Základy pravděpodobnosti

Nulová hypotéza, Normální rozložení, Další typy rozložení,Standardizace.

USUZOVÁNÍ - O CO JDE?

Pozorovaný výsledek představuje

- STATISTIKU (jak je to v našem výběrovém souboru),

z níž usuzujeme na

- PARAMETR (jak je to v populaci, z níž byl soubor vybrán).

PARAMETRY ZÁKLADNÍHO SOUBORU obvykle neznáme, ale můžeme je odhadovat z VÝBĚROVÝCH STATISTIK.

STATISTICKÁ INFERENCE - (USUZOVÁNÍ)

- Jde-li skutečně o VÝBĚROVÝ SOUBOR (při vyčerpávajícím šetření to nemá smysl).
- Jde-li o NÁHODNÝ VÝBĚR kdy každá jednotka dané populace má stejnou pravděpodobnost, že bude vybrána.
- Jde-li o NEZÁVISLÝ VÝBĚR (výběr žádné jednotky nezvyšuje ani nesnižuje pravděpodobnost výběru jiných jednotek).

Smysl otázky "JAK JSOU POZOROVANÉ VÝSLEDKY PRAVDĚPODOBNÉ?":

- TEORETICKÁ HYPOTÉZA vs. STATISTICKÁ HYPOTÉZA (neboli formální výroky o: neznámých parametrech základního souboru, o tvaru rozložení četností, o statistických vztazích mezi soubory či proměnnými v něm….)

- Lze, se zvolenou pravděpodobností předpokládat, že STATISTIKA jakožto pozorovaný výsledek reprezentuje nepozorovatelný PARAMETR?
- Není STATISTIKA v důsledku výběrové chyby přece jen příliš vzdálená PARAMETRU?
- V jakém intervalu kolem STATISTIKY můžeme s danou pravděpodobností očekávat výskyt PARAMETRU?

OBVYKLE SE TESTUJE:

- Zkoumaný výběrový soubor pochází ze základního souboru s určitým rozdělením (zda je výběr reprezentativní).
- Jak se odchyluje věkový průměr ve výběrovém souboru od známého věkového průměru populace.
- Jak se odchyluje struktura volebních preferencí ve výběrovém souboru od známé struktury těchto preferencí v populaci.
- Dva výběry pocházejí ze (stejného) základního souboru s určitým rozdělením.
- Liší se průměrné mzdy žen a mužů tak, že to nemůže být vysvětleno náhodou?
- Zda je možno považovat studovaný soubor za náhodně uspořádaný (zda mezi proměnnými neexistuje žádný vztah).
- Například distribuci proměnné lze považovat za náhodně uspořádánu, jestliže jsou všechny její kategorie stejně početné.
- Jak se hodnota odchyluje od určitého standardu
- Jak se odchyluje průměrná pracovní doba od zákonem stanovené délky pracovní doby.
- Jak se odchyluje vzdělanostní struktura čtenářů časopisu RESPEKT od vzdělanostní struktury populace.
- Jak se odchyluje průměrné IQ ve skupině delikventů od 100 bodů.

NULOVÁ HYPOTÉZA

Obvykle se testuje NULOVÁ HYPOTÉZA (H0) jako specifický model statistické hypotézy. NULOVÁ HYPOTÉZA PŘEDPOKLÁDÁ STAV „NEEXISTENCE“ (ROZDÍLU) ČI STAV SHODY.

Hypotéza se zamítá:

Hypotézy lze zásadně prohlásit za falešné (tedy zamítnout jejich platnost), nikoliv však dokazovat jejich platnost. Hypotéza nemůže být přímo dokázána, nýbrž může být jen zamítnuta jí odporující (nulová) hypotéza.

