Statistické metody - nástroj poznání a rozhodování anebo zdroj omylů a lží

Zdeněk Karpíšek

“Jsou tři druhy lží: lži, odsouzeníhodné lži a statistiky.”

“Statistika je logická a přesná metoda, jak nepřesně sdělit polopravdu.”

Od vědy o státu a hazardních her k matematické statistice

Statistika:

- číselné údaje o hromadných jevech
- sběr, zpracování a vyhodnocování statistických údajů
- teoretická disciplína

Tři etapy statistického zkoumání:

- získávání statistických údajů statistickým zjišťováním (šetřením)
- kontrola, verifikace, třídění a agregace statistických souborů
- vyhodnocování a rozbor získaných statistických údajů

Aplikace výsledků + zpětná vazba

Od vědy o státu a hazardních her k matematické statistice

Popisná statistika:

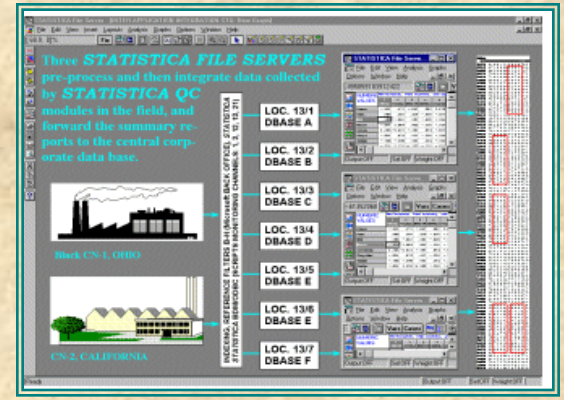
- vytvoření a třídění statistických souborů
- číselné a grafické zpracování

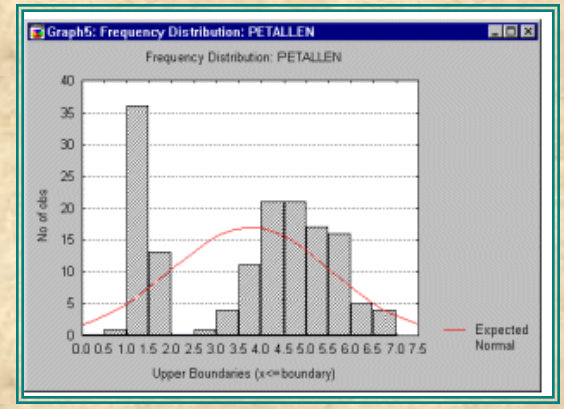
Teorie pravděpodobnosti:

- náhodné jevy a pravděpodobnost
- náhodné veličiny, vektory a procesy

Matematická statistika:

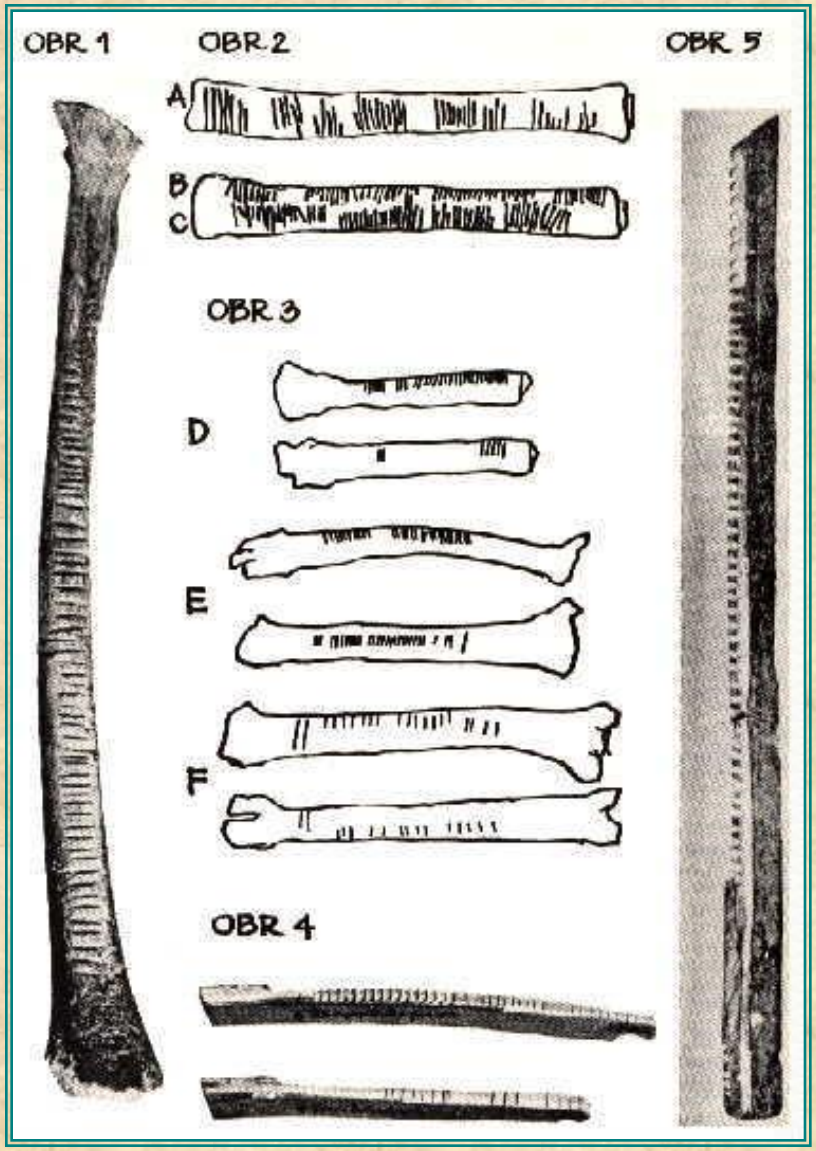
- náhodný výběr
- odhady parametrů a rozdělení
- testování statistických hypotéz

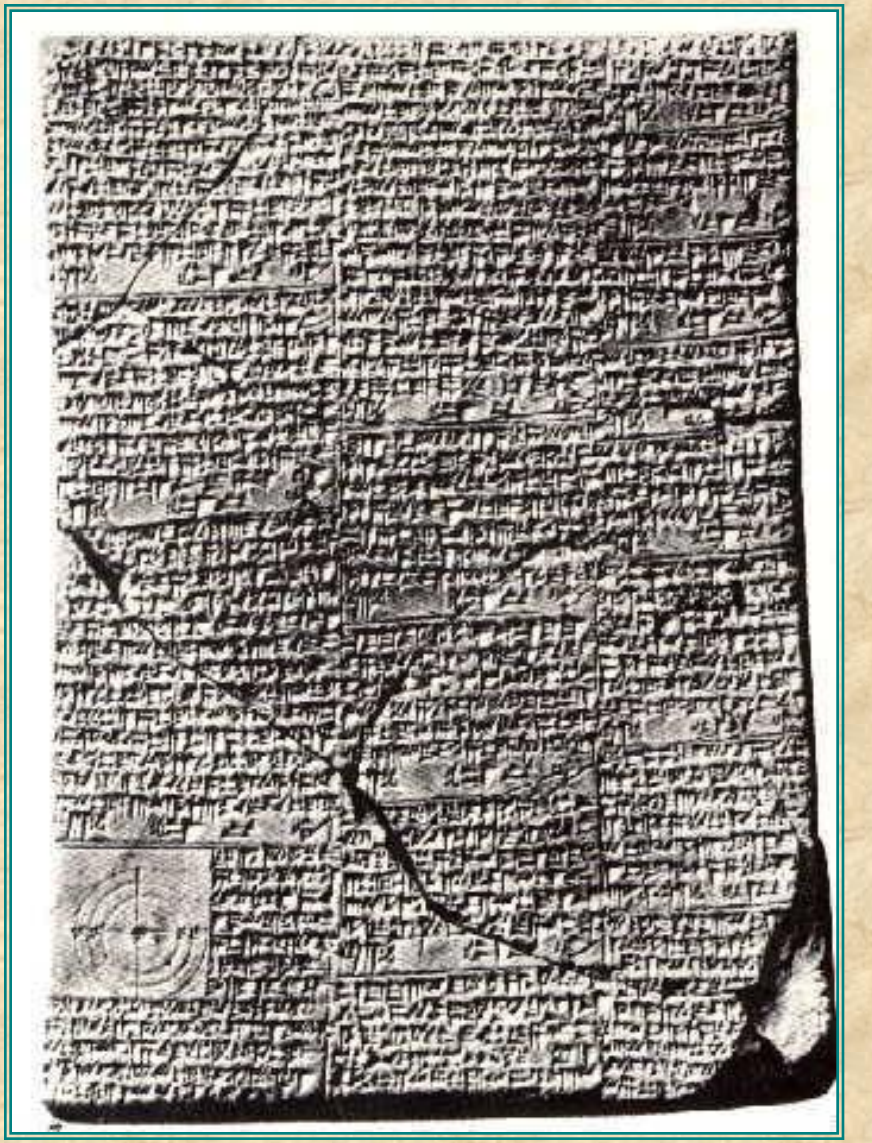




Od vědy o státu a hazardních her k matematické statistice

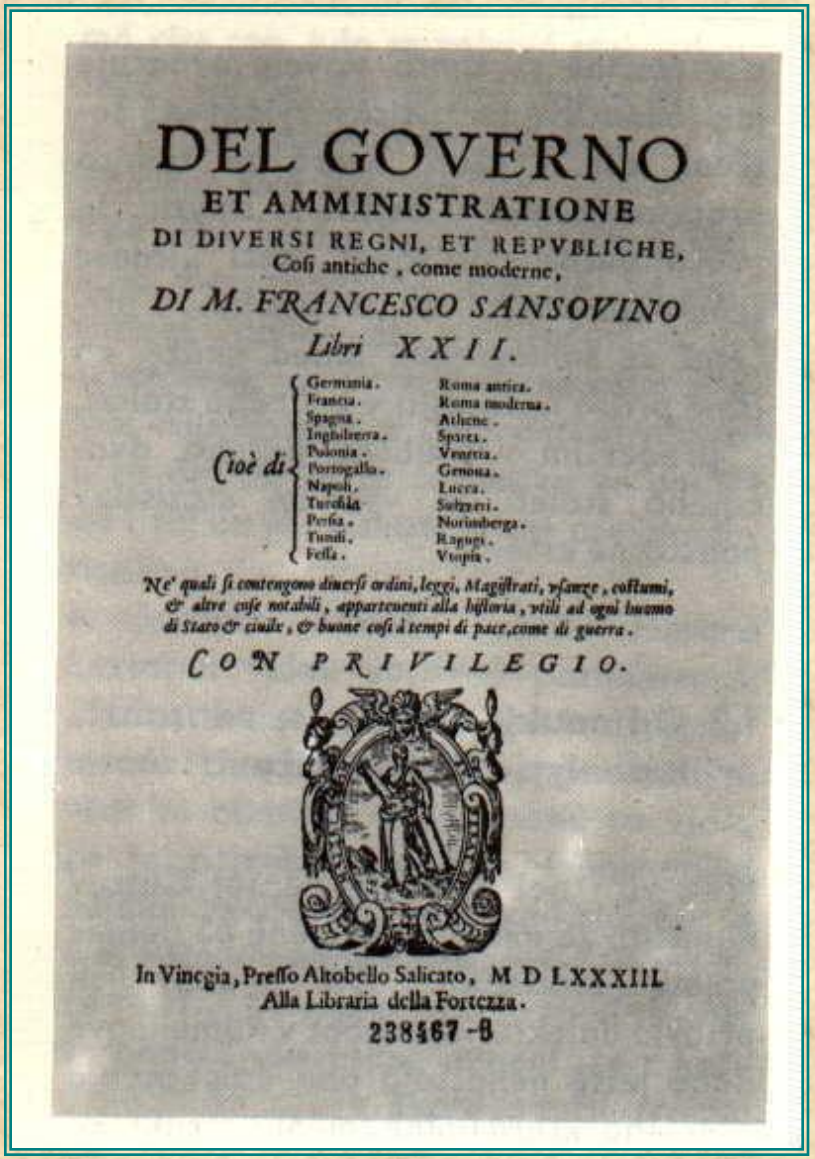
Historie:





Od vědy o státu a hazardních her k matematické statistice

Historie:



DEL GOVERNO
ET AMMINISTRATIONE
DI DIVERSI REGNI, ET REPVBLICHE,
Cosi antiche, come moderne,
DI M. FRANCESCO SANSOVINO
Libri XXII.

Cioè di

Germania.	Roma antica.
Francia.	Roma moderna.
Spagna.	Athene.
Inghilterra.	Sparta.
Polonia.	Venetia.
Portogallo.	Genoua.
Napoli.	Lucca.
Turchia	Suizzeri.
Persia.	Norimberga.
Tunisi.	Ragusi.
Fessa.	Vtopia.

Ne' quali si contengono diuersi ordini, leggi, Magistrati, vsanze, costumi, & altre cose notabili, appartenenti alla historia, vtili ad ogni huomo di Stato & ciuile, & buone cosi à tempi di pace, come di guerra.

CON PRIVILEGIO.

In Vinegia, Presso Altobello Salicato, M D LXXXIII
Alla Libraria della Fortezza.

238467-B

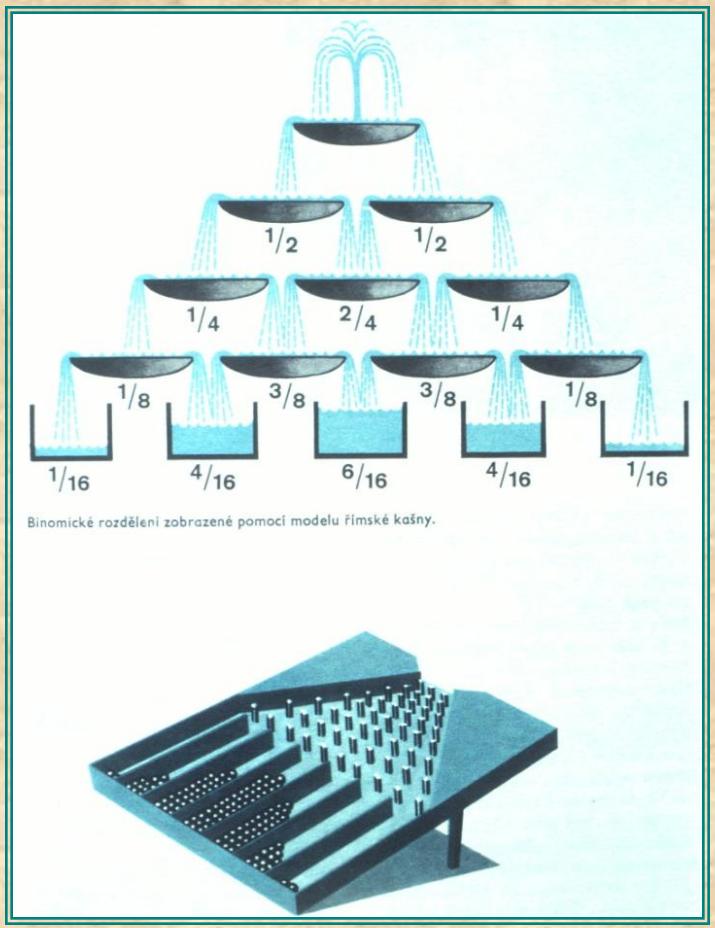
SCIENTIFIC AMERICAN

A WEEKLY JOURNAL OF PRACTICAL INFORMATION, ART, SCIENCE, MECHANICS, CHEMISTRY, AND MANUFACTURES.

