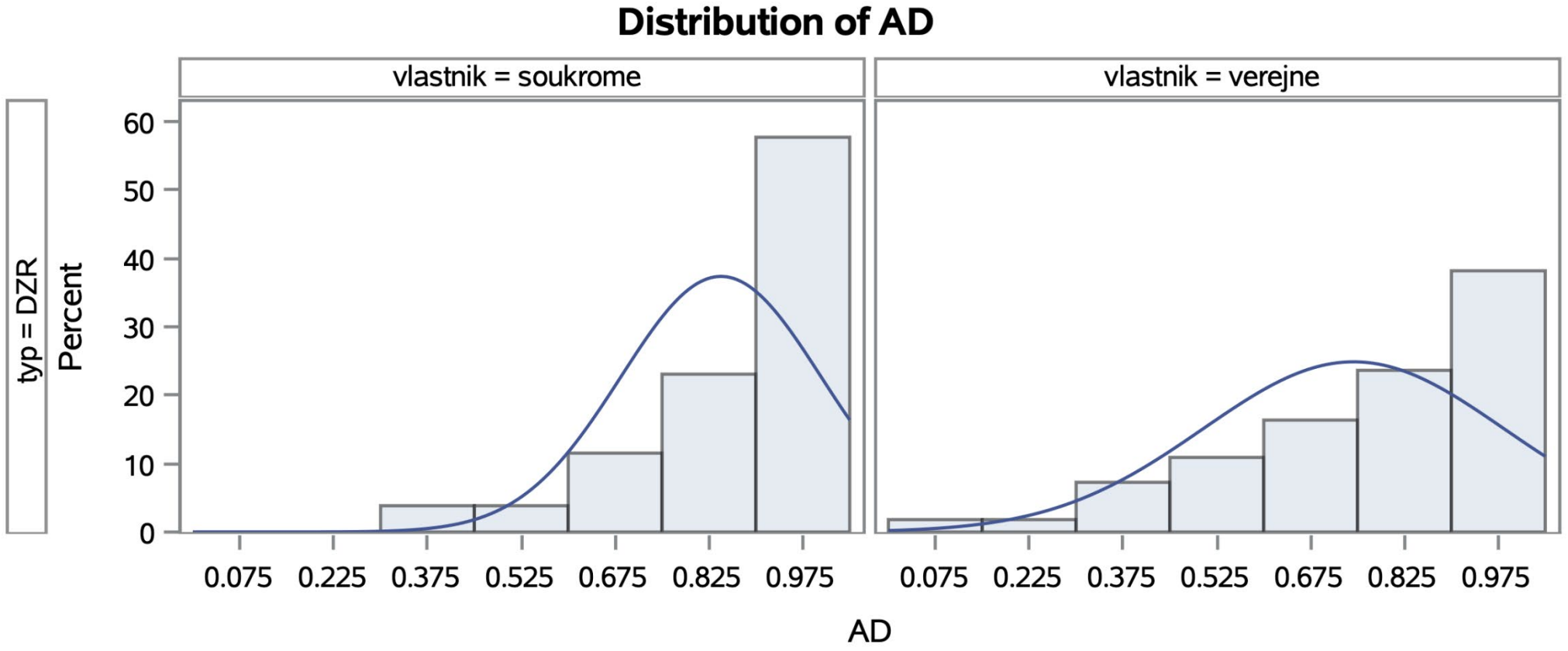
Distribúcia

- priemer, medián a štandardná odchýlka “nehovoria” celý príbeh
- rozdiely v distribúcii dát zobraziteľné tvarom **histogramu**
- Histogram ukazuje distribúciu dát - zobrazuje počty (alebo podiely) pozorovaní pre jednotlivé hodnoty alebo preddefinované skupiny hodnôt
- Boxplot je krabičkový graf – zobrazuje krajné hodnoty súboru (minimum, maximum), priemer a kvartily (25, 50 – medián a 75).

Histogram

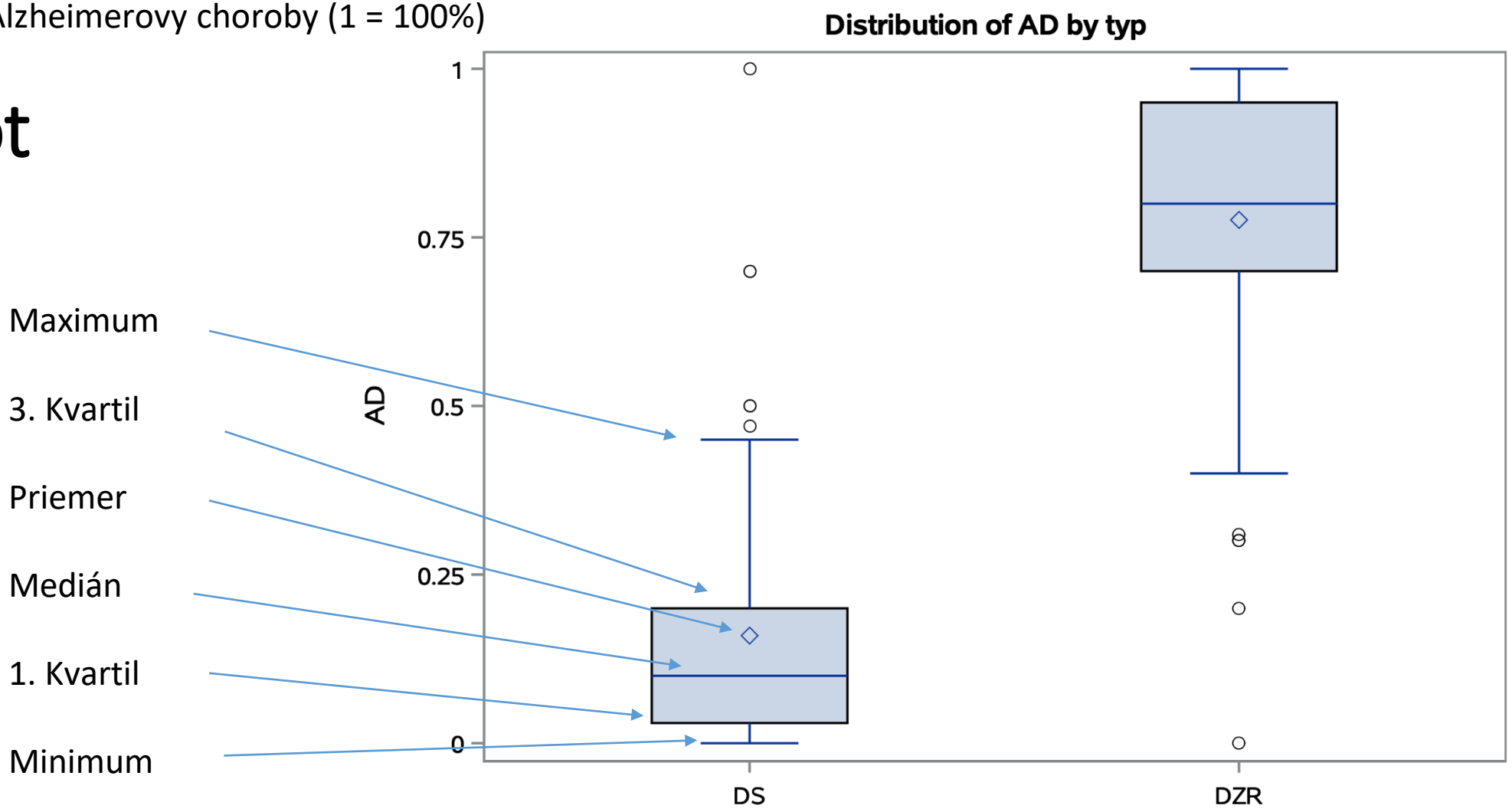

Distribution of kapacity
vlastnik = soukrome
vlastnik = verejne
typ = DZR
Percent
40
30
20
10
0
15
45
75
105
135
165
195
225
255
kapacity

Histogram

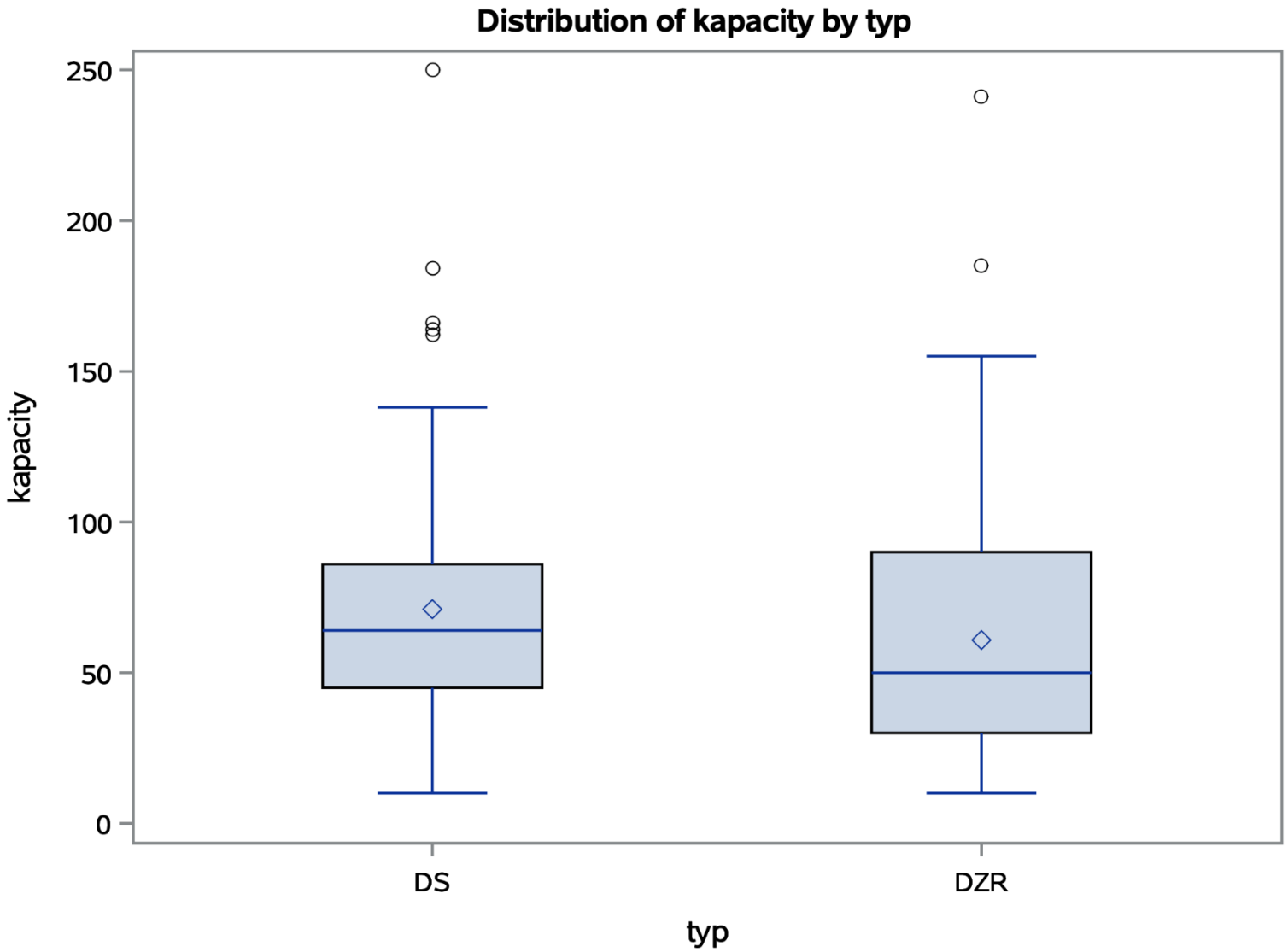

Distribution of AD
vlastnik = soukrome
vlastnik = verejne
typ = DZR
Percent
60
50
40
30
20
10
0
0.075
0.225
0.375
0.525
0.675
0.825
0.975
AD