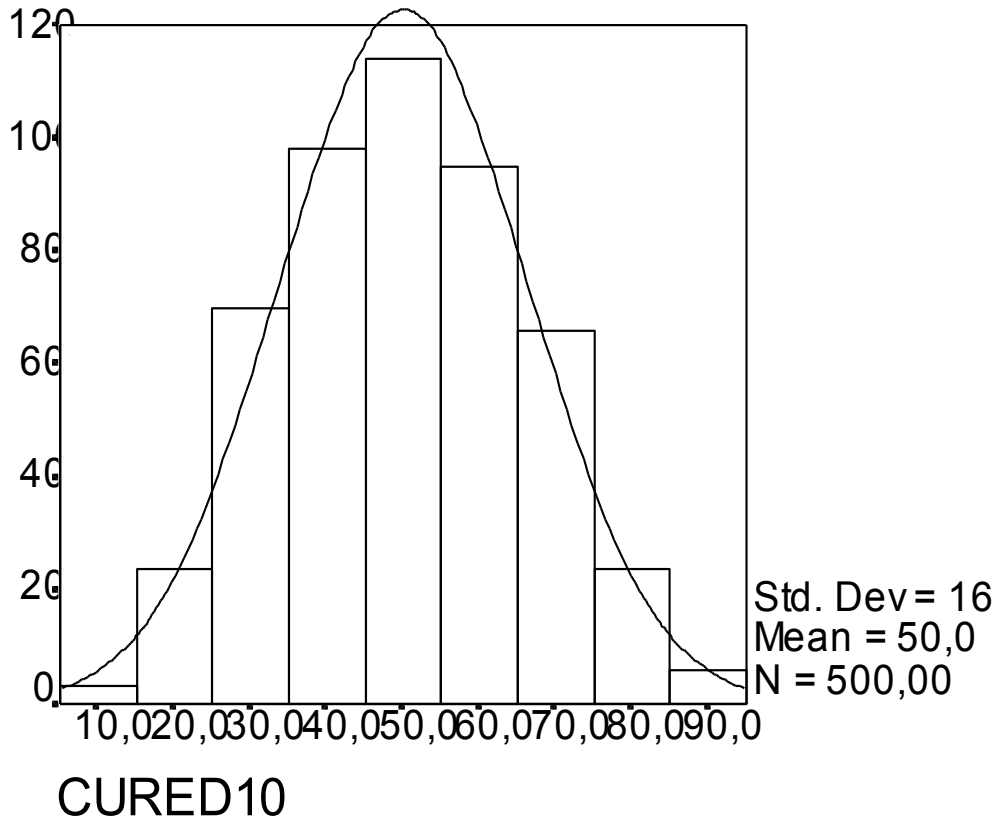
INFERENCE ZE STATISTIKY NA PARAMETR

- BODOVÝ ODHAD jako číslo, jehož hodnota je v nějakém (teoreticky) stanoveném smyslu optimálně určena.
- INTERVALOVÝ ODHAD, kdy hledáme interval (spolehlivosti), v kterém s určitou, předem zvolenou pravděpodobností neznámý populační parametr leží.

NORMÁLNÍ ROZLOŽENÍ - Gaussova křivka

(Karl Fridrich Gauss 1777-1855)

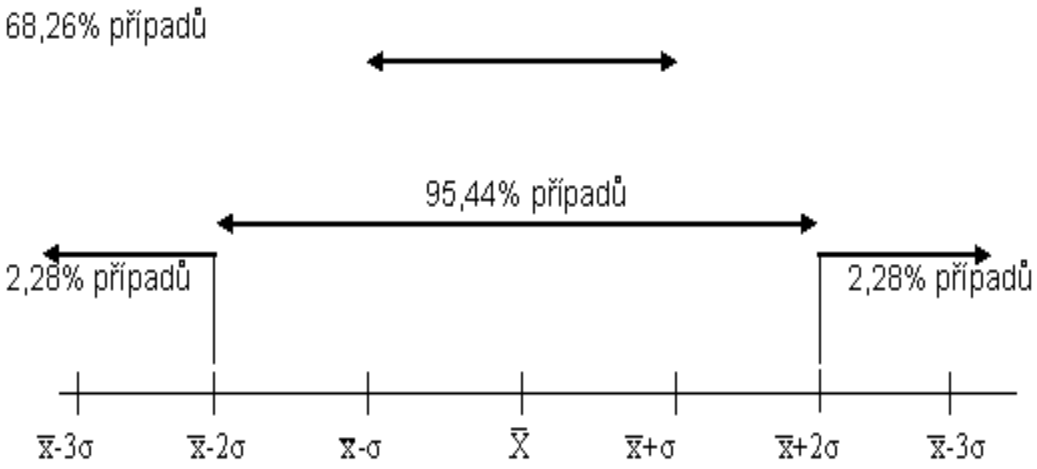
Pro mnoho náhodných veličin má rozdělení pravděpodobnostní tvar zvonu.



NORMÁLNÍ ROZLOŽENÍ

- Symetrie: polovina hodnot je větších než průměr a polovina hodnot je menší.
- Aritmetický průměr = medián = modus.
- Můžeme vždy vypočítat procento případů, spadajících do určitého intervalu kolem průměru.