NEW YORK, AUGUST 30, 1890.

Od vědy o státu a hazardních her k matematické statistice

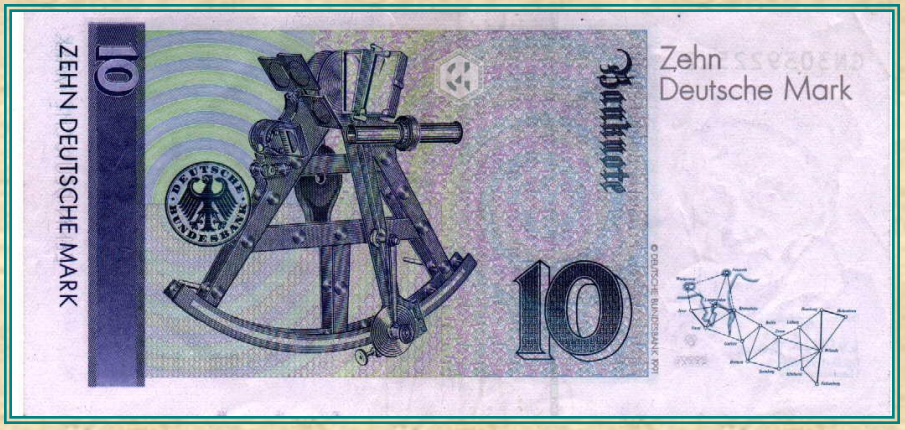
Historie:

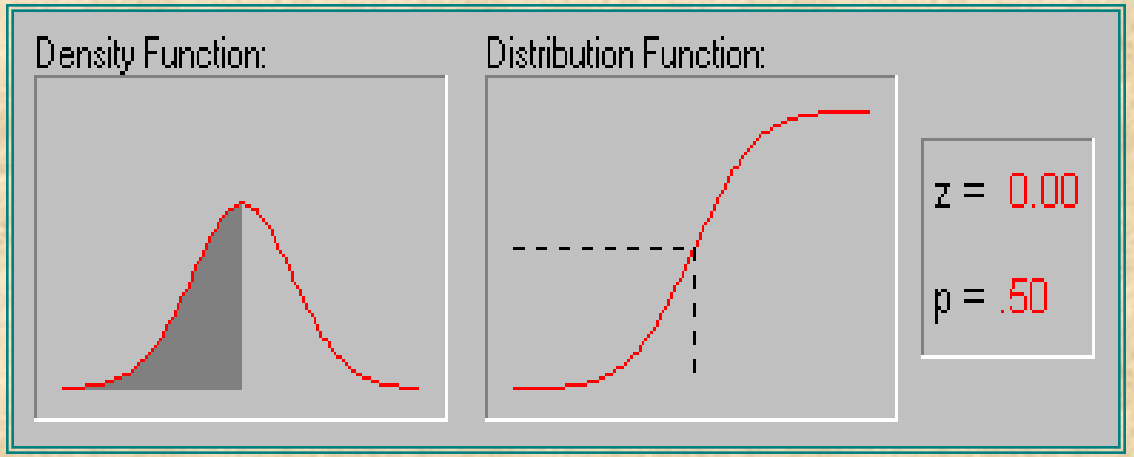


Binomické rozdělení zobrazené pomocí modelu římské kašny.

Od vědy o státu a hazardních her k matematické statistice

Historie:

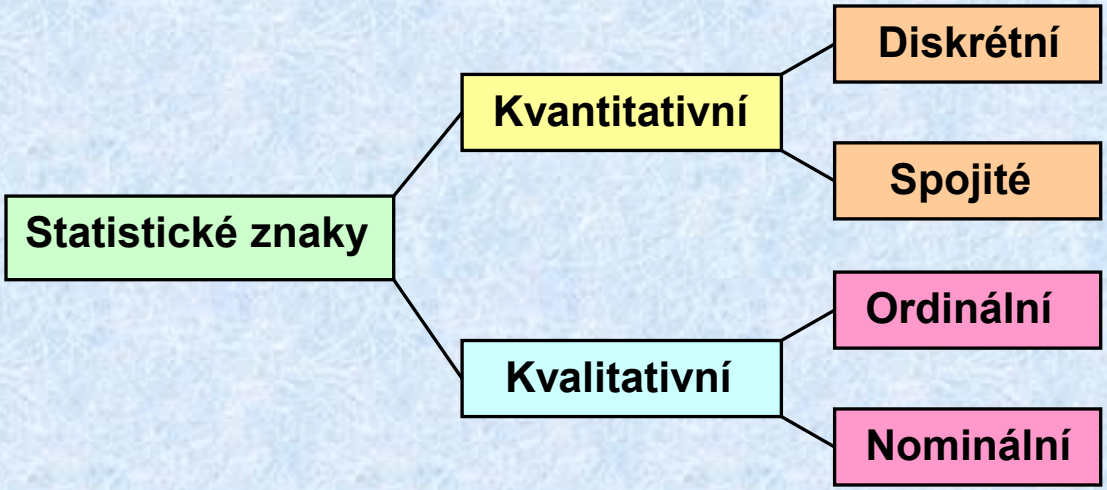


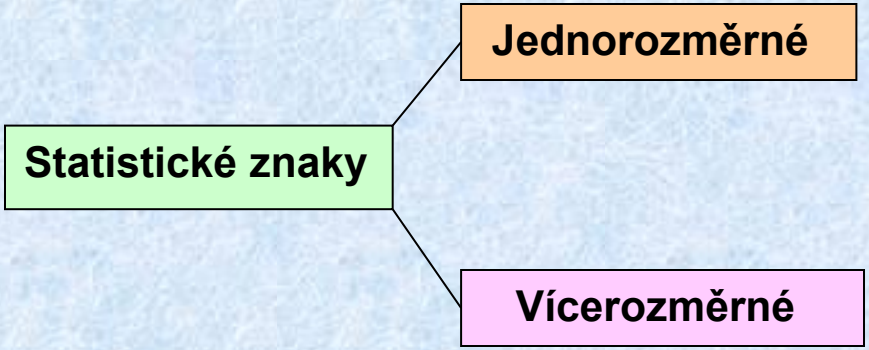


Populace, výběr, náhoda a neurčitost patří k sobě

Základní soubor (populace) = souhrn statistických jednotek

Statistické jednotky - statistické znaky - hodnoty





Populace, výběr, náhoda a neurčitost patří k sobě

Základní soubor → výběrový soubor, rozsah

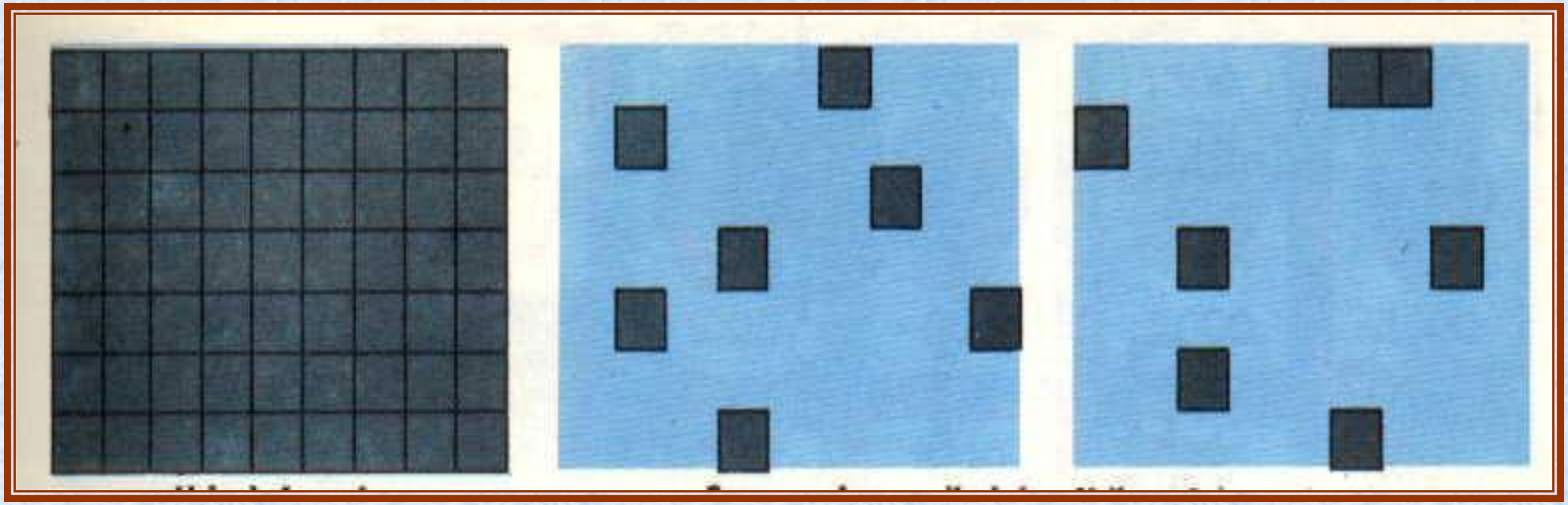
Výběry:

- malé (*obvykle do 30 až 50*)
- velké (*řádově stovky, tisíce i více*)

Požadavky na výběr:

- **reprezentativní** (*informace bez omezení*)
- **homogenní** (*bez vlivu dalších faktorů*)
- **náhodný**

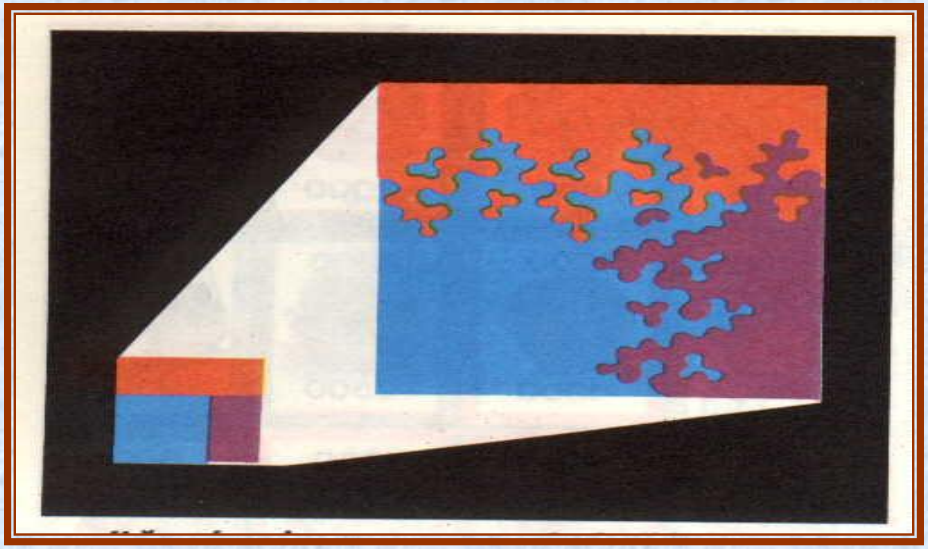
Neurčitost výběru = zkreslení informací o základním souboru



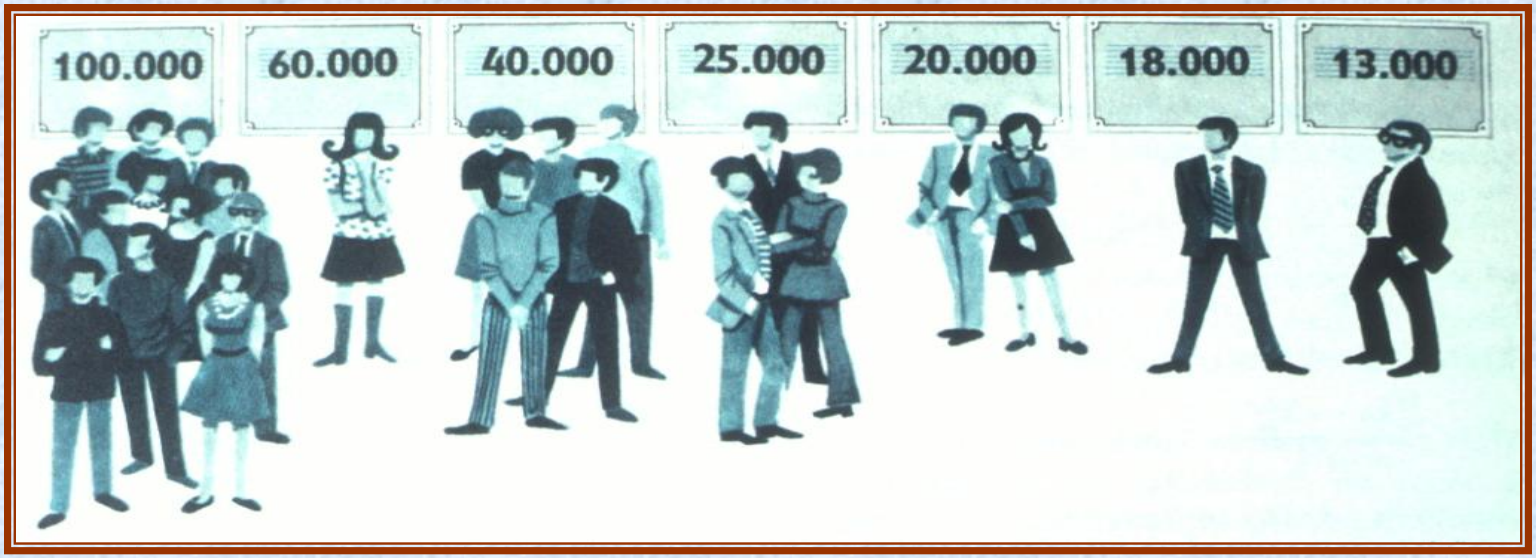
Populace, výběr, náhoda a neurčitost patří k sobě

Druhy výběrů:

- **bez opakování**
- **s opakováním**
- **záměrný**
- **oblastní**
- **mechanický**



Statistický soubor = soubor pozorovaných hodnot (x_1, x_2, …, x_n) znaku X na vybraných statistických jednotkách