Prevalence Alzheimerovy choroby (1 = 100%)

Boxplot

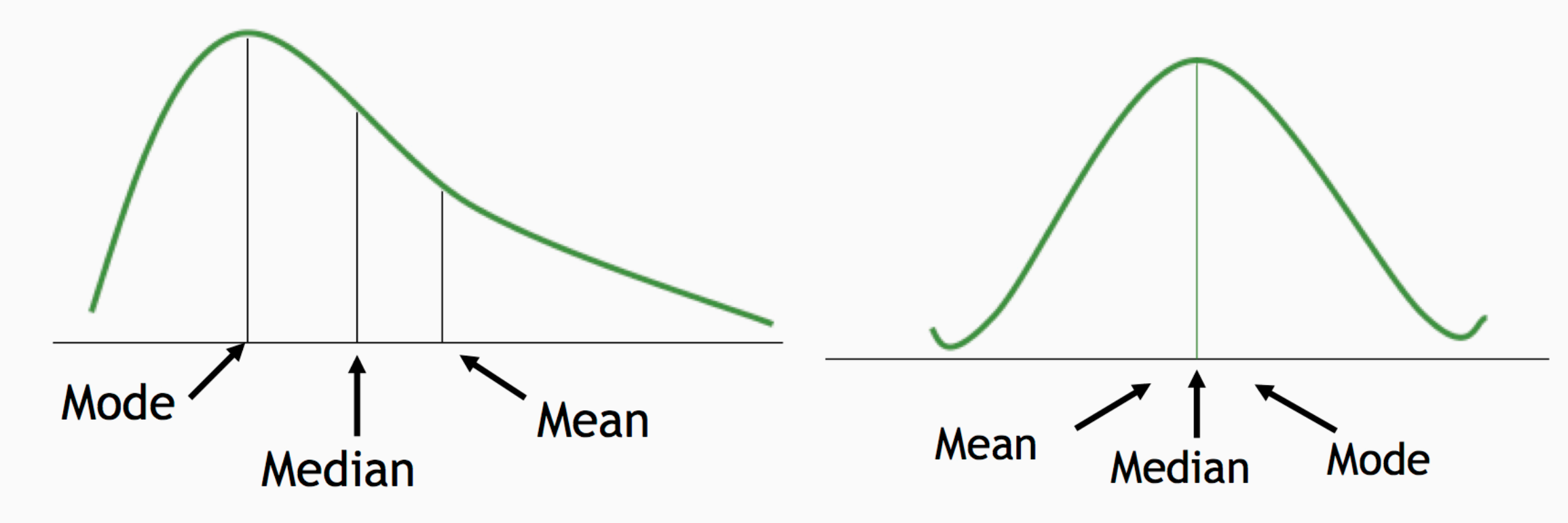


Boxplot


Distribution of kapacity by typ
kapacity
250
200
150
100
50
0
DS
DZR
typ

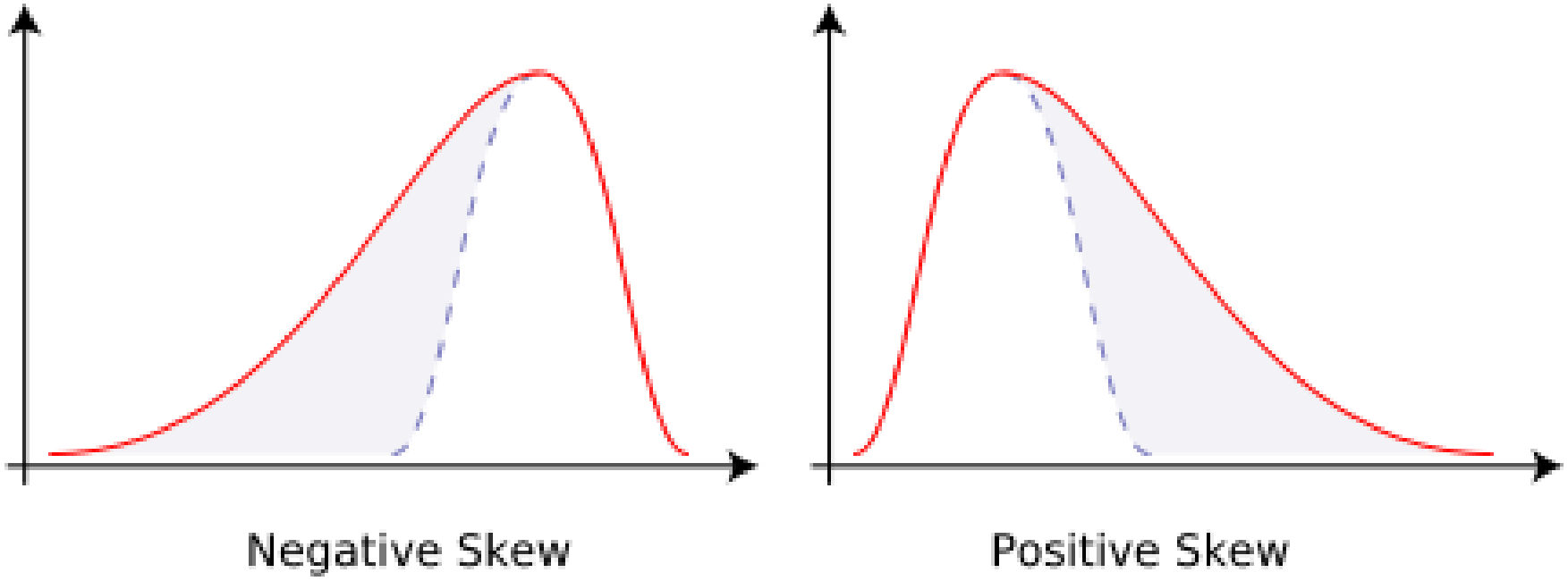
Distribúcia

- symetrická a asymetrická distribúcia (poloha priemeru, modusu a mediánu)



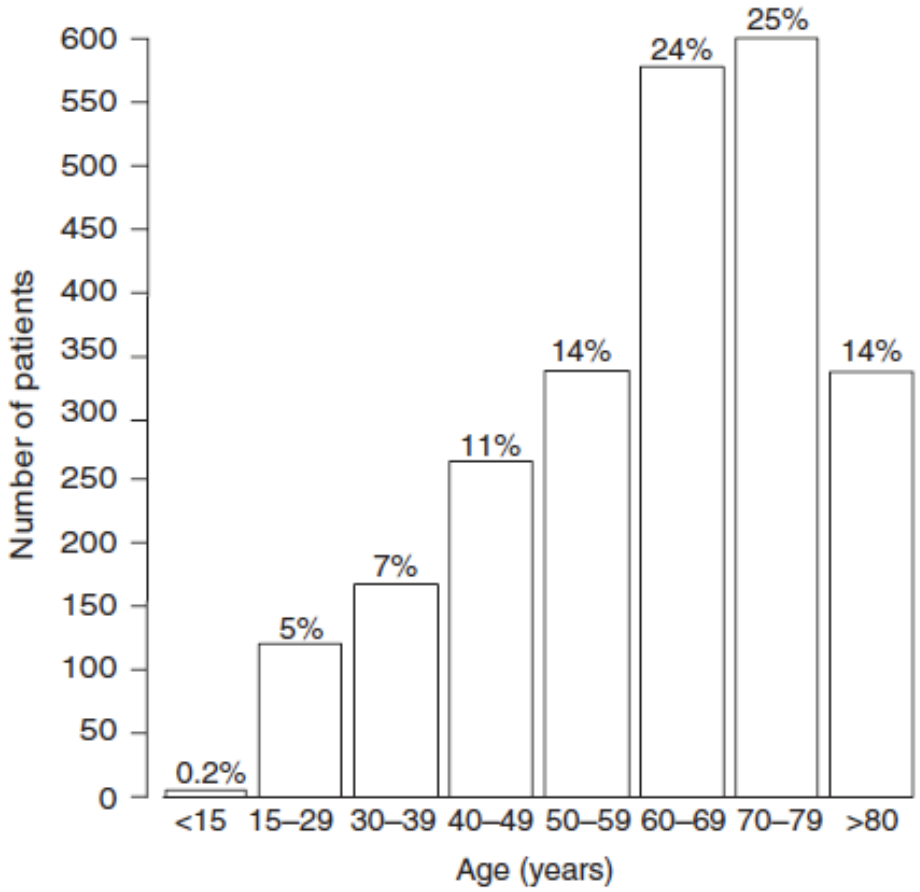
Distribúcia - šikmosť

- šikmosť (skew) - pozitívne alebo negatívne
  - ak je priemer < medián: negatívne šikmá
  - ak je priemer > medián: pozitívne šikmá