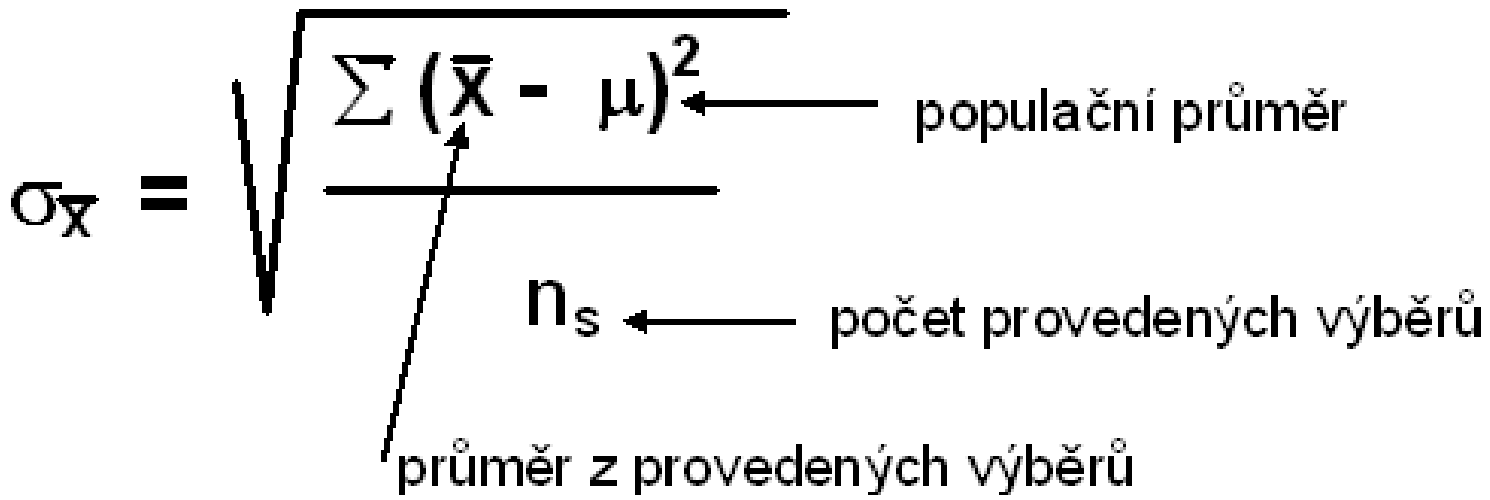
VLASTNOSTI


68,26% případů
95,44% případů
2,28% případů
2,28% případů
$\bar{x}$-3σ
$\bar{x}$-2σ
$\bar{x}$-σ
$\bar{X}$
$\bar{x}$+σ
$\bar{x}$+2σ
$\bar{x}$-3σ

Standardní odchylka

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$$

STANDARDNÍ CHYBA PRŮMĚRU



PROČ JE S.E. DŮLEŽITÁ?

S 95% pravděpodobností (5% riziko chyby) můžeme tvrdit, že:

průměr základního souboru (parametr)

= průměr výběrového souboru (statistika) ± 1,96 směrodatná chyby (často se zaokrouhluje na dvojnásobek)

S 99% pravděpodobností (1% riziko chyby) můžeme tvrdit, že:

průměr základního souboru (parametr)

= průměr výběrového souboru (statistika) ± 2,96 směrodatná chyby (často se zaokrouhluje na trojnásobek)

INTERVAL SPOLEHLIVOSTI

Protože pracujeme s výběrovými soubory, můžeme vypočítat statistiky, ale nevíme, jak tyto statistiky korespondují s parametry. Víme ovšem, že se - se zvolenou pravděpodobností - pohybují v intervalu (spolehlivosti), jehož obecný vzorec je:

$$C.I = X \pm z . \sigma X$$

X = vypočítaný výběrový průměr (statistika)

z = z-skóre korespondující s požadovanou úrovní pravděpodobnosti (hladinou významnosti). Pro HV=95% je to 1,96.

σX = standardní/směrodatná chyba distribuce výběrových průměrů

Interval spolehlivosti pro 95% HV znamená:

Jestliže bychom z populace opakovaně činili výběry stejné velikosti, v 95% z nich výběrů by se populační průměr nacházel uvnitř intervalu spolehlivosti (s 95% pravděpodobnost interval spolehlivosti tento populační průměr zahrnuje).

IS zavisí na:

- Velikosti souboru
- Heterogenitě souboru