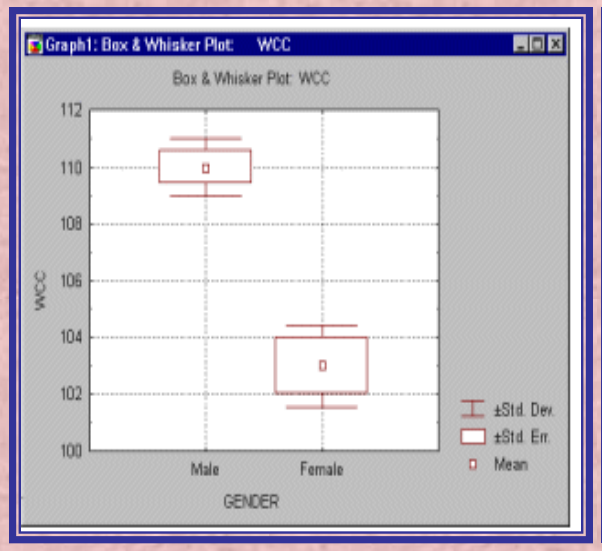


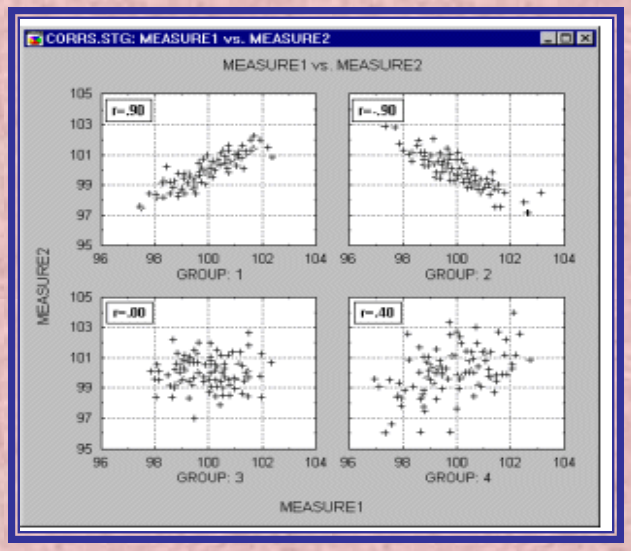
Průměry a jejich ctnosti i nectnosti

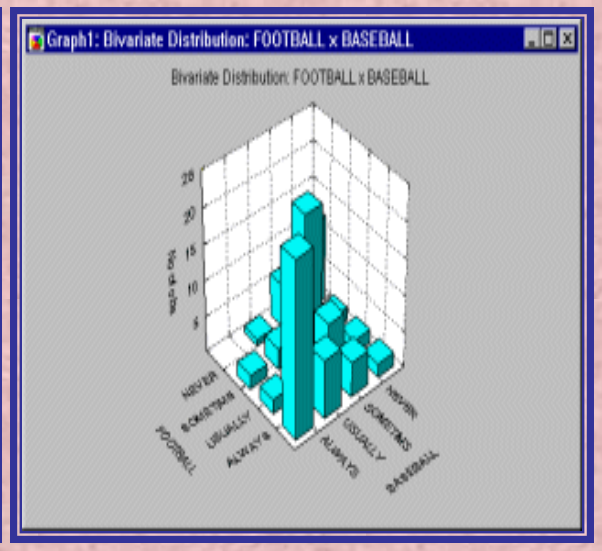
Zpracování statistického souboru = příprava + grafické znázornění + výpočet číselných charakteristik

Roztříděný soubor - třídy, střed a četnost

Grafy = vizuální informace o poloze, variabilitě, symetrii, modalitě,…







Průměry a jejich ctnosti i nectnosti

Číselné (empirické) charakteristiky = číselné informace o poloze, variabilitě, symetrii, modalitě,…

Aritmetický průměr = $(x_1 + x_2 + ...+ x_n)/n$

Medián = prostřední hodnota uspořádaného souboru

Příklad:

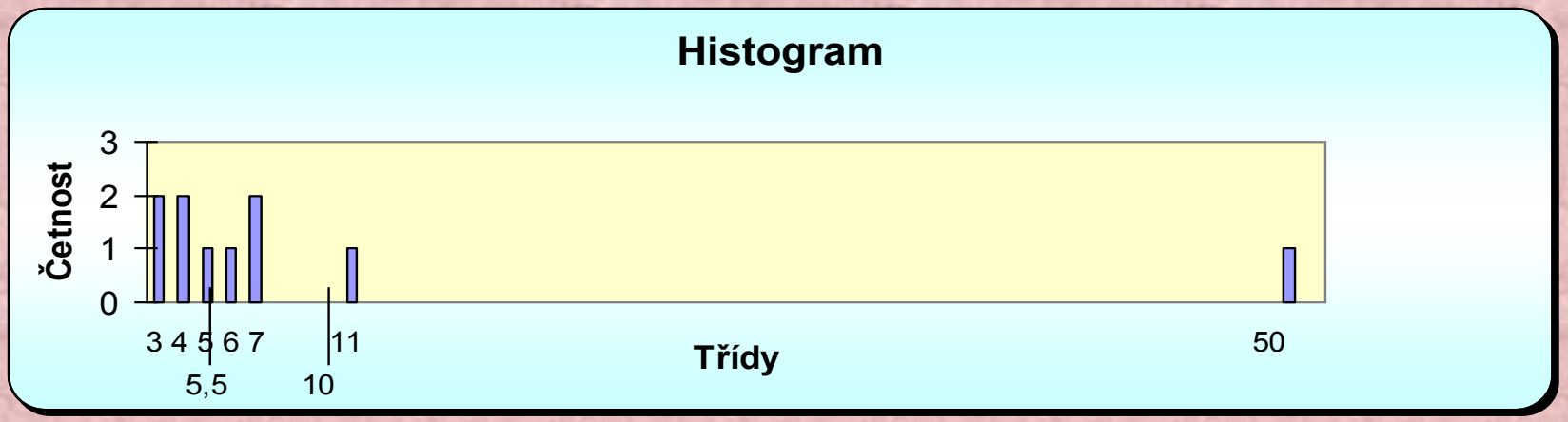
Měsíční platy 10 pracovníků (v tis. Kč): 3, 3, 4, 4, 5, 6, 7, 7, 11, 50.

Průměrný měsíční plat = 100/10 = 10.

Medián měsíčního platu = 5 až 6 $\approx$ (5 + 6)/2 = 5,5.

Průměrný měsíční plat po změně (50 na 100) = 150/10 = 15.

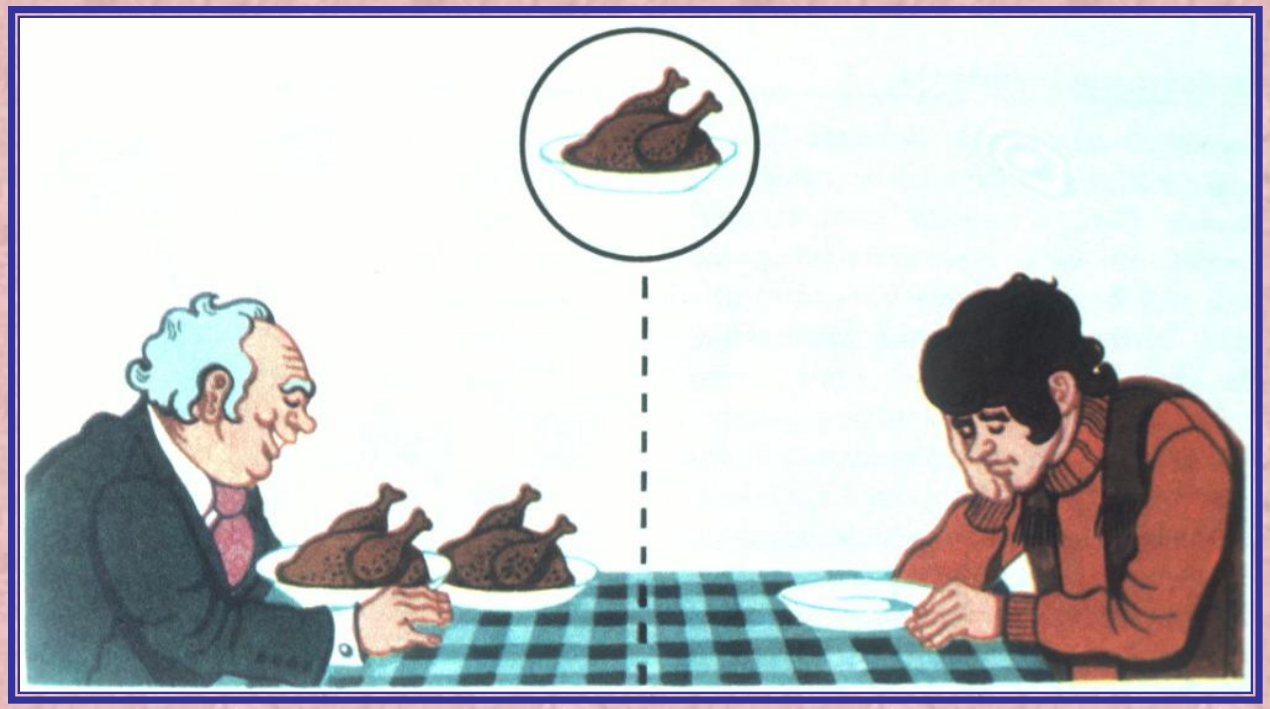
Medián měsíčního platu po změně (50 na 100) = 5 až 6 $\approx$ (5 + 6)/2 = 5,5.



Průměry a jejich ctnosti i nectnosti

Vlastnosti průměru:

- **poměrně citlivý na změnu hodnot souboru**
- **u kladně (záporně) asymetrických souborů je průměr větší (menší) než medián**
- **konvergence s rostoucím rozsahem souboru k průměru celé populace**
- **rychlá konvergence rozdělení pravděpodobnosti průměru k normálnímu rozdělení**



Měříme proměnlivost veličin a vztahy mezi nimi

Rozptyl (disperze) = průměr kvadrátů odchylek od průměru

Směrodatná odchylka = druhá odmocnina z rozptylu

Příklad:

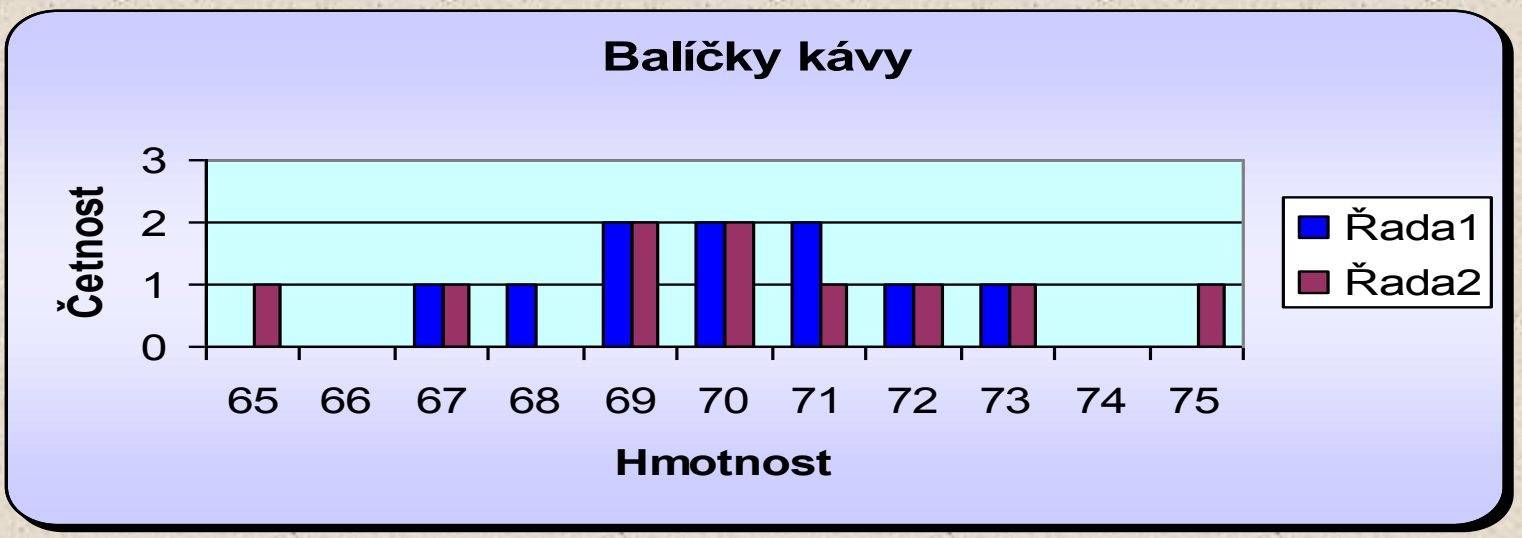
Dva soubory hmotnosti balíčku kávy (v gramech):

1. automat: 67; 68; 69; 69; 70; 70; 71; 71; 72; 73
2. automat: 65; 67; 69; 69; 70; 70; 71; 72; 73; 75