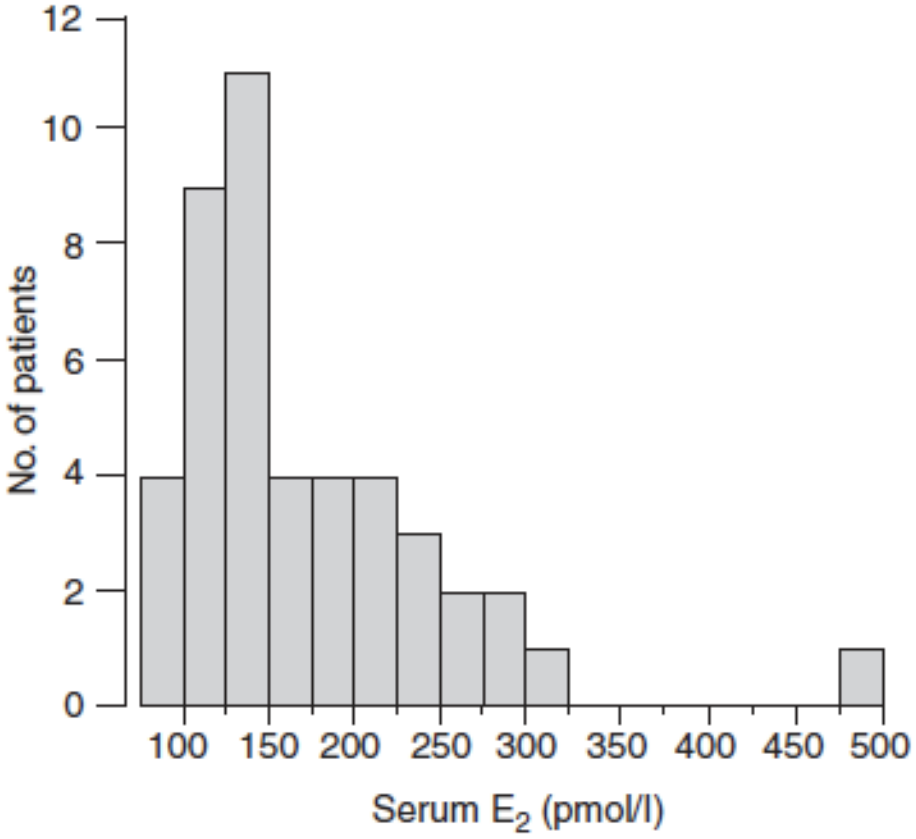


Príklad negatívne šikmej distribúcie

- distribúcia pacientov s akútnou pľúcnou embóliou

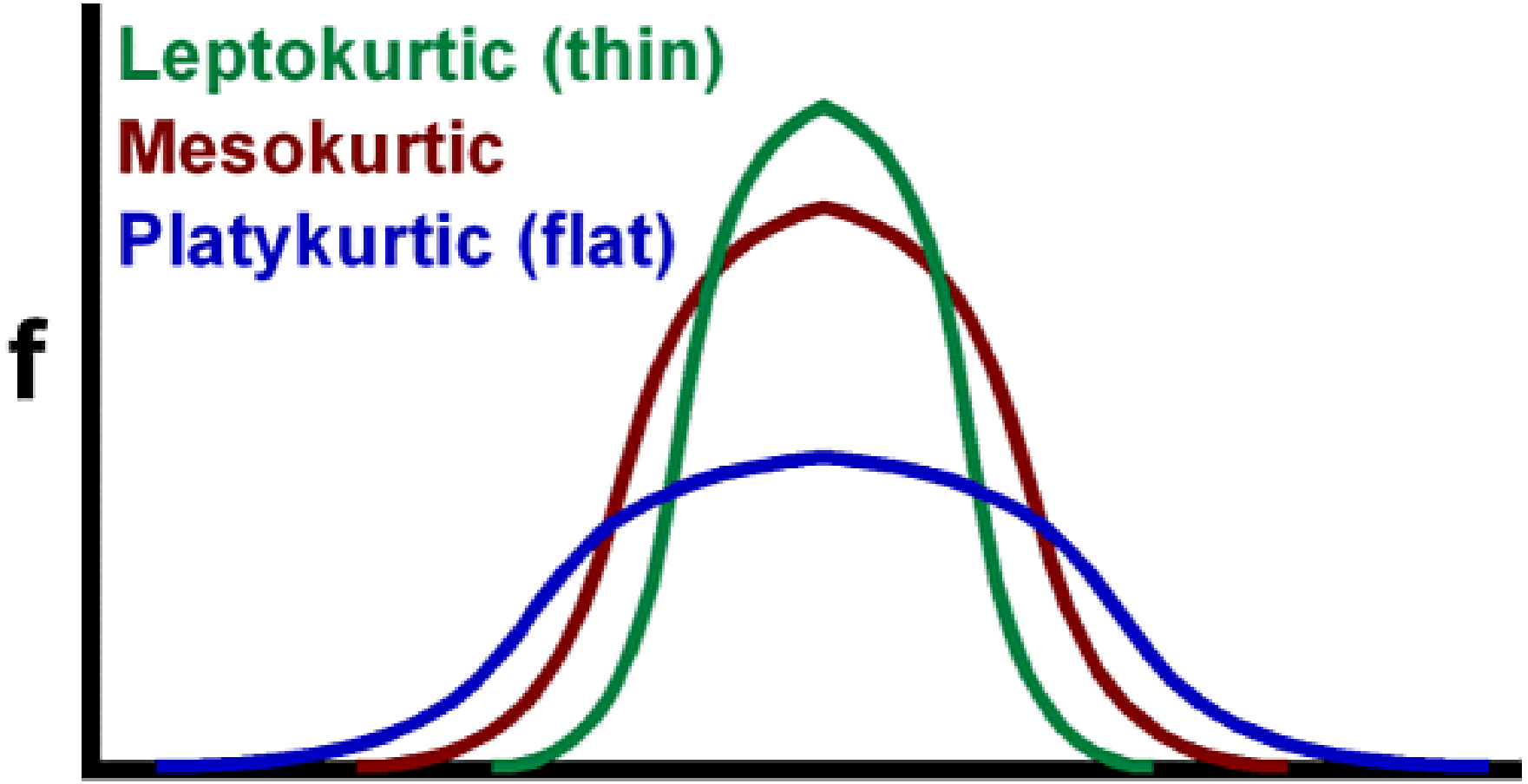


Príklad pozitívne šikmej distribúcie



- plazmatická hladina E_2 u pacientiek v prevencii osteoporózy

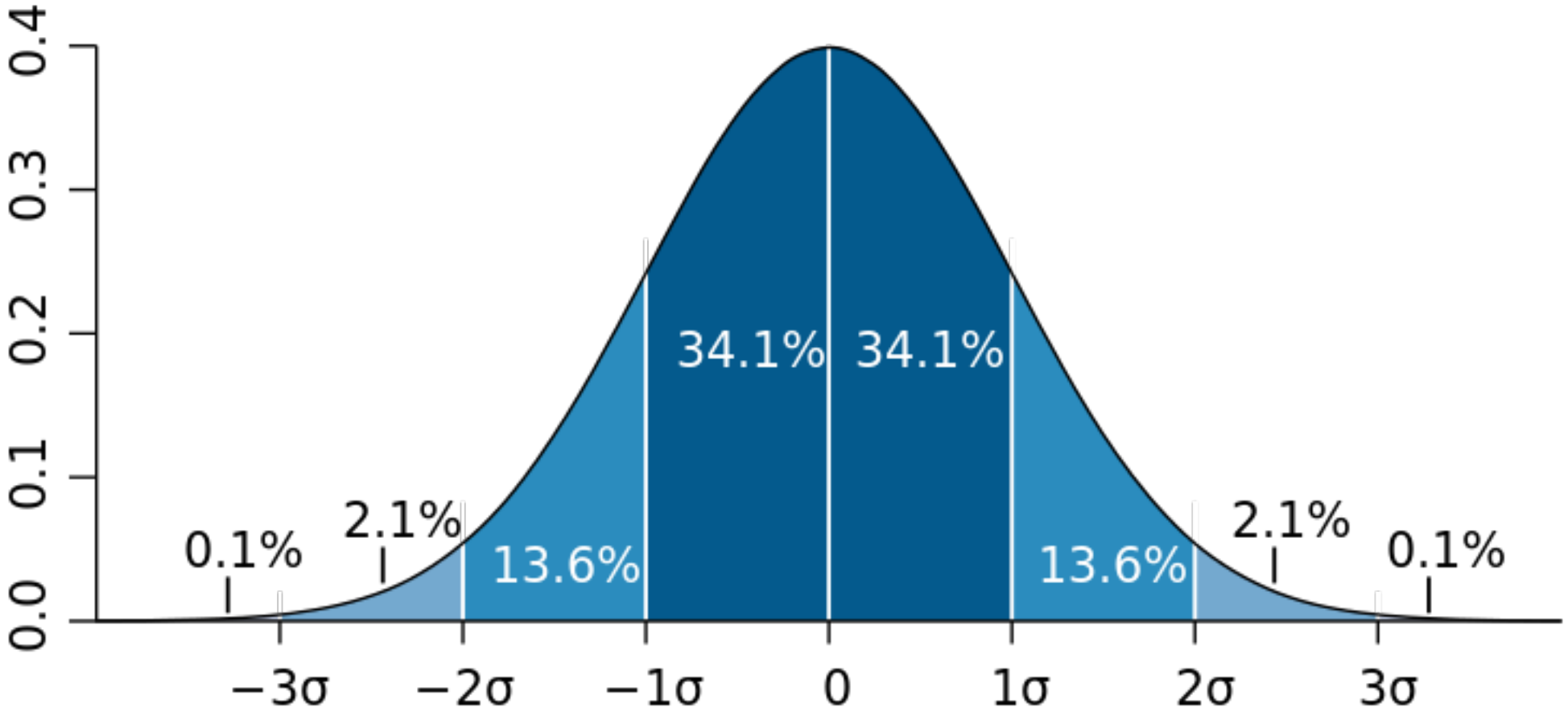
Distribúcia – špicatosť



3. Pravdepodobnosti
Normálne (Gaussove) rozdelenie

- koncept, ktorý je v štatistke veľmi dôležitý vzhľadom na niektoré špecifické charakteristiky:
  1) je symetrické
  2) priemer, medián aj modus nadobúdajú rovnakú hodnotu
     - najvyššia frekvencia prípadov je v strede (unimodálna) a klesá smerom k okrajom
  3) je asymptotické
     - pravý aj ľavý koniec sa nikdy nedotknú x-ovej (vodorovnej) osi