Průměrná hmotnost je stejná = 70

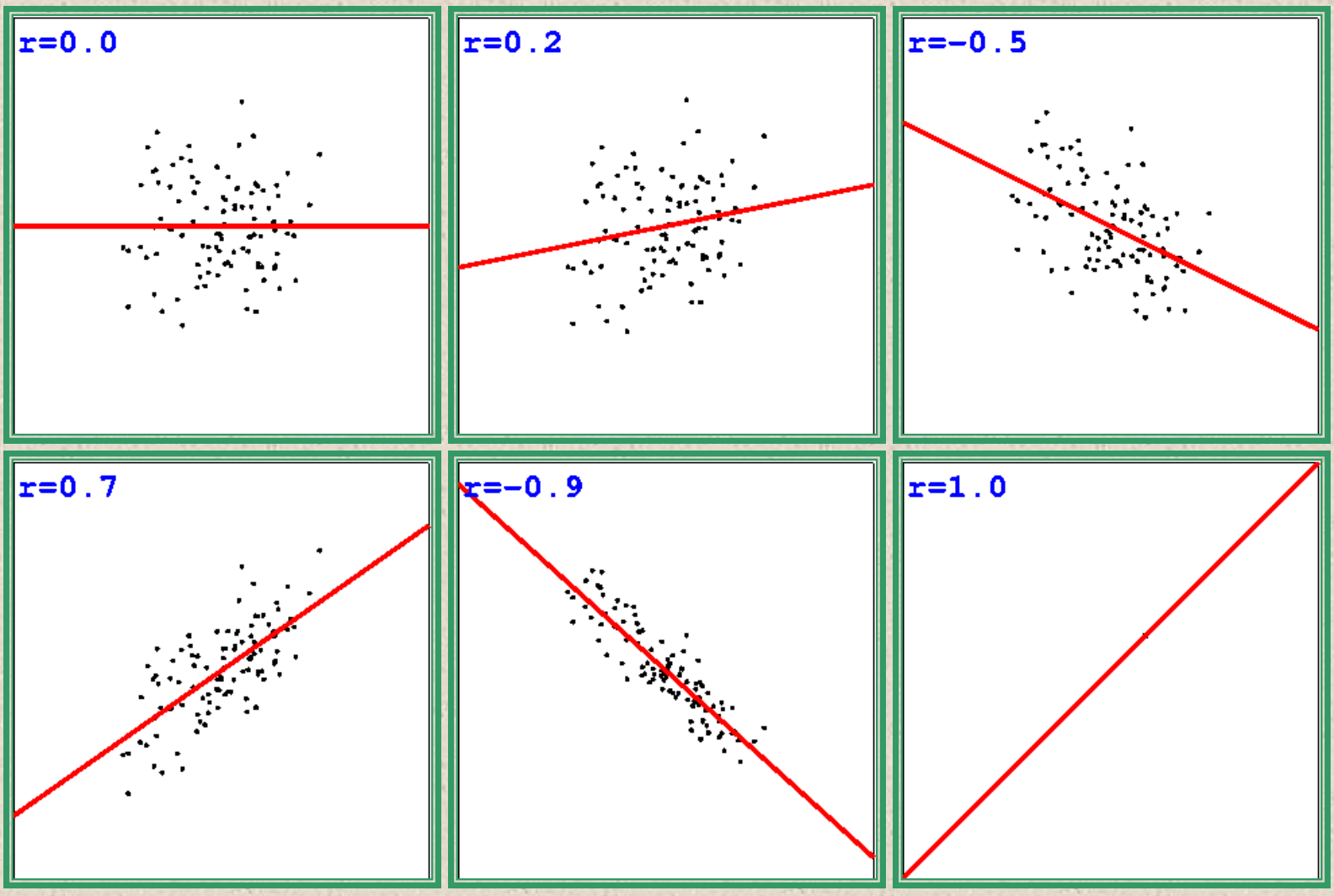
Rozptyl (směrodatná odchylka) pro 1. automat = 3,00 (1,732)

Rozptyl (směrodatná odchylka) pro 2. automat = 8,22 (2,867)



Měříme proměnlivost veličin a vztahy mezi nimi

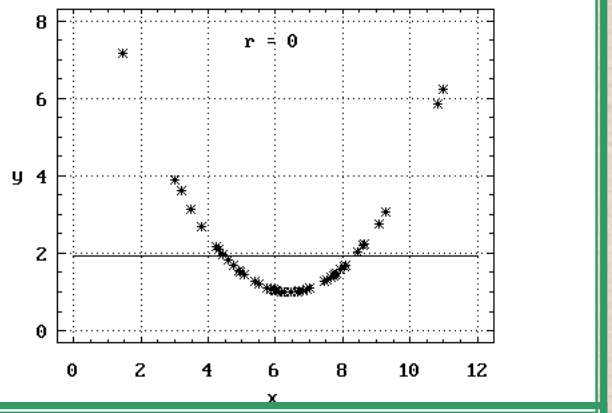
Koeficient korelace = normovaná kovariance = průměr součinů odchylek od průměrů/součin směrodatných odchylek (hodnoty od -1 do 1)

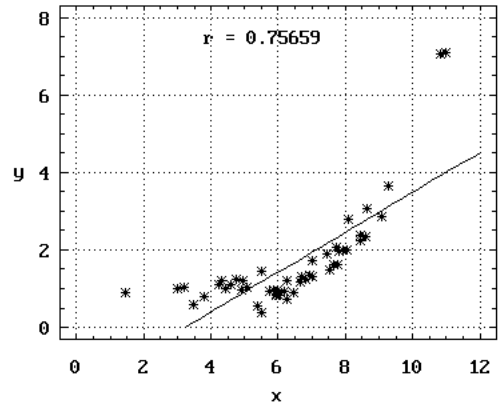


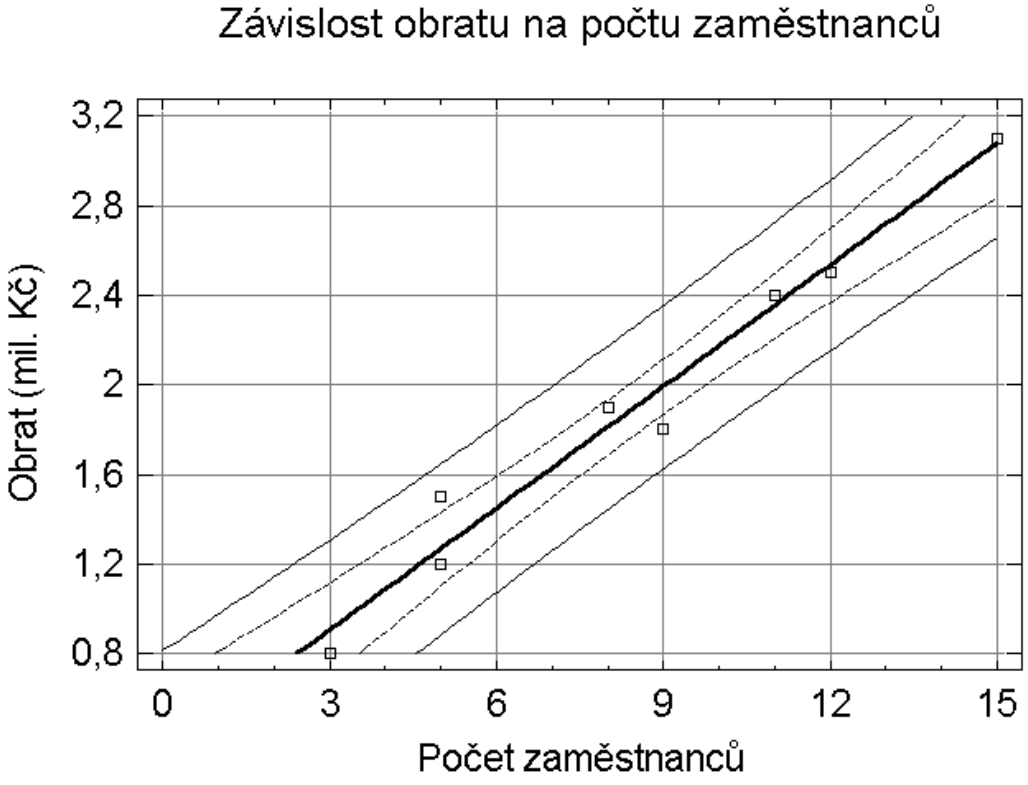
Měříme proměnlivost veličin a vztahy mezi nimi

Koeficient korelace = 0 nemusí znamenat nezávislost

Regresní analýza = "jemnější" vyjádření závislosti mezi znaky a predikce



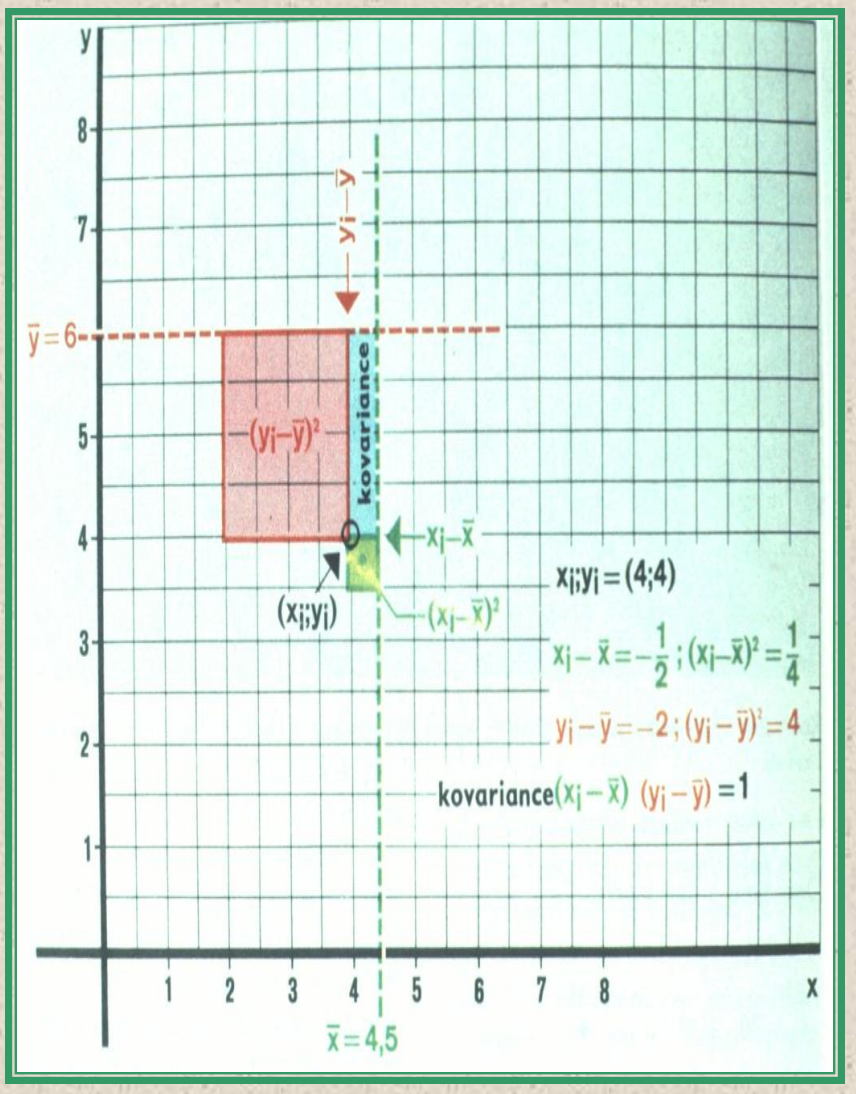


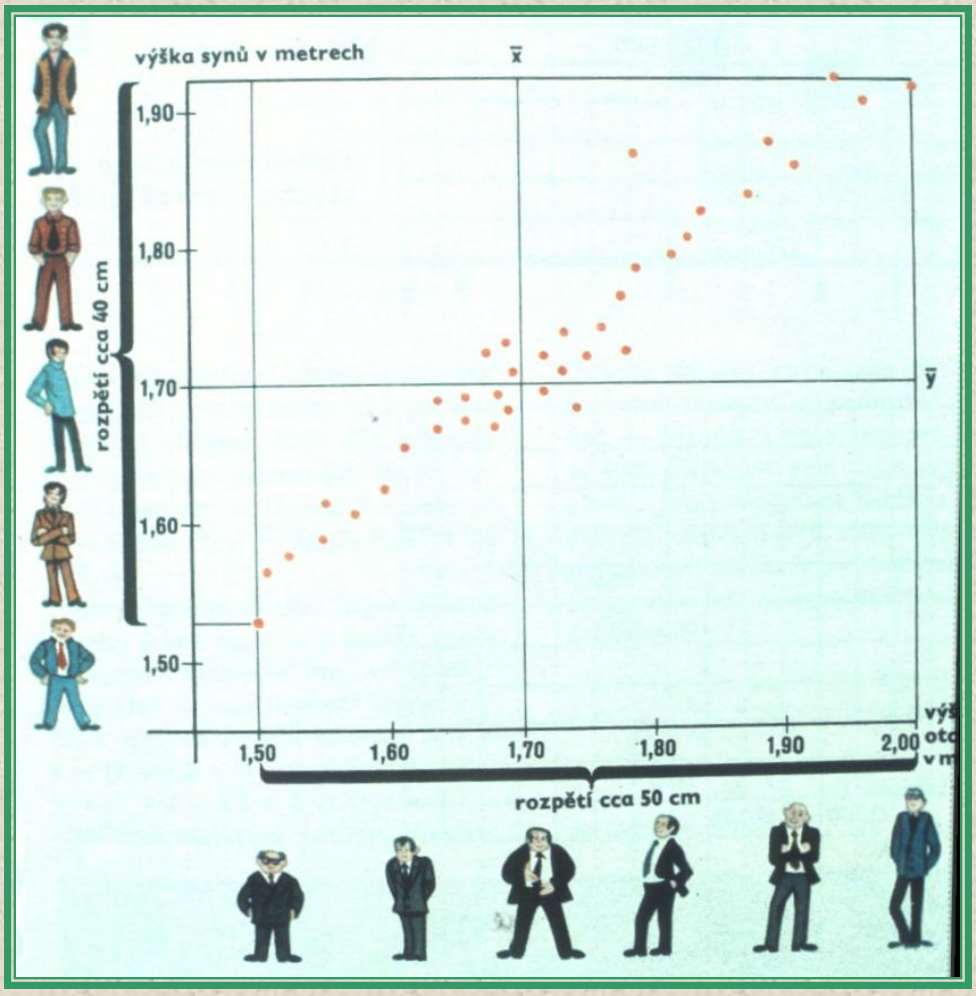


$y = 0{,}361 + 0{,}181x$; $y(10) \approx 2{,}172$; $r = 0{,}984798$

$y(10) \in \langle 2{,}040 ; 2{,}302 \rangle$; $y(10) \in \langle 1{,}804 ; 2{,}539 \rangle$

Měříme proměnlivost veličin a vztahy mezi nimi





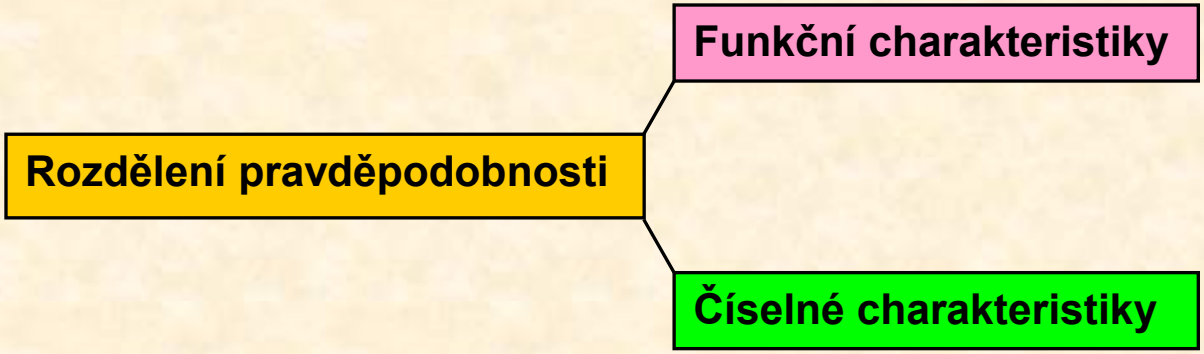
Rozdělení pravděpodobnosti a zákon velkých čísel jsou obrazem nás i okolního světa

Pravděpodobnost P(A) - teoretická míra možnosti nastoupení náhodného jevu A

Klasická definice: P(A) = m/n

- **m = počet příznivých případů jevu A**
- **n = počet všech možných případů**

Axiomatická definice - založená na teorii množin



Funkční charakteristiky: distribuční funkce, hustota aj.

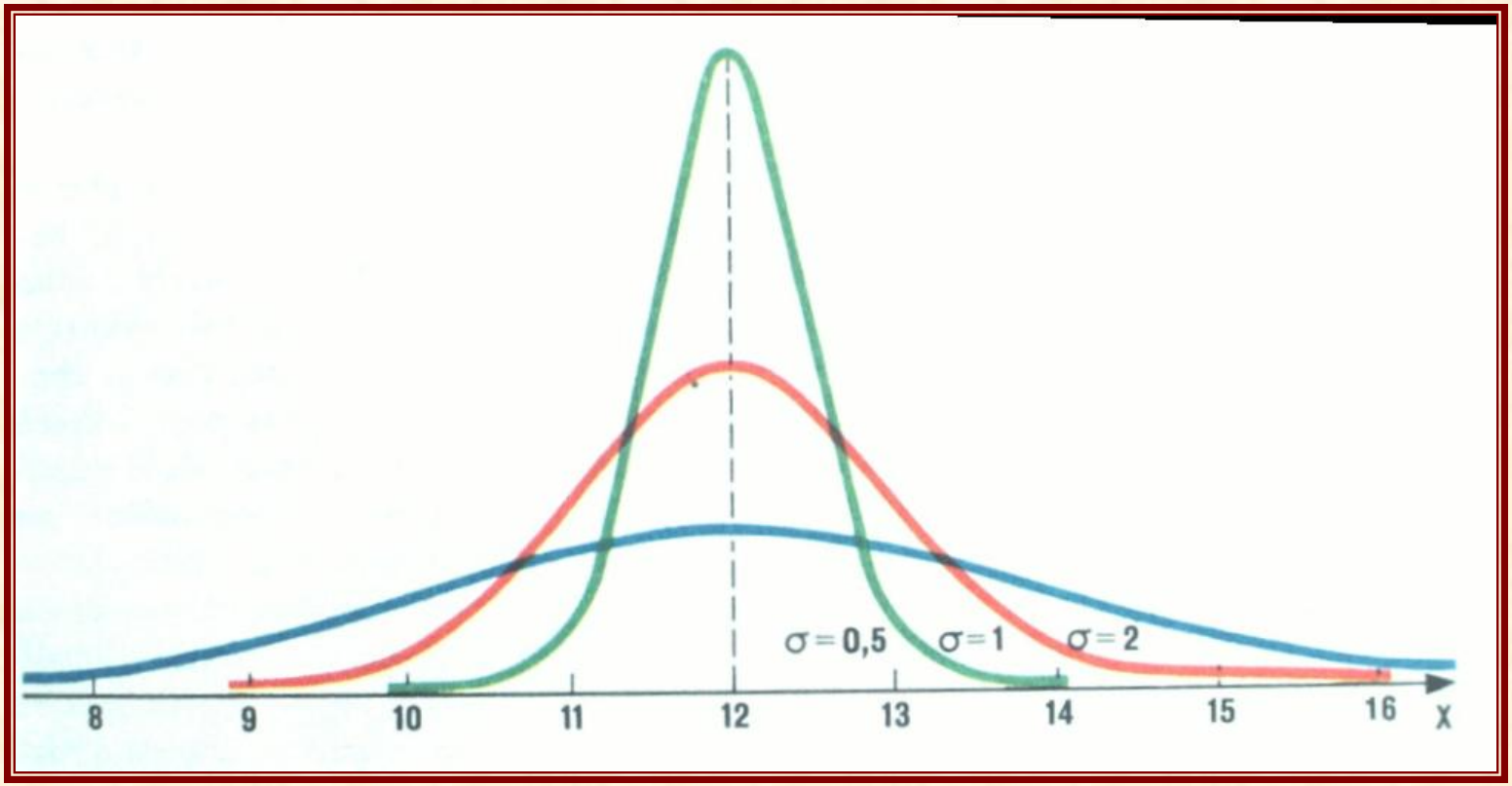
Číselné charakteristiky: střední hodnota, rozptyl aj.

Rozdělení pravděpodobnosti pro modelování reálných jevů: binomické, hypergeometrické, normální aj.

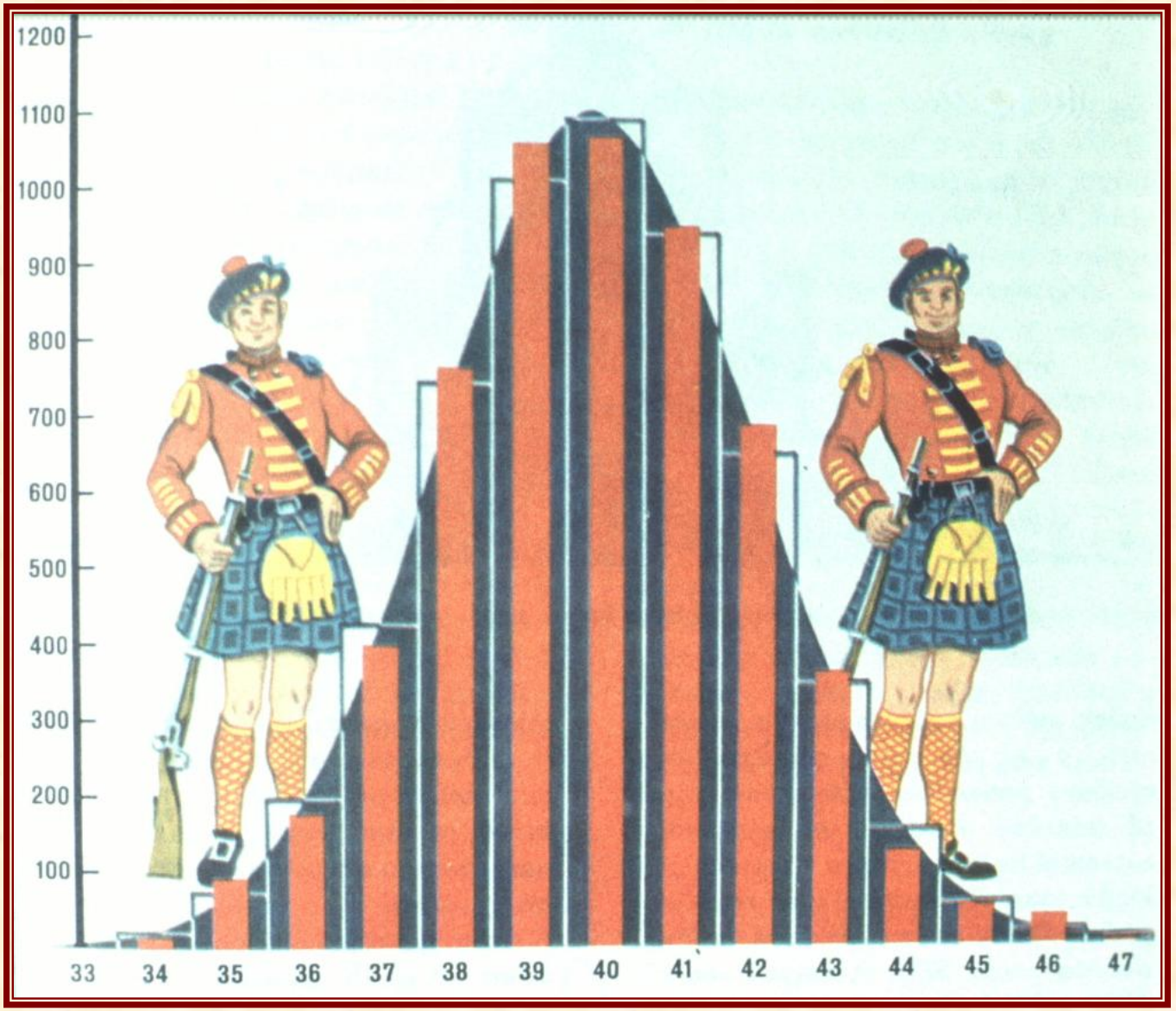
Rozdělení pravděpodobnosti a zákon velkých čísel jsou obrazem nás i okolního světa

Bernoulliův zákon velkých čísel - asymptotické chování relativní četnosti

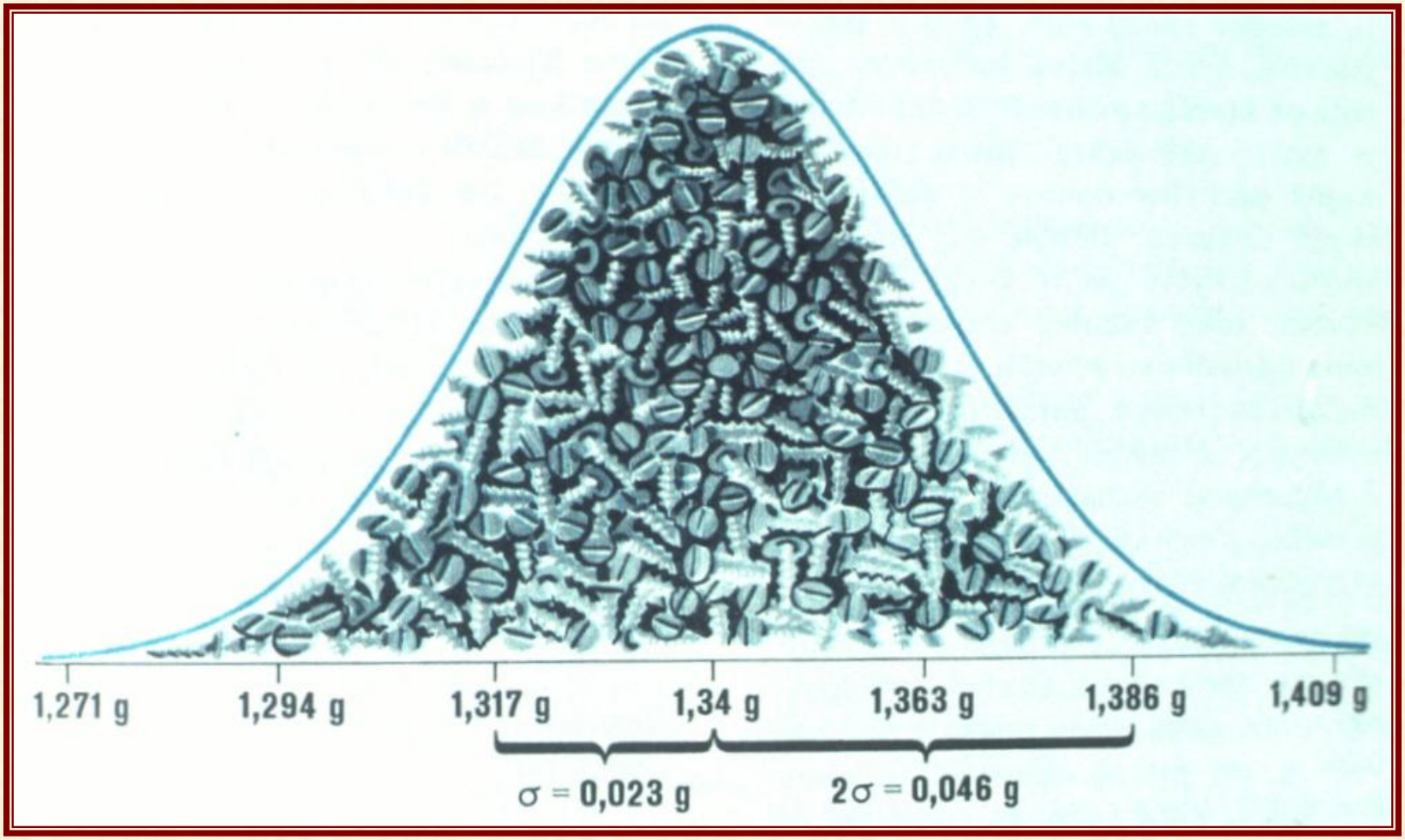
Normální rozdělení - významné postavení při modelování reálného světa



Rozdělení pravděpodobnosti a zákon velkých čísel jsou obrazem nás i okolního světa



Rozdělení pravděpodobnosti a zákon velkých čísel jsou obrazem nás i okolního světa



Výběrové charakteristiky a kletba statistikova

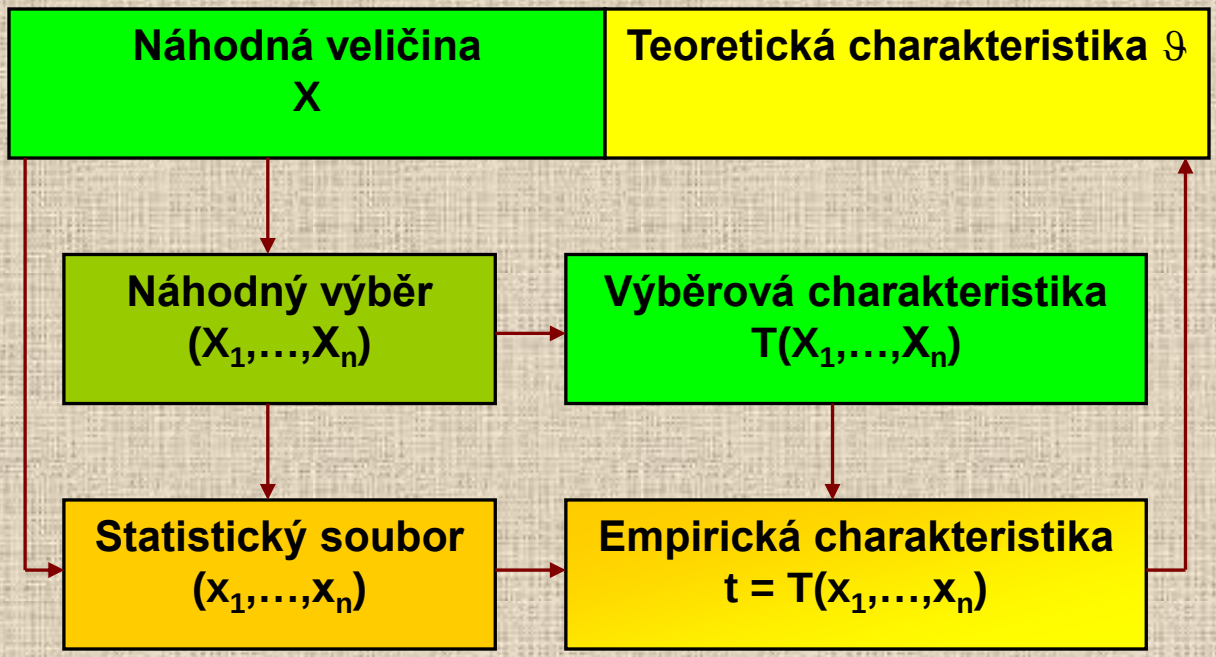
Základní úlohy matematické statistiky:

- odhady parametrů a rozdělení
- testování statistických hypotéz o parametrech a rozděleních

Principy matematické statistiky:

- hodnoty získané výběrem ze základního souboru jsou náhodné
- získaný statistický soubor je hodnotou náhodného výběru

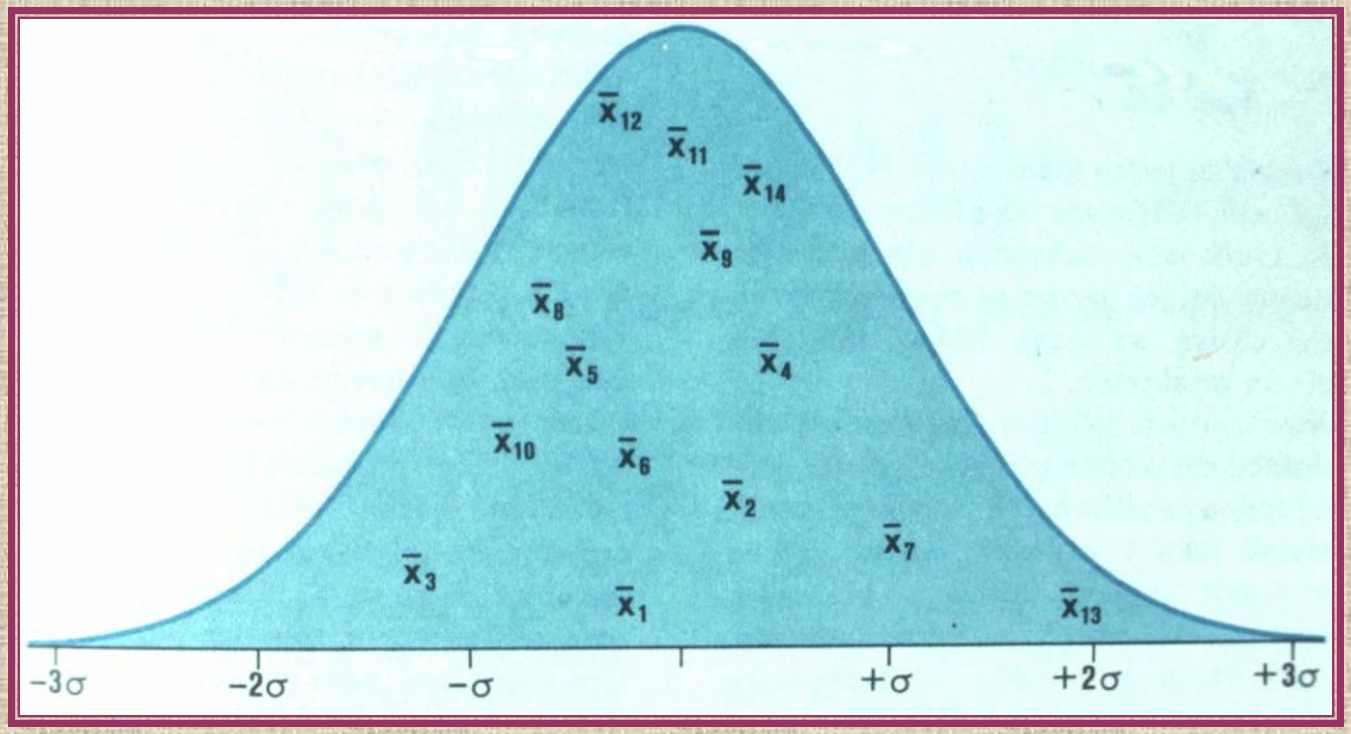
Statistická indukce:



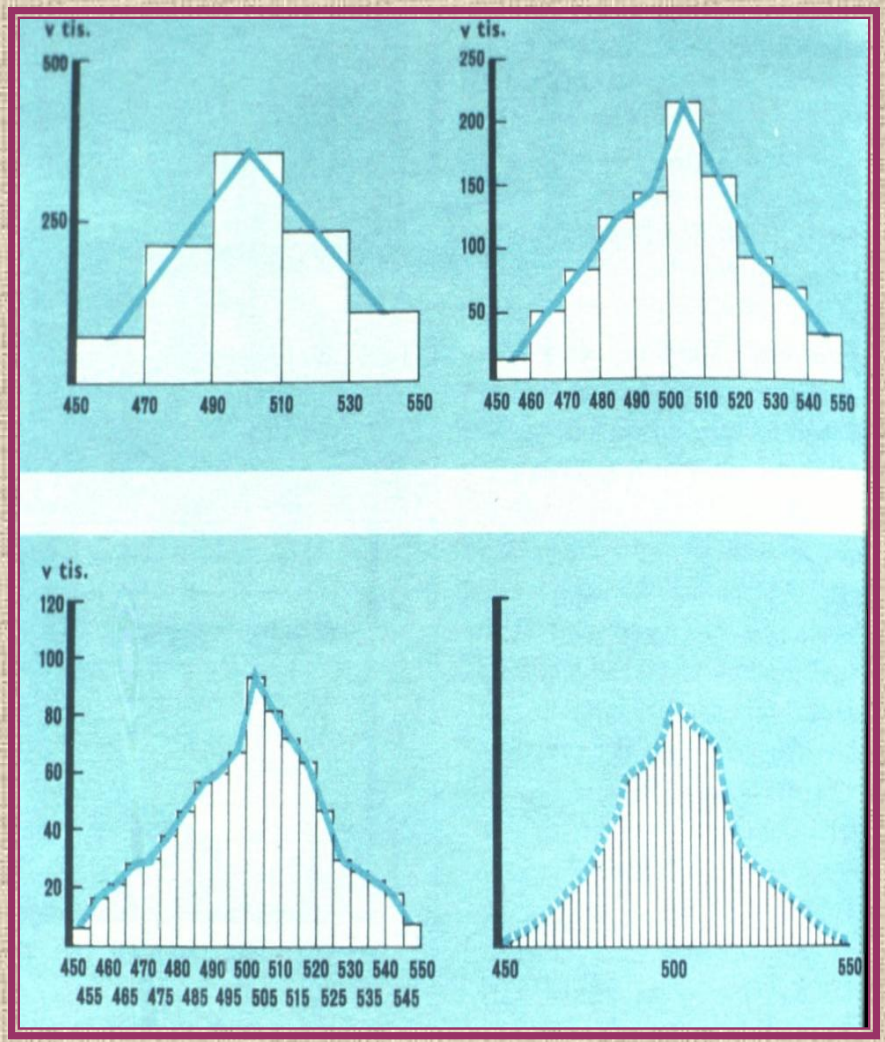
Výběrové charakteristiky a kletba statistikova

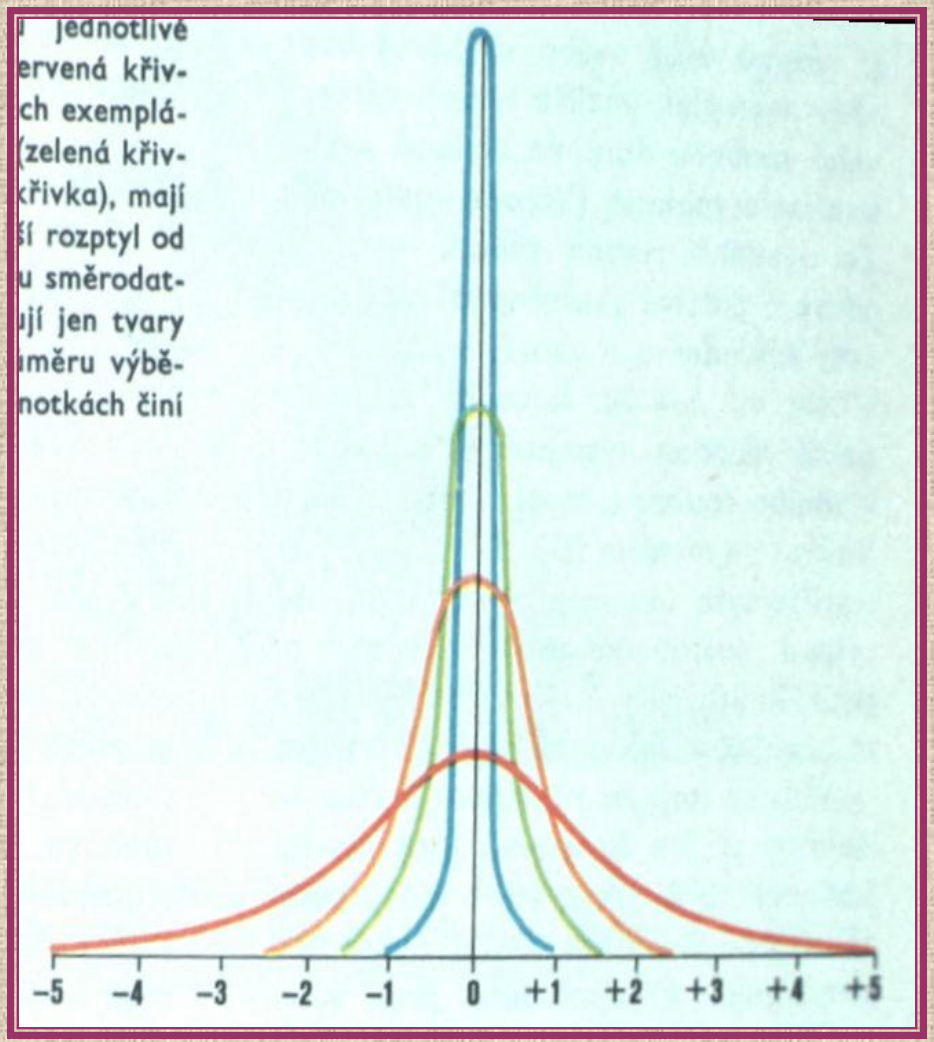
Např.:

Střední hodnota výběrového průměru = střední hodnota pozorované veličiny ("průměru" populace) a rozptyl výběrového průměru $\to 0$ pro $n \to \infty$, takže pro dostatečně velké n je takřka jistě průměr souboru blízký neznámé střední hodnotě; avšak tento rozptyl $\to 0$ s rychlostí $n^{1/2}$. Velmi často však rozdělení výběrového průměru konverguje k rozdělení normálnímu:



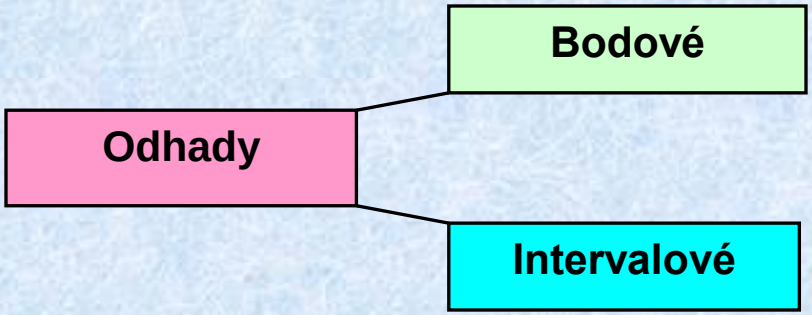
Výběrové charakteristiky a kletba statistikova





Je lepší odhad statistika anebo experta?

Odhad parametru ϑ = výběrová charakteristika $T(X_1,\ldots,X_n)$

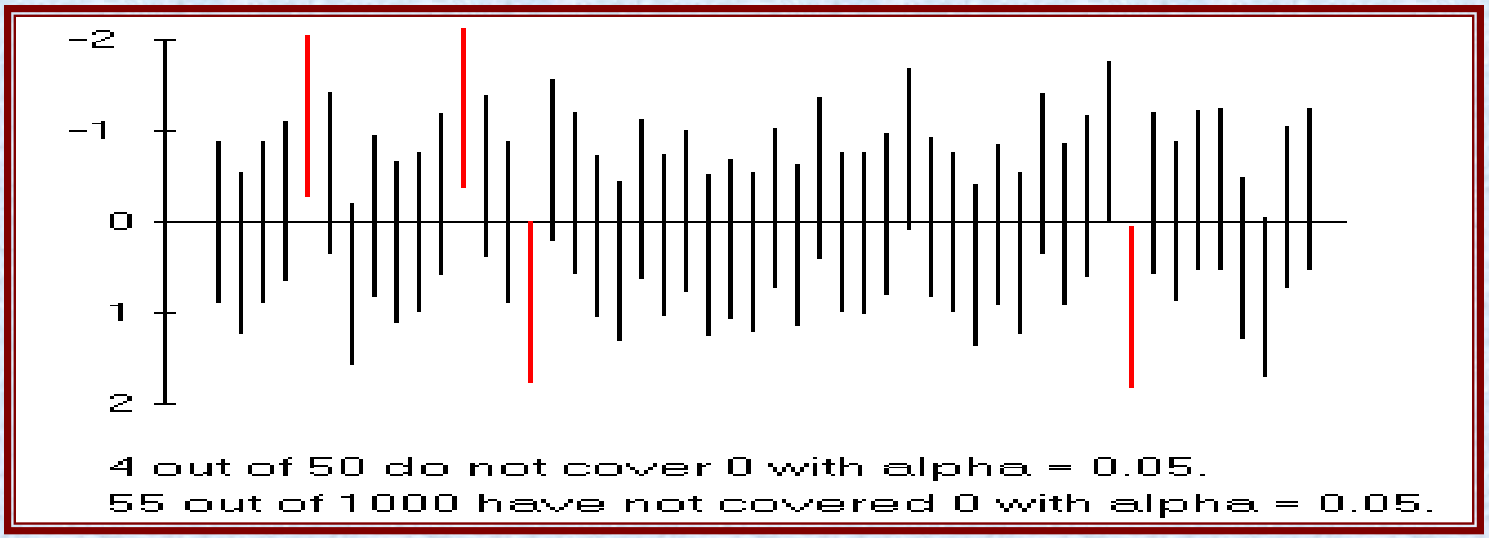


Bodový odhad ϑ: $t = T(x_1,..,x_n)$

Intervalový odhad ϑ se spolehlivostí 1 - α: konfidenční interval

$\langle T_1;T_2\rangle \Rightarrow \langle t_1;t_2\rangle$

Riziko chybného odhadu = α



Je lepší odhad statistika anebo experta?