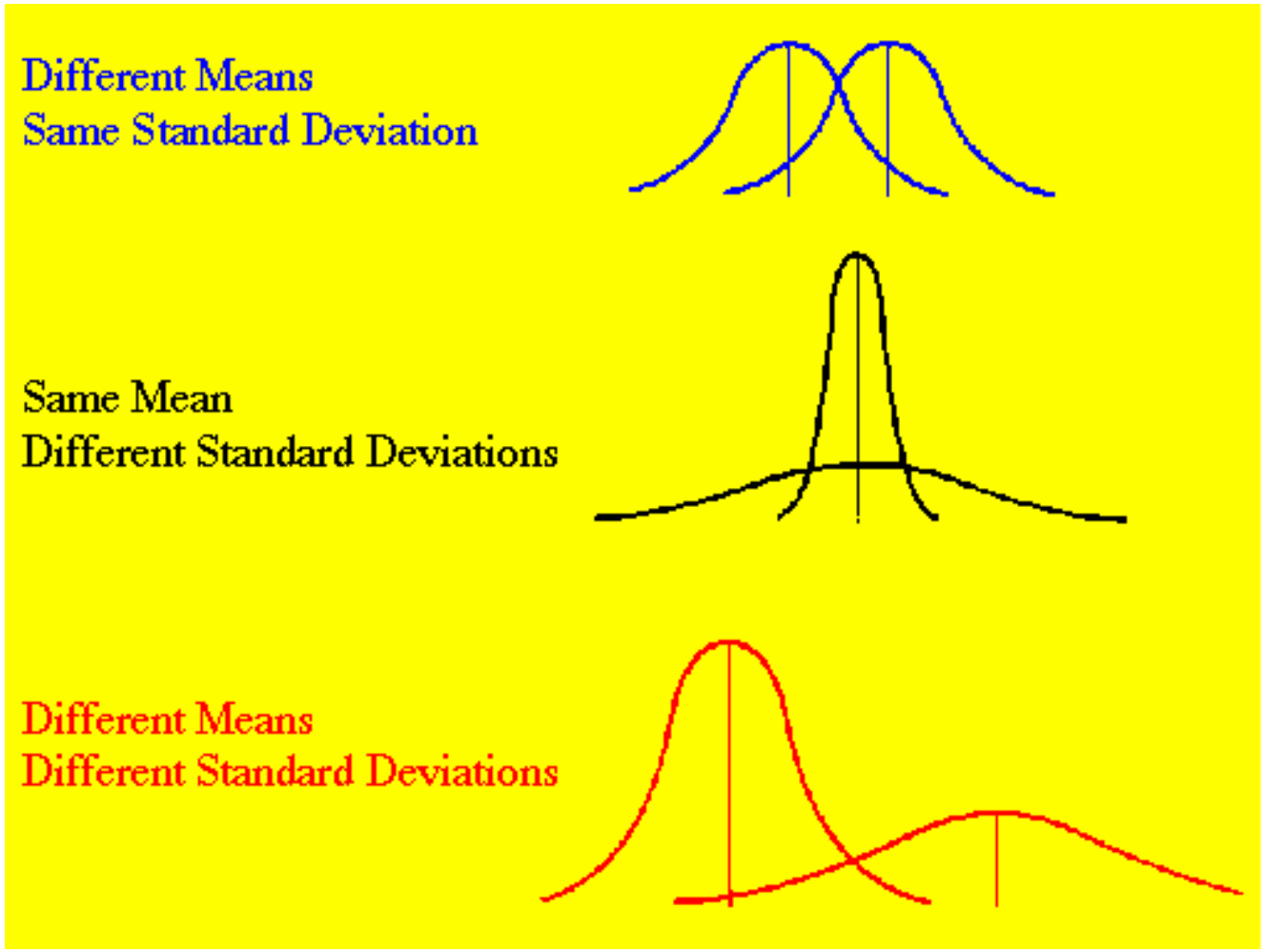
Normálne (Gaussove) rozdelenie



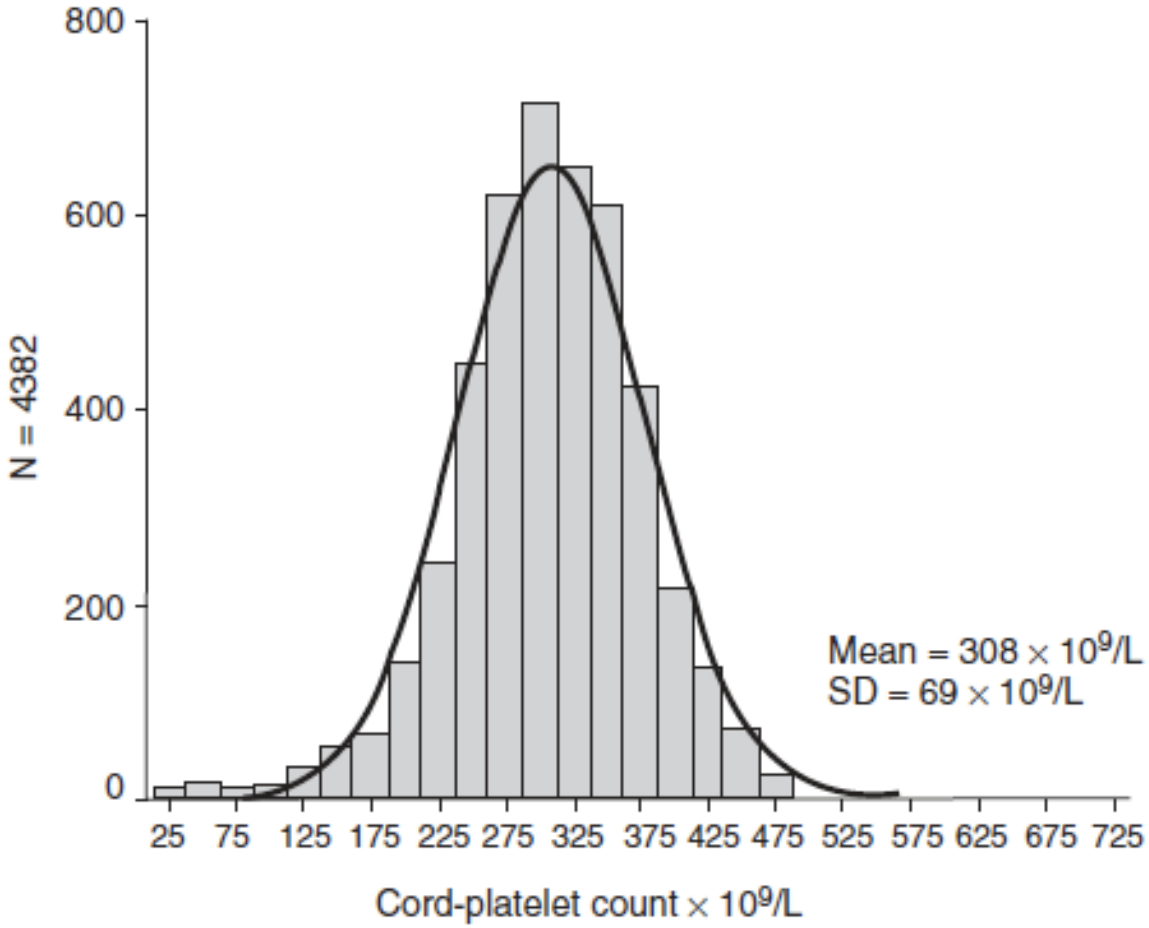
Normálne rozdelenie

- je definované 2 hodnotami:
  - priemerom a SD
- teoretická distribúcia
  - v skutočnom živote neexistuje, ale mnohé skutočné distribúcie sa mu dostatočne podobajú
- je dôležité v prediktívnej štatistike
  - ak má parameter (približne) normálnu distribúciu, máme dobré nástroje na odhadnutie javov v populácii

Normálne rozdelenie (Priemer a SD)



Normálne (Gaussove) rozdelenie v Praxi



Štandardizácia a z skóre

- priemer a štandardná odchýlka poskytujú užitočné informácie, keď chceme vedieť, ako vyzerá celá distribúcia výsledkov,
- ale občas potrebujeme vedieť aj informácie o individuálnom konkrétnom výsledku
  - na to sa používa **štandardizácia a z skóre**
  - ...ak máme normálne rozloženie hodnôt

Štandardizácia a z skóre

- štandardizácia – proces konvertovania každého výsledku v distribúcii na jednotky štandardnej odchýlky
- umožňuje nám porovnávať jablká s hruškami

Štandardizácia a z skóre

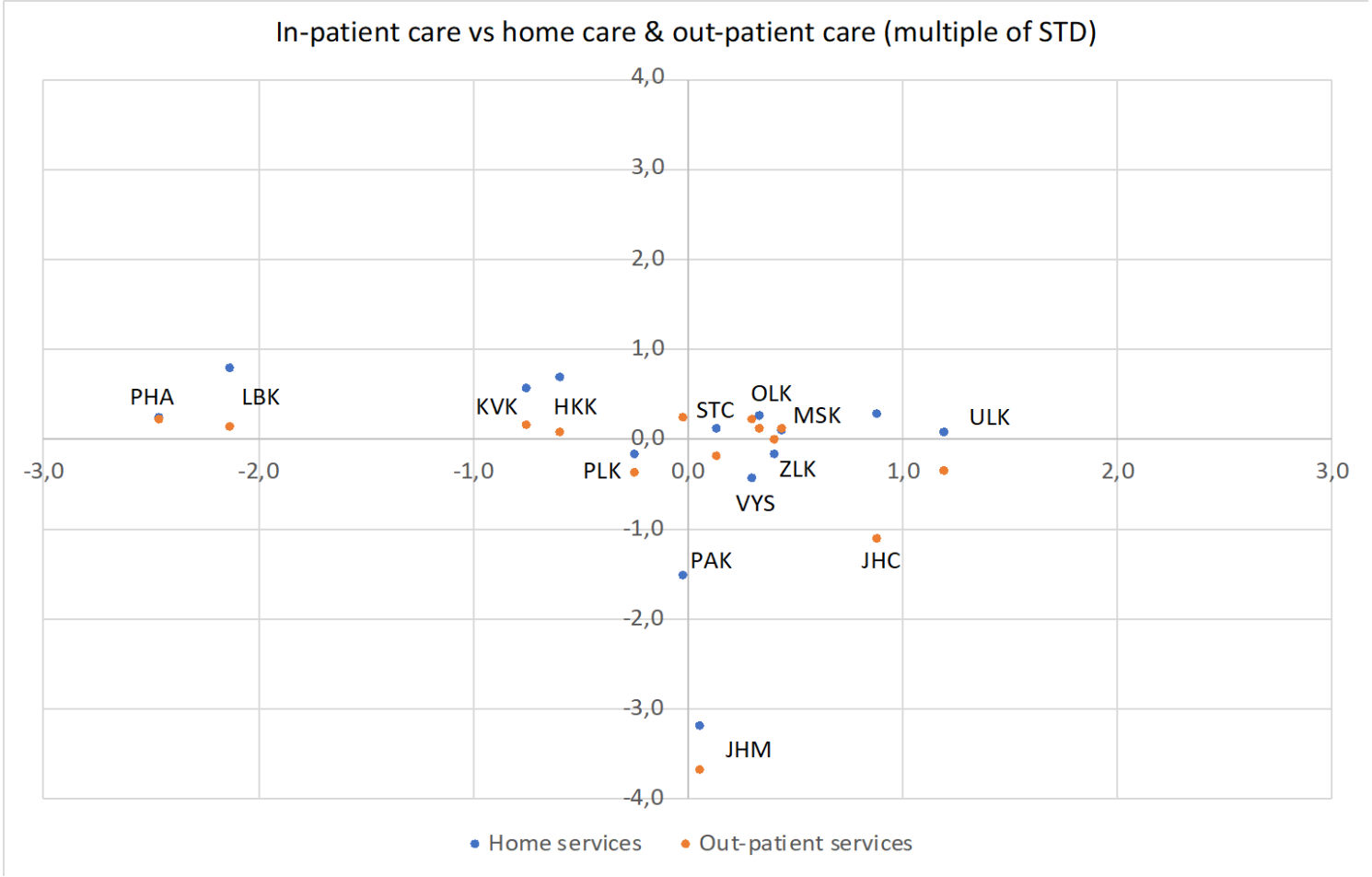
- z skóre – číslo, ktoré hovorí o tom, ako ďaleko od priemeru je dané skóre v distribúcii v jednotkách štandardnej odchýlky
  - hovorí, ako veľký alebo malý je výsledok v porovnaní s ostatnými výsledkami v distribúcii

$$z_i = \frac{X_i - x}{s}$$

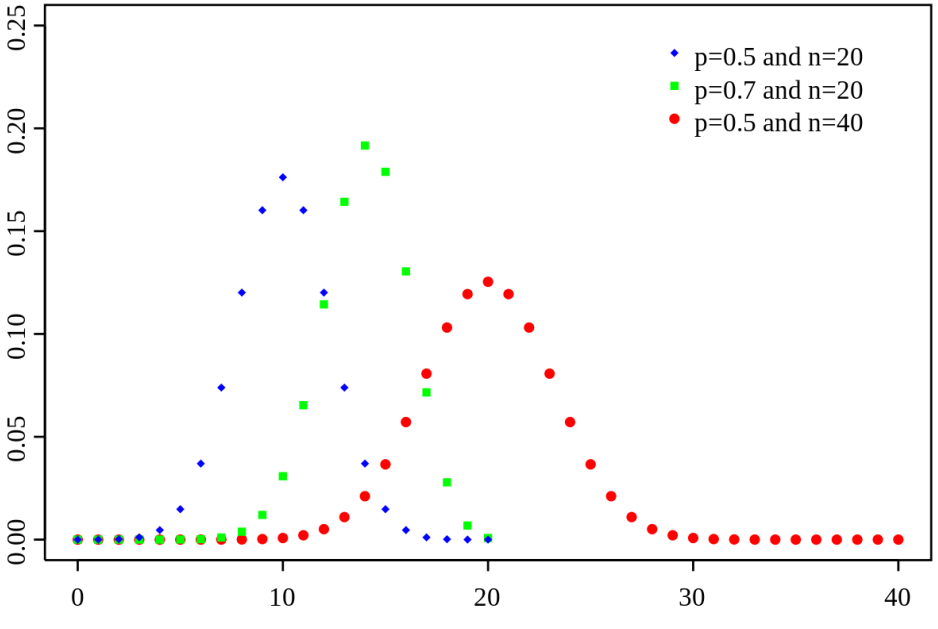
Štandardizácia a z- skóre v praxi

Region	EP	Current Capacity			Patients/current capacity			Difference to average			Difference to average/STD (Z-score)		
		I	H	O	I	H	O	I	H	O	I	H	O
PHA	17,681	4,409	969	351	4.0	18.2	50.4	-1.3	2.2	34.7	-2.4	0.2	0.2
VYS	7,155	2,812	285	140	2.5	25.1	51.1	0.2	-4.6	34.0	0.3	-0.4	0.2
MSK	16,432	6,650	838	245	2.5	19.6	67.1	0.2	0.8	18.0	0.4	0.1	0.1
OLK	8,922	3,533	497	130	2.5	17.9	68.8	0.2	2.5	16.3	0.3	0.2	0.1
PAK	7,167	2,641	199	152	2.7	36.1	47.2	0.0	-15.6	37.9	0.0	-1.5	0.2
HKK	8,068	2,671	587	107	3.0	13.7	75.3	-0.3	6.7	9.8	-0.6	0.6	0.1
LBK	5,953	1,552	470	93	3.8	12.7	63.7	-1.1	7.8	21.4	-2.1	0.7	0.1
ULK	10,822	5,229	546	73	2.1	19.8	148.0	0.6	0.7	-62.9	1.2	0.1	-0.4
PLK	8,128	2,868	364	55	2.8	22.3	148.9	-0.1	-1.9	-63.8	-0.2	-0.2	-0.4
STC	16,925	6,426	874	142	2.6	19.4	119.2	0.1	1.1	-34.2	0.1	0.1	-0.2
JHC	8,834	3,949	495	32	2.2	17.8	272.2	0.5	2.6	-187.1	0.9	0.2	-1.1
ZLK	8,260	3,317	371	93	2.5	22.3	89.1	0.2	-1.8	-4.0	0.4	-0.2	0.0
KVK	4,026	1,297	272	66	3.1	14.8	61.1	-0.4	5.6	24.0	-0.7	0.5	0.1
JHM	16,419	6,133	309	23	2.7	53.1	703.7	0.0	-32.6	-618.6	0.1	-3.1	-3.6
Total	144,791	53,487	7,076	1,702									
Average	10,342	3,821	505	122	2.7	20.5	85.1						
STD					0.5	10.5	173.1						