Příklad:

Při průzkumu názoru z dotázaných n osob řeklo "ano" x osob. Pro spolehlivost 0,95:

n	x	Bodový odhad (%)	Intervalový odhad (%)	
			Od	Do
400	80	20	16,08	23,92
1600	320	20	18,04	21,96
6400	1280	20	19,02	20,98

Nezamítnutí statistické hypotézy ještě není potvrzení její správnosti

Statistické hypotézy = tvrzení o vlastnostech pozorovaného statistického znaku

Nulová hypotéza H_0 $\leftrightarrow$ Alternativní hypotéza H_A

Algoritmus testování hypotézy pomocí statistického souboru:

1. **Stanovení hypotéz H_0 a H_A.**
2. **Volba testového kritéria $T(X_1,\ldots,X_n)$.**
3. **Výpočet hodnoty testového kritéria $t = T(x_1,..,x_n)$.**
1. **Stanovení hladiny významnosti α a kritického oboru W_α.**
2. **Rozhodnutí o hypotézách H_0 a H_A.**

Hladina významnosti:

α = obvykle 5% anebo 1%

Nezamítnutí statistické hypotézy ještě není potvrzení její správnosti

Rozhodnutí:

- $t \in W_\alpha \Rightarrow H_0$ zamítáme a H_A nezamítáme
- $t \notin W_\alpha \Rightarrow H_0$ nezamítáme a H_A zamítáme

H_0	PLATÍ	NEPLATÍ
ZAMÍTÁME	CHYBA 1. DRUHU	-------
NEZAMÍTÁME	-------	CHYBA 2. DRUHU

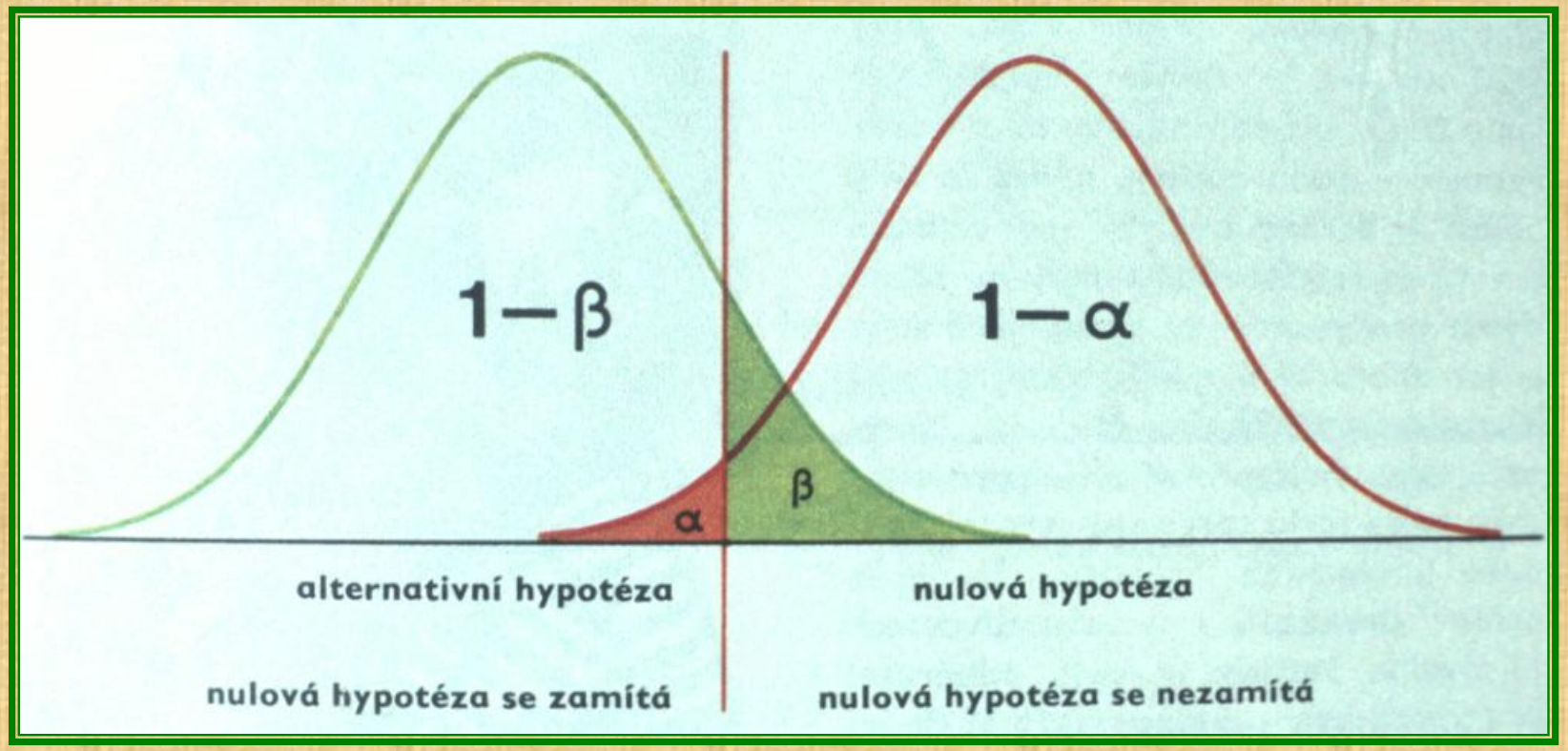
Rizika :

- pravděpodobnost chyby 1. druhu = hladina významnosti α
- pravděpodobnost chyby 2. druhu snižujeme zvýšením rozsahu

Aspekty:

- nezamítnutí hypotézy neznamená vždy její přijetí $\Rightarrow$
  $\Rightarrow$ zvýšíme rozsah výběru a znovu testujeme
- nezamítnutí nebo přijetí hypotézy není potvrzení její platnosti

Nezamítnutí statistické hypotézy ještě není potvrzení její správnosti

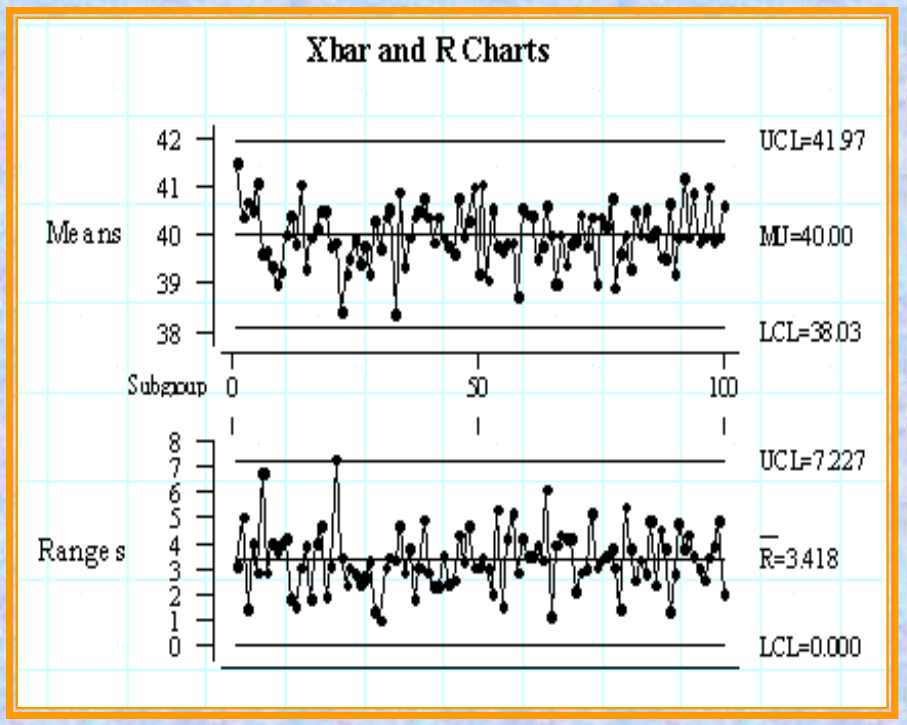


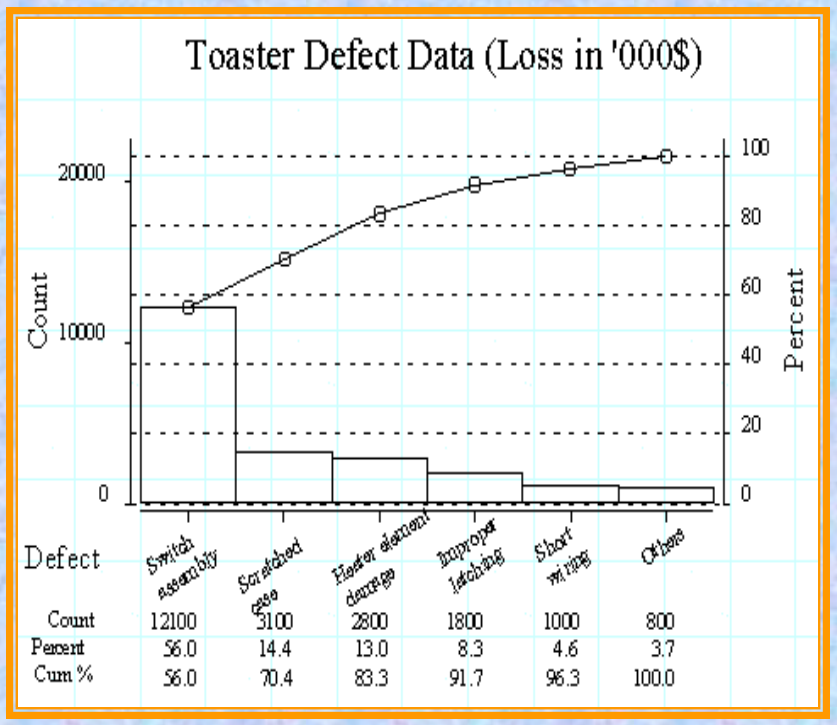
Používání statistických metod při řízení jakosti technologických procesů není samoúčelné ani módou

Statistické metody (SPC) = nástroje pro hodnocení a řízení jakosti výroby (QC)

Shewhartovy regulační diagramy:

- **regulace měřením (normální rozdělení)**
- **regulace porovnáváním (binomické a Poissonovo rozdělení)**





Používání statistických metod při řízení jakosti technologických procesů není samoúčelné ani módou

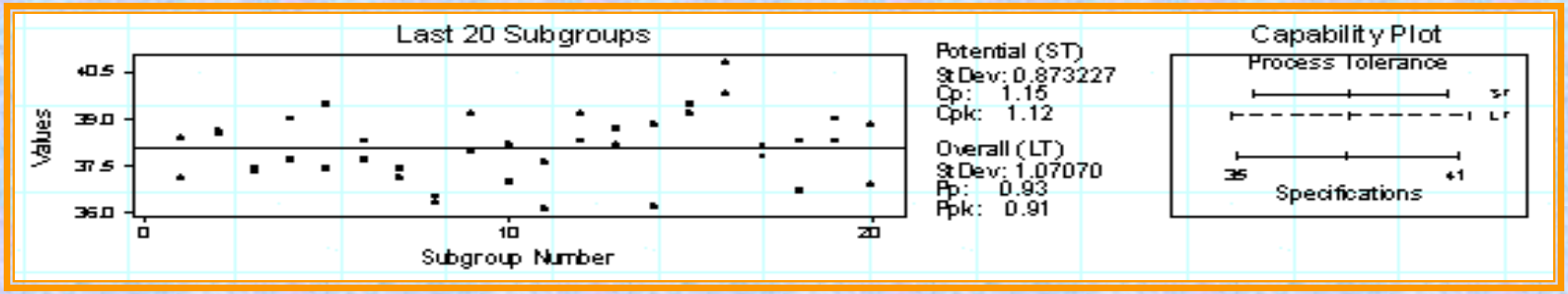
Aplikace:

- **výběry relativně malých skupin sledovaných výrobků ke kontrole během výroby ve stanovených časových intervalech**
- **grafické zpracování číselných charakteristik pro informace o negativních vlivech na kvalitu výroby**
- **při vybočení sledované charakteristiky mimo statisticky určené meze zásah nebo vyhodnocení pro přípravu další výroby**

Důsledek : finanční efekt pro výrobce a zvýšení důvěry odběratele

Nyní: SPC součástí norem pro QC na státní a nadstátní úrovni i obsahem příruček jakosti u jednotlivých firem (certifikace)

Další metody: způsobilost výrobních procesů, statistické přejímky, optimalizace nákladů aj.



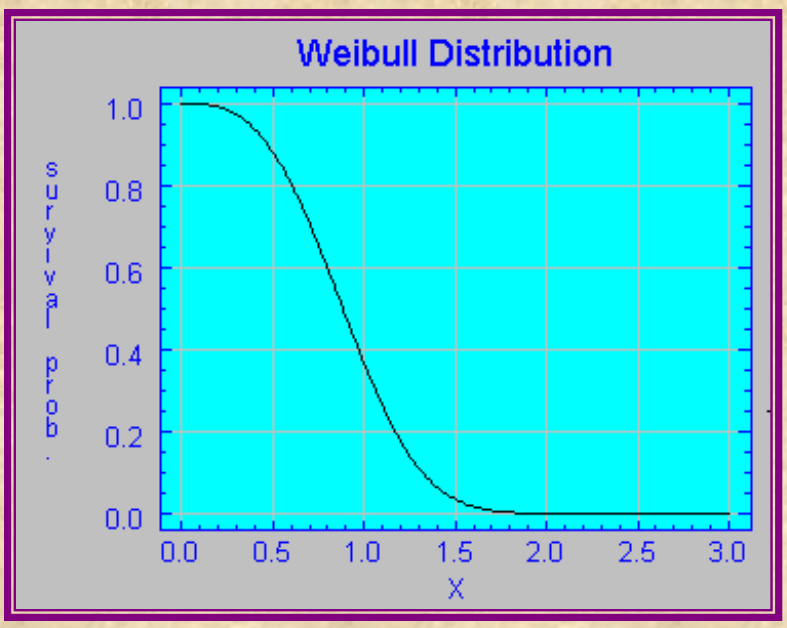
Spolehlivost výrobku se dá měřit a úspěšně využívat