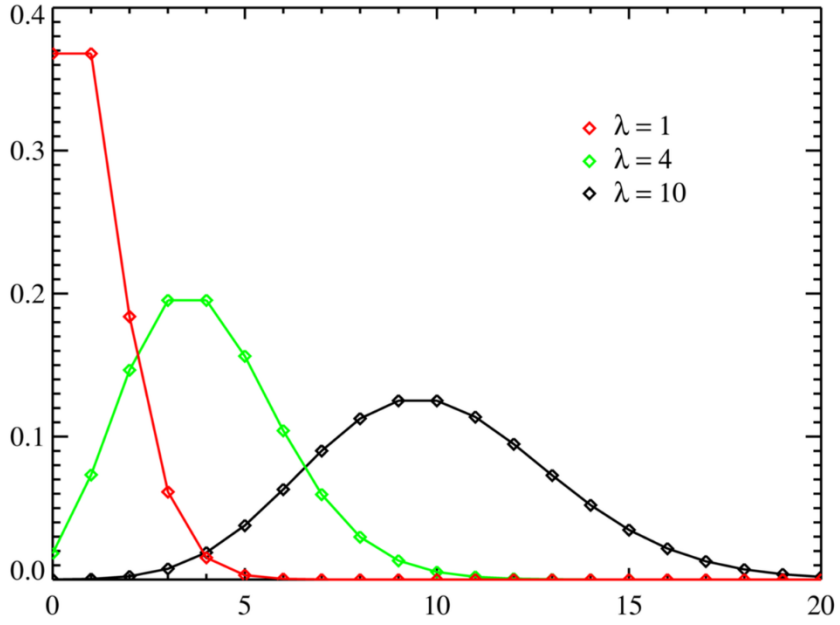
Štandardizácia a z- skóre v praxi



Diskrétne rozdelenia



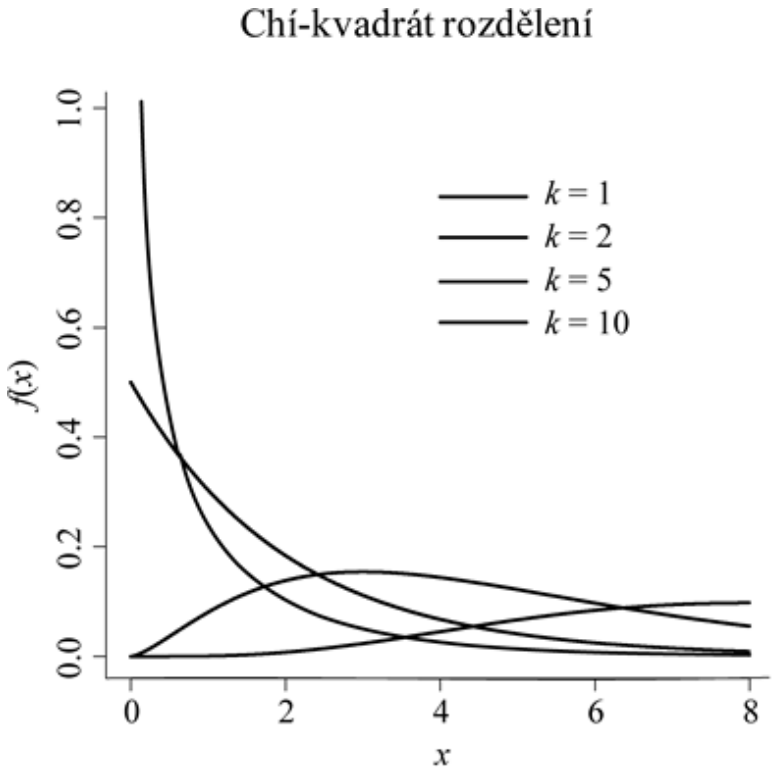
Binomické rozdelenie

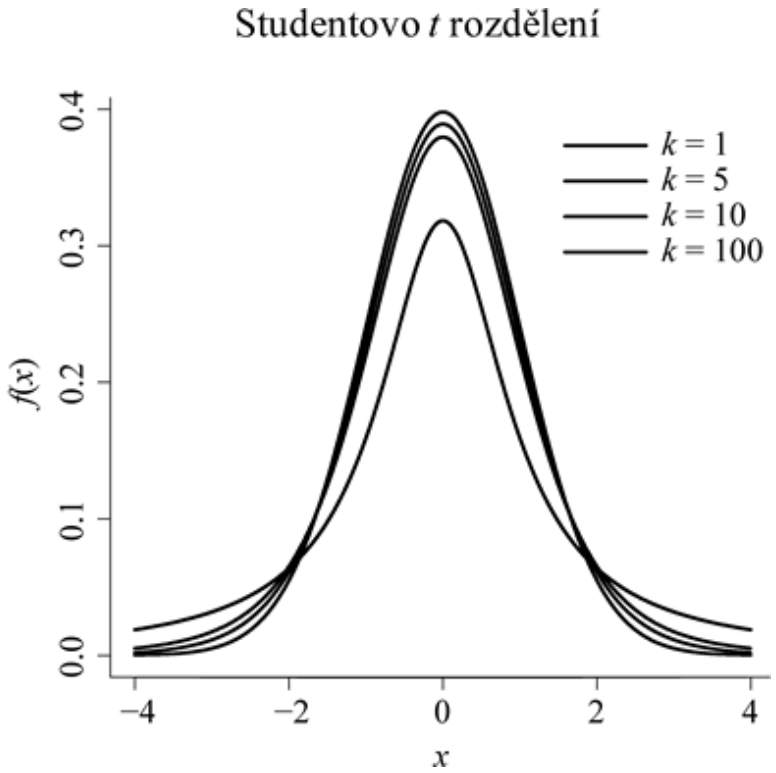


Poissonovo rozdelenie

- Hypergeometrické rozdelenie
- Rovnomerné rozdelenie

Spojité rozdelenia



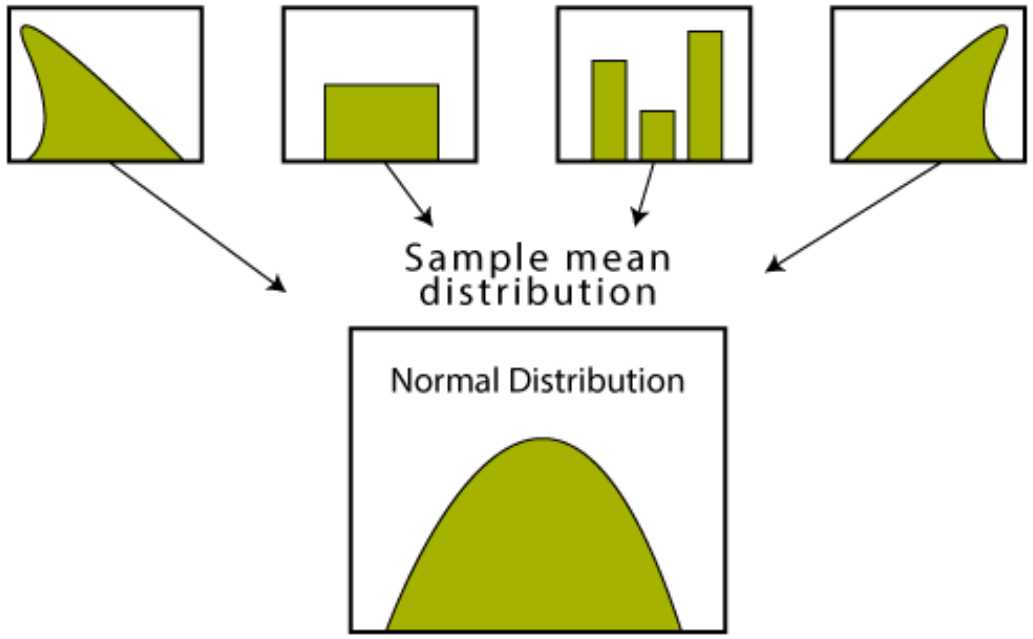


- Rovnomerné spojité rozdelenie
- Normálne rozdelenie
- Logaritmicko-normálne rozdelenie
- Exponenciálne rozdelenie
- Fisherovo-Snedecorovo rozdelenie (F-rozdelenie)

Centrálna limitná veta

- Centrálna limitná veta (Central Limit Theorem)
  - hovorí, že pokiaľ máme dostatočne veľký počet meraní, distribúcia priemerov rôznych výberov bude mať normálne rozdelenie (pri dostatočne veľkom n) a to aj v prípade, že distribúcia samotných hodnôt nemá normálne rozdelenie
- stredom tejto distribúcie je skutočný priemer populácie

Centrálna limitná veta



Hovorí, že pokiaľ máme dostatočne veľký počet meraní, distribúcia priemerov rôznych výberov bude mať normálne rozdelenie (pri dostatočne veľkom n) a to aj v prípade, že distribúcia samotných hodnôt nemá normálne rozdelenie

Štandardná chyba

- štandardná odchýlka (SD) distribúcie týchto priemerov (z viacerých výberov) sa nazýva štandardná chyba (SE) priemeru
  - SD hovorí o typickom rozdiele medzi hodnotou a priemerom
  - SE priemeru hovorí o typickom rozdiele medzi nameraným priemerom v jednom výbere a skutočným priemerom
  - vyjadruje, ako si môžeme byť istí, že priemer nášho výberu reprezentuje skutočný priemer v populácii

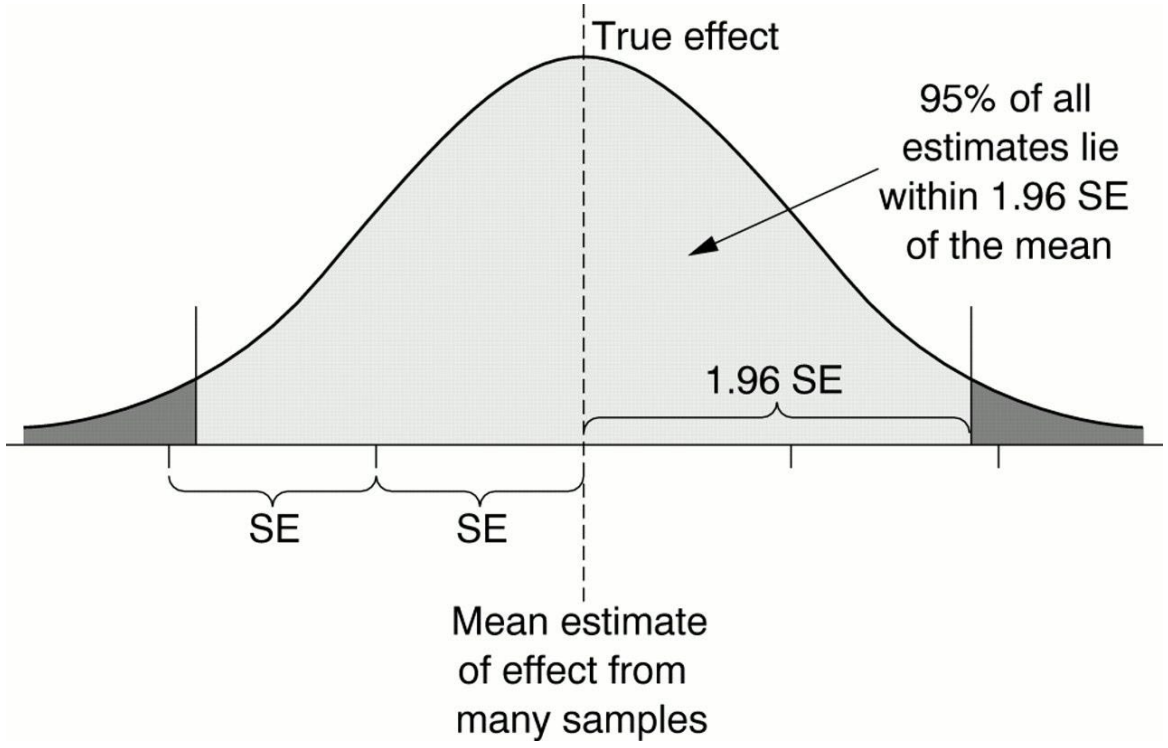
Štandardná chyba (SE)

- veľkosť SE závisí od:
  - aká veľká je vzorka (čím väčšia vzorka, tým menšia chyba a tým presnejší je odhad)
  - aká je veľká štandardná odchýlka
    (ak sú jednotlivé hodnoty vo vzorke príliš rozdielne, aj hodnoty v skutočnej populácii budú rozdielne - teda čím väčšia SD vo vzorke, tým väčšia bude variácia hodnôt v populácii a teda aj SE)

$$SE = \frac{SD}{\sqrt{n}}$$

Konfidenčný interval

- konfidenčný interval, interval spoľahlivosti
  - udáva oblasť hodnôt, v ktorej sa nachádza s pravdepodobnosťou 95% skutočný priemer
  - „sme si na 95% istý, že skutočný priemer je v tomto intervale“
  - 95% udáva rozmedzie: priemer ± 1,96 SE



Veľkosť vzorky a presnosť výsledkov

- zväčšovaním vzorky zvyšujeme presnosť výsledkov
  - odhad priemeru (a iných parametrov) na vzorke s počtom 1000 pozorovaní je presnejší než so vzorkou so 100 pozorovaniami
- ale so zväčšovaním vzorky zároveň stúpa pravdepodobnosť výskytu extrémnych hodnôt (outliers)
  - vo veľkých vzorkách sa však nízke a vysoké extrémy vzájomne vyrovnávajú a ich vplyv na priemer a SD je malý
  - v menších vzorkách môže výskyt jednej extrémnej hodnoty priemer a SD výrazne ovplyvniť

4. Testovanie hypotéz

- Štatistická hypotéza odráža konkrétny problém a vychádza z výskumnej hypotézy
- Preto vždy rozlišujeme:
  - Štatistickú významnosť (plynie z testu štatistickej hypotézy)
  - Skutočnú významnosť (plynúcu z povahy riešeného problému)
- Hypotézy
  - Nulovú hypotézu označujeme jako H_0
  - Alternatívnu hypotézu ako H_1
  -

H_0: $\pi = 0{,}9$

H_1: $\pi < 0{,}9$ jednostranná

H_2: $\pi = 0{,}9$ obojstranná

Testovanie hypotéz

- Testové kritérium – jeho hodnotu počítáme z náhodného výberu
  - Obor prijatia (označujeme aj jako V)
  - Kritický Obor (označujeme aj ako W)
- Obor prijatia V – obsahuje také hodnoty testovej štatistiky, ktoré nie sú v rozpore s platnosťou nulovej hypotézy a pre ktoré nulovú hypotézu nezamietame
- Kritický obor W – obsahuje hodnoty testovej štatistiky, ktoré sú v rozpore s nulovou hypotézou
  (hovorí v prospech alternatívnej hypotézy) a vedie k zamietnutiu nulovej hypotézy

Testovanie hypotéz

- Prijatie/zamietnutie nulovej hypotézy
  - Pokiaľ došlo k zamietnutiu nulovej hypotézy, nazveme **test štatisticky významným**
  - V opačnom prípade hovoríme, že **test je nevýznamný**

- Vzhľadom k tomu, že sa rozhodujeme na základe náhodného výberu, a teda iba s obmedzenými informáciami, môže i správne zvolený a použitý test hypotézy viesť k mylnému rozhodnutiu

Chyby pri testovanie hypotéz

náš úsudok / skutočnosť	nezamietneme H_0	zamietneme H_0
H_0 je pravdivá	máme pravdu Skutočne negatívny výsledok	chyba I. typu (α) Falošne pozitívny výsledok
H_1 je pravdivá	chyba II. typu (β) Falošne negatívny výsledok	máme pravdu Skutočne pozitívny výsledok

Testovanie hypotéz

- Bolo by dobre, aby obe pravdepodobnosti chybných rozhodnutí boli malé
- Bohužiaľ však platí, že keď znižujeme pravdepodobnosť jedného zlého rozhodnutia, zvyšuje sa pravdepodobnosť druhého špatného rozhodnutia

- V zdravotníctve hovoríme o senzitivite a špecificite diagnostického testu

Senzitivita a špecificita (DG test)

DG test / Prítomnosť ochorenia	nezamietneme H_0 (test je negatívny)	zamietneme H_0 (test je pozitívny)
H_0 je pravdivá (nie - zdravý)	máme pravdu Skutočne negatívny výsledok	chyba I. typu (α) Falošne pozitívny výsledok
H_1 je pravdivá (áno - chorý)	chyba II. typu (β) Falošne negatívny výsledok	máme pravdu Skutočne pozitívny výsledok

Senzitivita a špecificita (DG test)

DG Test / Prítomnosť ochorenia	nezamietneme H_0 (test je negatívny)	zamietneme H_0 (test je pozitívny)	Spolu
H_0 je pravdivá (nie - zdravý)	65 Skutočne negatívny výsledok	10 Falošne pozitívny výsledok	75 Zdraví spolu
H_1 je pravdivá (áno - chorý)	5 Falošne negatívny výsledok	20 Skutočne pozitívny výsledok	25 Chorí spolu
Spolu	70 Negatívny test spolu	30 Pozitívny test spolu	100 Spolu

Senzitivita

- *Senzitivita* je definovaná ako pravdepodobnosť, že test bude pozitívny u chorého subjektu.
- Keď počítame senzitivitu, zaujímame sa iba o „chorú“ časť populácie.
- Test správne diagnostikoval 20 z 25 subjektov, ktorí majú dané ochorenie. Takže senzitivita tohoto testu je 80%

$$\text{senzitivita} = \frac{\text{počet skutečně pozitivních}}{\text{počet skutečně pozitivních+počet falešně negativních}} = \frac{20}{20+5} = 80\%$$

Špecifita

- *Špecificita* je definovaná ako pravdepodobnosť, že test bude negatívny u subjektu, ktorý nemá dané ochorenie.
- Keď sa zaujímame o špecificitu, uvažujeme iba „zdravú“ časť populácie.
- Test správne identifikoval 65 zo 75 subjektov, ktorí netrpeli daným ochorením. Takže špecificita tohoto testu je 87%

$$\text{specificita} = \frac{\text{počet skutečně negativních}}{\text{počet skutečně negativních+počet falešně pozitivních}} = \frac{65}{65 + 10} = 87\%$$

Senzitivita a špecifita u COVID-19 testov

Recommended diagnostic method for COVID-19

Real-time PCR method is recommended by WHO for COVID-19 diagnosis.

COVID-19 TEST	Antigen-based immunoassy	Antibody-based immunoassay	Real-time PCR
Analyte	Antigen	Antibody	Gene
Detectable period	From a few days after onset of symptoms	From 7-28 days after onset of symptoms	All of stages
Sensitivity	50-70% (expected)*	More than 95%	More than 95%
Specificity	50-70%*	Not clear*	More than 95%
Detection of asymptomatic infection	Depending on the amount of viral antigen	At the later stage of infection	From the early stage of infection
Status of use	Only some regions	Only some regions	All of countries (recommended by WHO & CDC)

*Based on conventional antigen-based immunoassay and affected by seasonal coronaviruses
Reference: Oral presentation from online forum of KOFST(The Korean Federation of Science and Technology Societies)

Pozitívna prediktívna hodnota

- Uvažujme situáciu, že výsledok diagnostického testu je u pacienta pozitívny.
- Potom si môžeme položiť otázku: aká je pravdepodobnosť, že je tento pacient naozaj (v skutočnosti) chorý?
- Pri výpočte PPV sa pozrieme len na tú časť našej populácie, ktorá mala pozitívny výsledok testu.
- Vidíme, že 20 z 30 pozitívnych výsledkov je správnych. Takže pozitívna prediktívna hodnota tohto testu je 66%.

Negatívna prediktívna hodnota

- Pri výpočte NPV uvažujeme len tú časť z našej populácie, ktorá mala negatívny výsledok testu.
- V našom prípade je 65 zo 70 negatívnych výsledkov správnych.
- Z toho vyplýva, že negatívna prediktívna hodnota je rovná 93%.