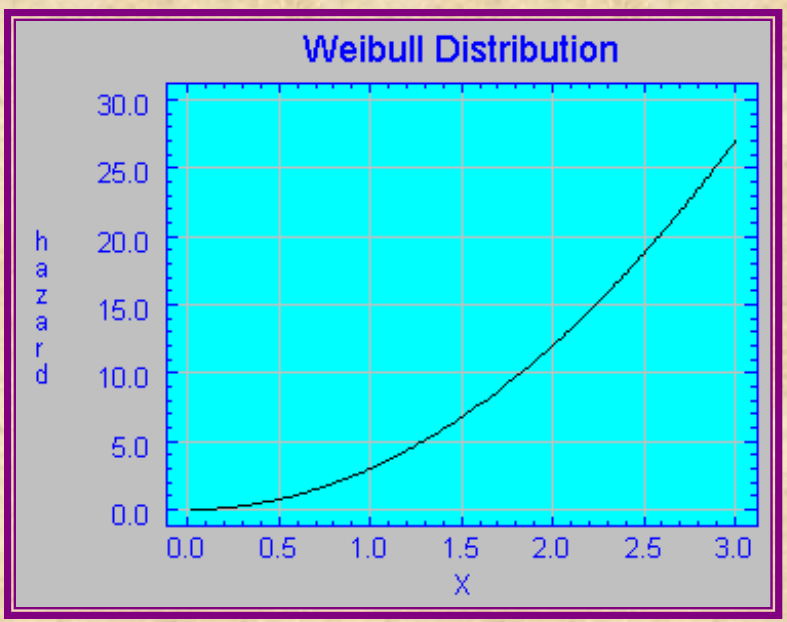
Spolehlivost výrobku = jeho schopnost plnit požadované činnosti

Statistický princip: doba bezporuchového stavu, doba opravy apod. sledovaného objektu jsou náhodné veličiny

Charakteristiky spolehlivosti:

- **funkční: funkce spolehlivosti, intenzita poruch, hustota obnov aj.**
- **číselné: střední doba do poruchy, koeficient pohotovosti aj.**



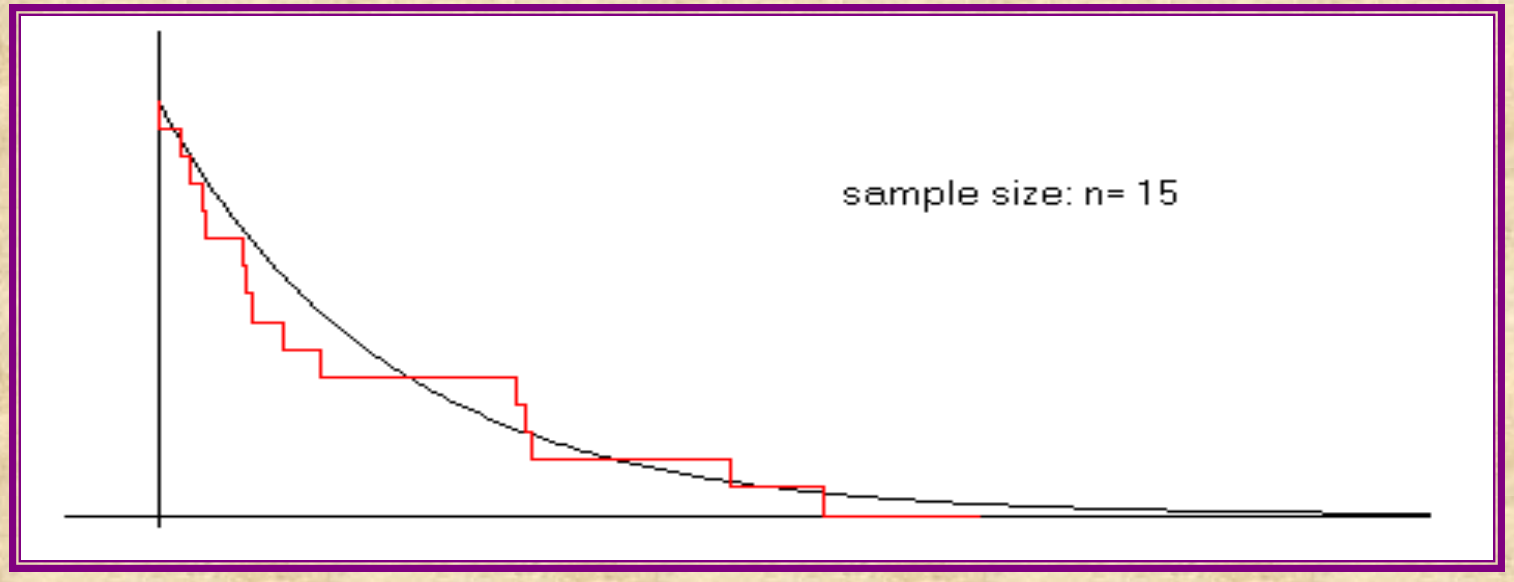


Spolehlivost výrobku se dá měřit a úspěšně využívat

Postupy:

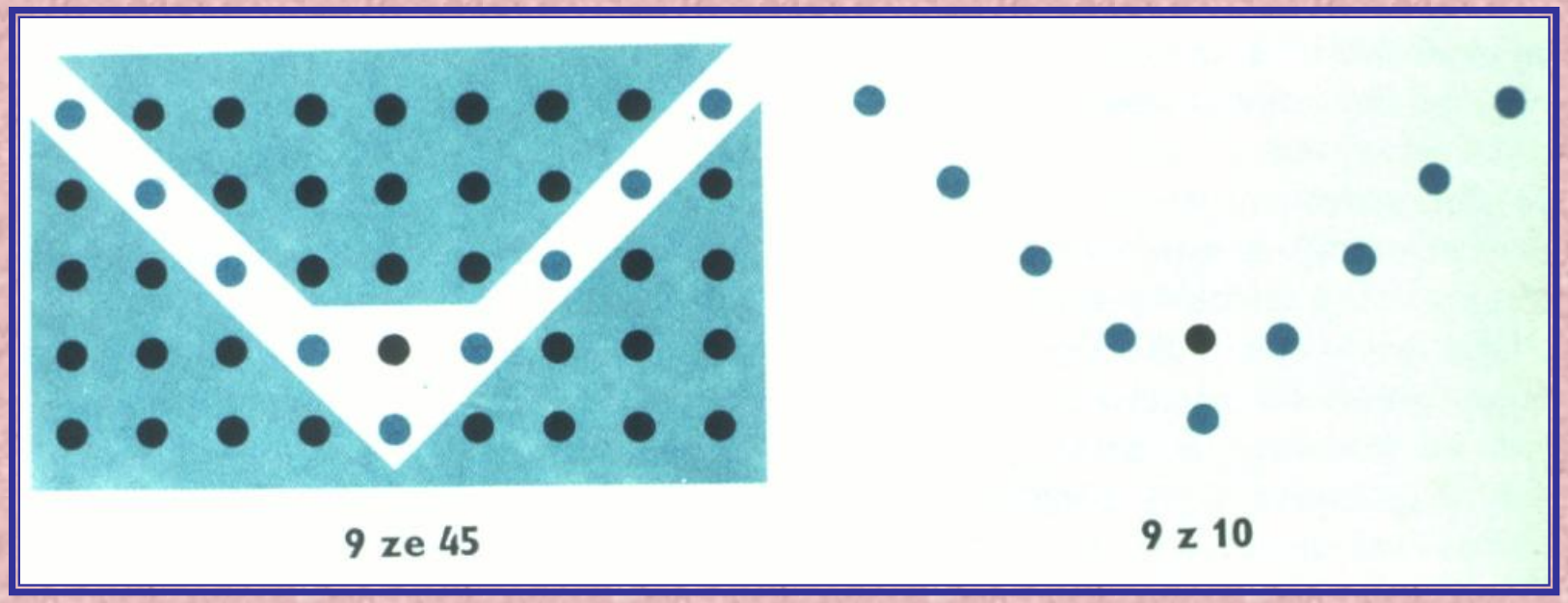
- **odhady charakteristik ze statistických souborů často speciálně orientovanými metodami podle realizace zkoušek (např.cenzorované výběry)**
- **plány údržby**
- **pravděpodobnostní modely systémů - celků (např. výrobní linky, energetické bloky)**
- **optimalizace spolehlivosti vzhledem k nákladům**

Efekty spolehlivosti: užitná hodnota, bezpečnosti, cena

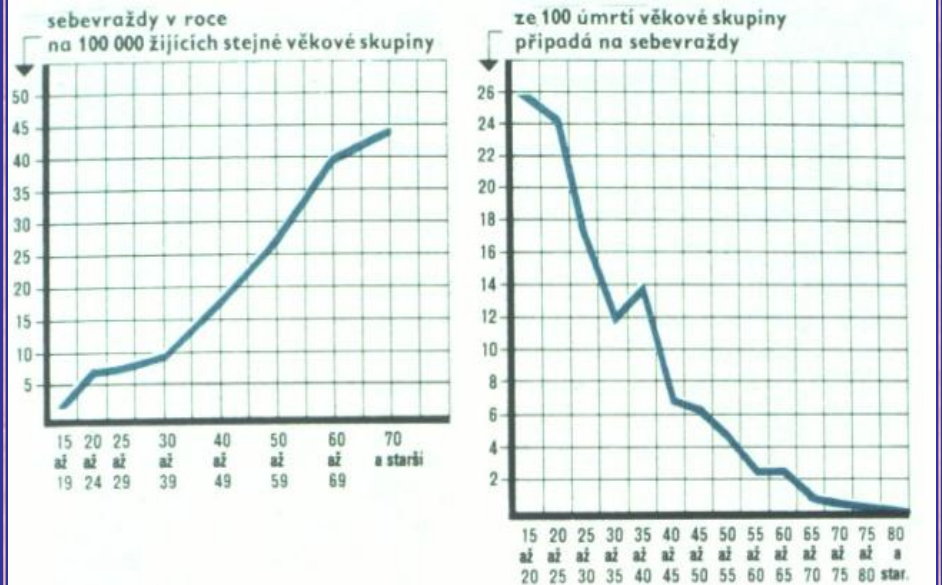


Proč a jak někdy statistiky lžou, a co dál…

No to snad ne…!?



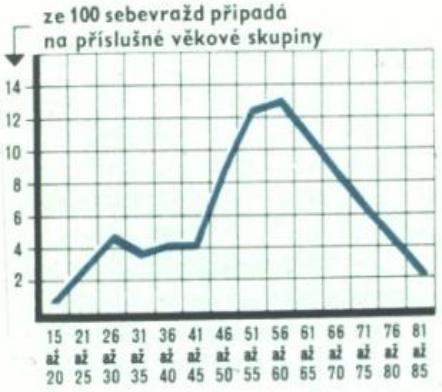
Proč a jak někdy statistiky lžou, a co dál…

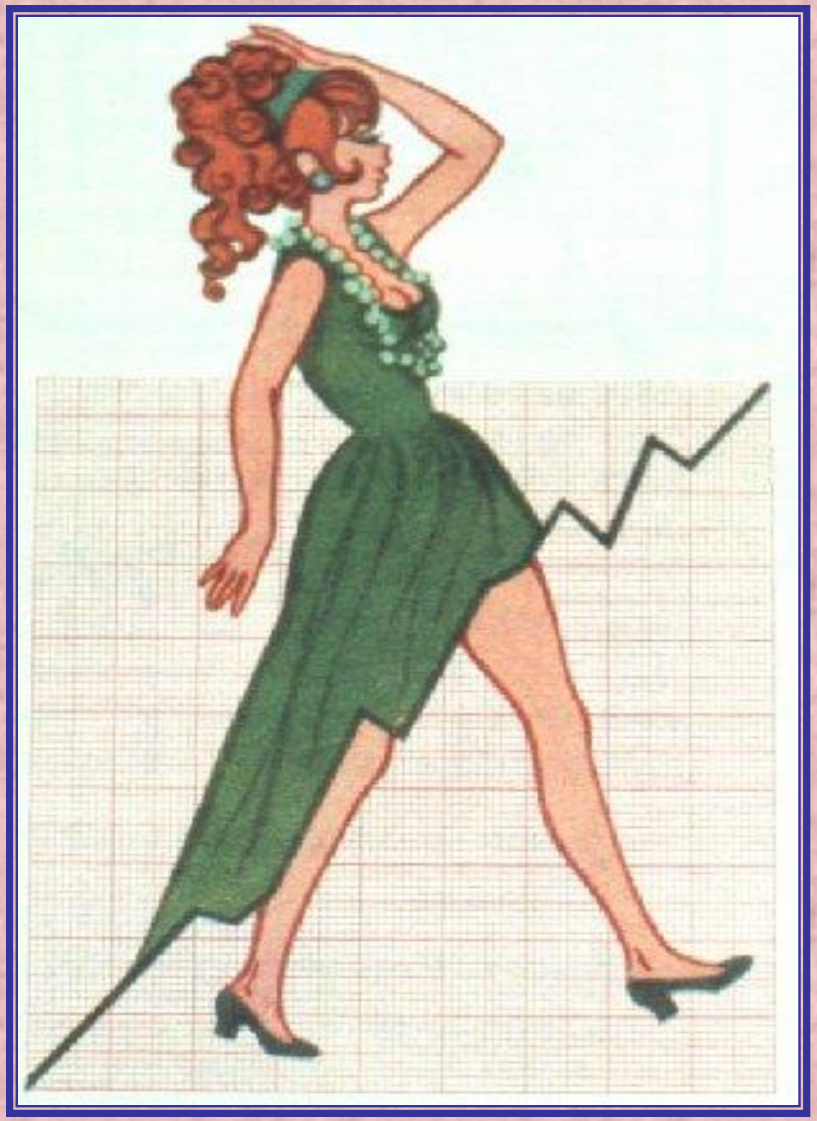


Tak se „lže" statistikou.

První „statisticky dokázané zjištění": sebevražednost se zvyšuje s rostoucím věkem. Průběh křivky grafu zřetelně ukazuje, že počet sebevrahů v roce a dané věkové skupině se plynule zvyšuje a ve věkové skupině „70 a starší" činí více než desetinásobek skupiny mladistvých.

Druhé „statisticky dokázané zjištění": ve středním věku dosahuje sebevražednost vrcholu, vzácné jsou naproti tomu mezi mladými a nejstaršími. Průběh křivky grafu zřetelně ukazuje, že z každých 100 sebevražd připadá asi $^1/_4$ na věkovou skupinu 51 až 60, avšak na sedmdesátileté (71—80) jen asi 10 % a na skupinu 21 až 30 let pouze 8 %.





Proč a jak někdy statistiky lžou, a co dál…

Říká se, že statistiky někdy lžou. Pojmy, metody a postupy matematické statistiky jsou součástí matematiky, tedy produktem našeho abstraktního myšlení. Jejich aplikace (tzv. statistiky) jsou výsledkem konkrétní činnosti lidí, která může být seriozní anebo naopak neseriozní. Ve druhém případě pak může jít buď o nevědomost anebo záměr. Tak či onak za to samy statistiky nemohou - lhát může jen člověk. Statistiky proto nelžou, avšak horší to bývá s jejich realizátory a vykladači, kteří z nich někdy polopravdy, nepravdy, případně až lži vytvářejí. Racionální a zodpovědné používání statistických metod naopak přináší pozitivní výsledky tím, že rozšiřuje naše poznání okolního světa i sebe sama a umožňuje nám účelně rozhodovat o našem dalším konání. A v tom lze vidět i jejich možný přínos v budoucnosti.