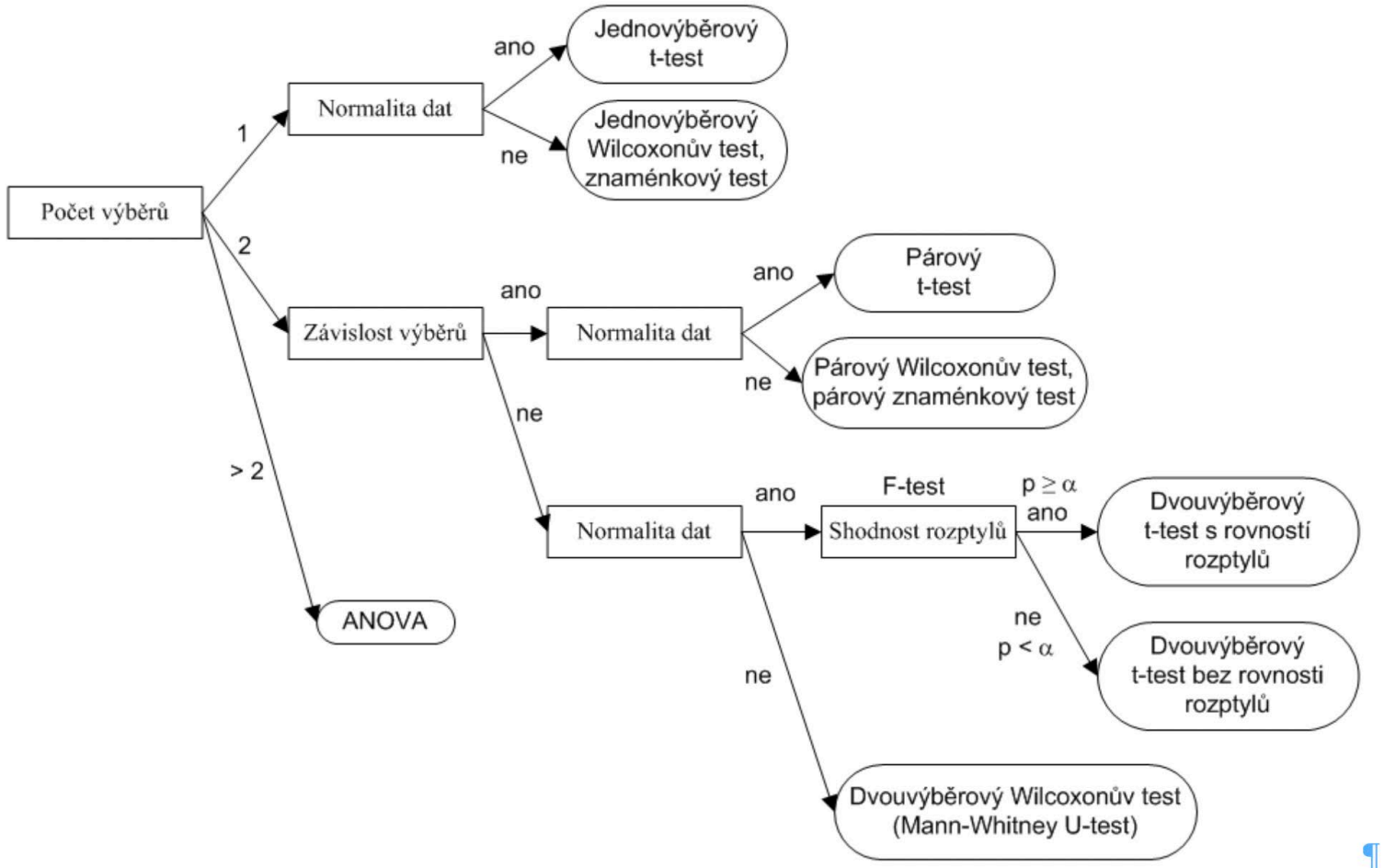
Zhrňme si, že ...

- **Senzitivita**: ak pacient má dané ochorenie, ako často je test pozitívny.
- **Špecificita**: ak je pacient nemá dané ochorenie, ako často bude výsledok testu negatívny.
- **Pozitívna prediktívna hodnota**: ak je test pozitívny, aká je pravdepodobnosť, že pacient má dané ochorenie.
- **Negatívna prediktívna hodnota**: ak je test negatívny, aká je pravdepodobnosť, že pacient nemá dané ochorenie.

Vyhodnotenie hypotéz

- Prostredníctvom software (ja osobne používam SAS University Edition)
- Pre rozhodnutie je používaná p-hodnota
- P-hodnotu určíme ako pravdepodobnosť, že za platnosti nulovej hypotézy obdržíme hodnotu testového kritéria získanú náhodným výberom
- Je to tiež minimálna hladina významnosti, pre ktorú ešte zamietame nulovú hypotézu
- Pri rozhodnutí o hypotéze porovnávame p-hodnotu so zvolenou hladinou významnosti alfa:
  - p-hodnota <= alfa, zamietame nulovú hypotézu v prospech alternatívy
  - p-hodnota > alfa, nulovú hodnotu nezamietame

Ako vybrať správny test?



¶

Chí-kvadrát test

- chí-kvadrát test dobrej zhody
- na porovnávanie výskytu kategorických premenných (pozorované frekvencie) vo vzorkách oproti očakávanému výskytu vo frekvenciách medzi skupinami (očakávané frekvencie)
- zisťujeme, či sú pozorované frekvencie štatisticky významne odlišné od očakávaných frekvencií
- Napríklad porovnanie podielu cisárskych rezov v českých pôrodniciach (25%) s odporúčaním WHO (15%)
- umožňuje porovnávanie aj viacerých skupín alebo typov výsledkov

5. Analýza závislosti

- Analýza kontingenčných tabuliek
- Analýza rozptylu
- Regresná analýza
- Korelačná analýza

Analýza kontingenčných tabuliek

- Dvojrozmerné triedenie dát (vytvorenie kontingenčnej tabuľky)
- Test nezávislosti v kontingenčnej tabuľke
  - Chí-kvadrát test nezávislosti v kontingenčnej tabuľke
  - Vyhodnocujem P-hodnotu voči hladine významnosti alfa
  - Keď p-hodnota menšia ako 0,05, existuje medzi premennými významná asociácia

namerané	základné	stredné	vysokoškolské	spolu
podváha	68	16	8	92
normálna	39	67	71	177
nadváha	5	14	19	38
obezita	0	1	6	7
spolu	112	98	104	314
očakávané	základné	stredné	vysokoškolské	spolu
podváha	33	29	30	92
normálna	63	55	59	177
nadváha	14	12	13	38
obezita	2	2	2	7
spolu	112	98	104	314
chi-kvadrat	**0,00**			
existuje štatisticky významná asociácia medzi vzdelaním a BMI				

Analýza rozptylu (ANOVA)

Level of kraj	N	AD	
		Mean	Std Dev
Jihocesky	5	0.88000000	0.13038405
Jihomoravsky	5	0.73000000	0.14832397
Karlovarsky	5	0.77000000	0.22803509
Kralovehradecky	7	0.90000000	0.09574271
Liberecky	5	0.63000000	0.22803509
Moravskoslezsky	5	0.60000000	0.27386128
Olomoucky	5	0.89000000	0.11401754
Pardubicky	5	0.58000000	0.16431677
Plzensky	5	0.81600000	0.24296090
Praha	5	0.97000000	0.04472136
Stredocesky	5	0.85000000	0.15811388
Ustecky	6	0.61666667	0.35449495
Vysocina	13	0.80153846	0.25218532
Zlinsky	5	0.77000000	0.10954451

Dependent Variable: AD AD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	1.07553299	0.08273331	1.93	0.0418
Error	67	2.87062256	0.04284511		
Corrected Total	80	3.94615556			

R-Square	Coeff Var	Root MSE	AD Mean
0.272552	26.67659	0.206991	0.775926

Source	DF	Type I SS	Mean Square	F Value	Pr > F
kraj	13	1.07553299	0.08273331	1.93	0.0418

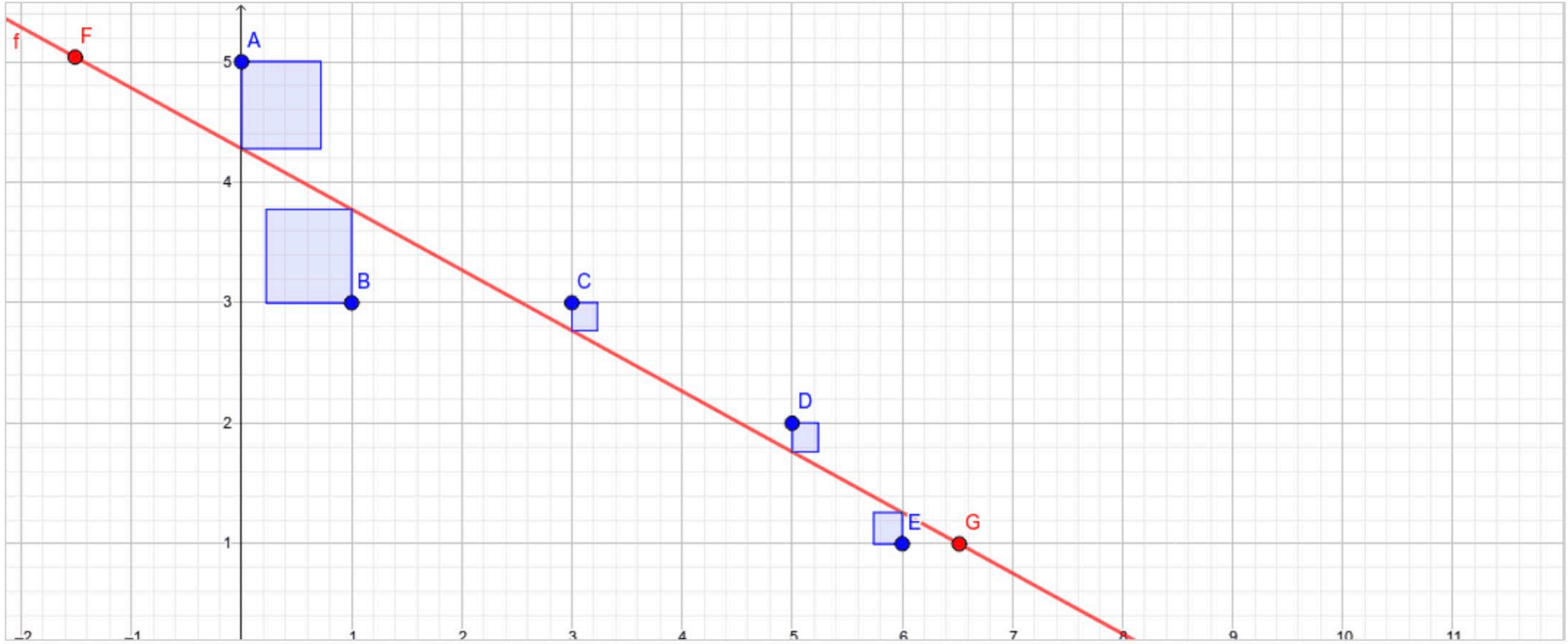
Source	DF	Type III SS	Mean Square	F Value	Pr > F
kraj	13	1.07553299	0.08273331	1.93	0.0418

Regresná analýza

- Predpokladajme dvojicu znakov
  - Veľkosť zariadenia
  - Prevalenciu AD
- Hodnoty zobrazíme v grafe
- Pokúšame sa vysvetliť závislosť medzi znakmi pomocou regresnej funkcie
- Pri konštrukcii grafu sme sa museli rozhodnúť, ktorú premennú dáme na ktorú os:
  - Premenná x – vodorovná os – vysvetľujúca premenná (nezávislá premenná)
  - Premenná y – zvislá os – vysvetľovaná premenná (závislá premenná)

Regresná analýza

- Priamková regresia – funguje na báze metódy najmenších štvorcov



Párový korelačný koeficient

- Nazývaný aj Pearsonov korelačný koeficient
- Má hodnoty < -1; +1>
- Hodnoty -1 má v prípade, že všetky body ležia na jednej klesajúcej priamke, ktorá má zápornú smernicu
- Hodnoty 1 má v prípade, že všetky body ležia na rastúcej priamke a má kladnú smernicu
- Pokiaľ je korelačný koeficient = 0, považujeme premenné za lineárne nezávislé (v takomto prípade je regresná priamka rovnobežná s vodorovnou osou)
- Pozor, korelačný koeficient sa vzťahuje iba k lineárnej závislosti, približne nulová hodnota koeficientu môže znamenať, že veličiny sú silne závislé, ale závislosť nie je lineárna (môže byť napr. Parabola)

Ďalšie typy regresných funkcií

- Vystihujú vzťahy, pre ktorých popis sa nehodí priamka (čiže lineárna regresia)
- Polynomická regresia
- Hyperbolická regresná funkcia
- Logaritmická regresná funkcia
- Exponenciálna regresná funkcia
- Mocninová regresná funkcia

Regresný model

- V regresnej analýze sa snažíme závislosti popísať regresnými funkciami
- Môžeme teda testovať štatistické hypotézy týkajúce sa regresného modelu
- Cieľom regresnej analýzy je hlbšie vniknutie do všeobecných rysov skúmaných závislostí a nájdenie takého matematického zápisu, ktorý by umožňoval všeobecné úvahy o podstate sledovaných javov

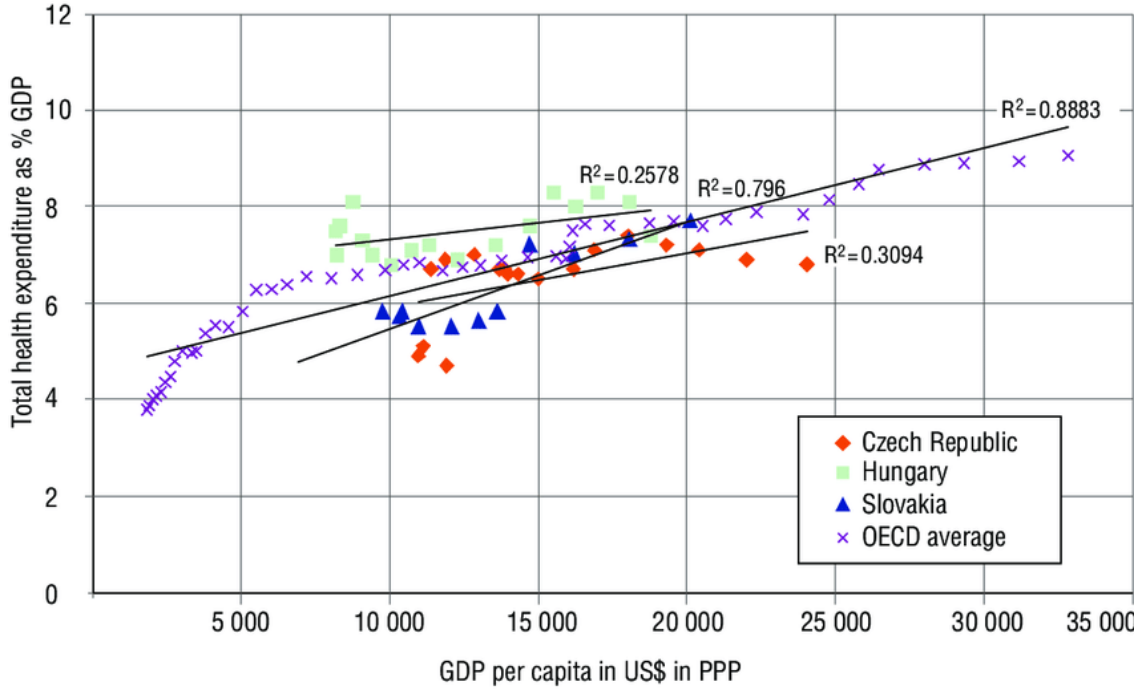
Regresný model – otázky ku konštrukcii

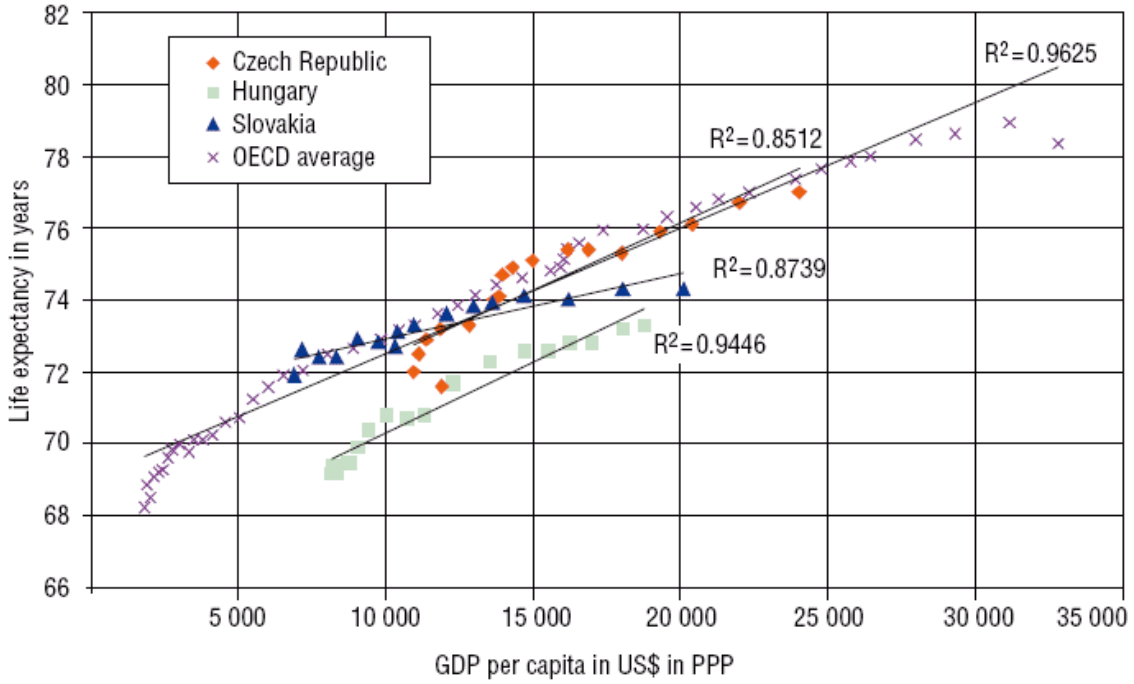
- Máme vhodný model regresnej funkcie?
- Máme správny výber vysvetľujúcich premenných?
- Aký je význam jednotlivých vysvetľujúcich premenných zaradených do regresnej funkcie?

Kvalita regresného modelu

- Kvalitu regresného modelu posudzujeme koeficientom determinácie (R^2)
  - Interpretujeme ako percento variability vysvetľovanej premennej vysvetlenej regresným modelom
  - $R^2 = 0$ (nezávislé premenné)
  - $R^2 = 1$ (regresný model vysvetlil 100% variability vysvetľovanej premennej)

Regresný model a R^2





Source: OECD, 2009.

Regresný model - predikcie

- Regresný model môžeme použiť aj na odhady stredných hodnôt vysvetľovanej premennej pre zvolenú hodnotu x, alebo individuálnej hodnoty y v tomto bode
- Takýto postup nazývame predikcia (alebo predpoveď)
- **Interpolácia** – hodnota nezávislej premennej je obsiahnutá v intervale medzi najmenšou a najväčšou hodnotou pozorovanej nezávislej premennej
- **Extrapolácia** – hodnota nezávislej premennej je mimo interval medzi najmenšou a najväčšou hodnotou pozorovanej nezávislej premennej
- **Predikčný pás** – interval spoľahlivosti pre individuálne hodnoty sledovanej premennej. Pás okolo regresnej priamky by mal v priemere obsahovať 95% z n pozorovaní

Viacnásobná regresia

- Skúma závislosť medzi jednou vysvetľovanou premennou a viacerými vysvetľujúcimi premennými (voláme ich regresory)
- Opäť hľadáme matematický zápis, ktorý uvedenú závislosť vysvetľuje

Multikolinearita

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	-7.17762	2.46418	-2.91	0.0036	.	0
dom	dom	1	0.00911	0.00104	8.77	<.0001	0.41724	2.39673
kraj	kraj	1	-0.00030126	0.00009176	-3.28	0.0010	0.96624	1.03493
rok	rok	1	0.00351	0.00123	2.86	0.0042	0.99833	1.00167
vel	vel	1	0.00236	0.00081155	2.91	0.0037	0.94449	1.05877
deti	deti	1	0.01266	0.00289	4.39	<.0001	0.65494	1.52687
duch	duch	1	0.05185	0.00724	7.16	<.0001	0.21053	4.74997
ea	ea	1	-0.04894	0.00545	-8.98	<.0001	0.23483	4.25838
ost	ost	1	0.01263	0.00744	1.70	0.0896	0.53069	1.88433
vek_p	vek_p	1	0.00443	0.00021292	20.79	<.0001	0.46394	2.15543

- Odstraňujeme premenné s VIF (Variance Inflation Factor) vyšším ako 5 až dovtedy, pokiaľv modeli nezostanú premenné s VIF menším ako 5
- Toto je dôležité preto, aby boli eliminované tie premenné, ktoré sú vzájomne lineárne závislé s inou premennou

- Napr. v tomto modeli boli eliminované tieto premenné
- Osob (VIF = 18,82)
- Vek_m (VIF = 10,86)

Volba regresnej funkcie

- Software
  - Excell umožňuje výber pri scatter plote (x,y- graf) a následne počíta R^2
  - SAS University Edition (ja osobne používam) umožňuje predovšetkým pri viacnásobných regresiách hlbšie možnosti

Volba regresnej funkcie

- Pri hľadaní regresnej funkcie (najprv diagnostické testy)
- Následne môžem použiť krokové metódy (vykoná samotný software)
  - Forward selection – začíname s modelom, ktorý obsahuje konštantu a postupne pridávame premenné, až pokiaľ nezaradené premenné už neprinášajú žiadnu novú hodnotu pre vysvetlenie premennej y
  - Backward elimination – Model začína so všetkými uvažovanými premennými a postupne vynecháva menej dôležité premenné, nakoniec v modeli ponechá len tie, ktoré významne prispievajú do regresného modelu
  - Stepwise selection (kombinácia oboch prístupov) – Začína s prázdnym modelom (konštanta), v každom ďalšom kroku vždy vyhľadá premennú, ktorú by bolo vhodné pridať a súčasne skúma, či už nejaké premenná zaradená do modelu nie je nadbytočná
- Pozor! Rôzne postupy môžu priniesť rôzne výsledné regresné modely! Následne pozeráme R^2
- Pozor! Krokové metódy neodstraňujú multikolinearitu

Zhrnutie

1. Úvod do štatistiky
   - štatistika **hovorí o priemernom alebo typickom prípade**
   - štatistika je užitočná, ak chceme **porovnať dve alebo viaceré skupiny.** Odpovedá na otázky, či sú dve charakteristiky na sebe **závislé** a snažia sa robiť **predpovede**
2. Deskriptívna štatistika = Sumarizácia a popis dát
3. Pravdepodobnosti – prediktívna štatistika
4. Testovanie hypotéz
   - Štatistická hypotéza odráža konkrétny problém a vychádza z výskumnej hypotézy
   - Nulová a alternatívna hypotéza
5. Analýza závislostí
   - Analýza kontingenčných tabuliek, Analýza rozptylu, Regresná analýza, Korelačná analýza

Zdroje

- OECD, 2009
- DRUMMOND, M F. *Methods for the economic evaluation of health care programmes.* Oxford: Oxford University Press, 2015. ISBN 978-0-19-966587-7.
- Hindls a kol.: Statistika v ekonomii, 2018, Professional publishing
- Szalay et al: Slovakia, Healthcare in Transition, 2011
- PAŽITNÝ, Peter, Daniela KANDILAKI a Lenka KOMÁRKOVÁ, 2019. Current Capacity Gap in Dementia/AD in the Czech Republic. In SOUKUPOVÁ, N. -- MATĚJČKOVÁ, M. (ed.). Proceedings of the 13th International Scientific Conference INPROFORUM 100 Years of the Koruna. České Budějovice: University of South Bohemia in České Budějovice, Faculty of Economics, 2019, s. 251--258. ISBN 978-80-7394-776-7. URL: http://ocs.ef.jcu.cz/files/site/INPROFORUM2020.pdf
- HANZALOVÁ, Markéta; PAŽITNÝ, Peter; KANDILAKI, Daniela. Prevalence of Alzhei-mer's Disease in Retirement Homes and Homes with a Special Regime in the Czech Republic. 2020.

Ďakujem veľmi pekne za pozornosť