

## Statistické metody a zpracování dat

### I. Úvod, základní pojmy

Petr Dobrovolný



### Obsah přednášky

1. Úvod, základní pojmy
2. Základní vyjadřovací prostředky ve statistice
3. Základní popisné statistické charakteristiky
4. Úvod do počtu pravděpodobnosti, teoretická rozdělení
5. Odhady parametrů a intervaly spolehlivosti
6. Testování statistických hypotéz
7. Měření závislosti náhodných veličin
8. Analýza kategoriálních dat
9. Úvod do analýzy rozptylu
10. Úvod do analýzy časových řad
11. Úvod do vícerozměrných statistických metod I, Faktorová analýza
12. Úvod do vícerozměrných statistických metod II, Shluková analýza

### Základní literatura

**Brázdil a kol.** (1995): Statistické metody v geografii. MU Brno, 177 s.

Prezentace z přednášek – doplňky

**Hendl, J.** (2004): Přehled statistických metod zpracování dat. Portál, Praha, 583 s.

**Rogerson, P. A.** (2001): *Statistical methods for Geography*. Sage Publications, London, 236 s.

**Heřmanová, E.** (1991): Vybrané vícerozměrné statistické metody v geografii. SPN, Praha, 133 s.

Cvičení – zadání, podkladová data - přes IS

Termíny písemných testů: 9.11. 21.12.

### STATISTIKA - definice

**Statistika je vědní obor zabývající se zkoumáním jevů, které mají hromadný charakter.**

*Statistika* je v určitém smyslu jazykem pro **shromažďování, zpracování, rozbor, hodnocení a interpretaci** hromadných jevů

#### Co je typické pro statistiku

- Zkoumá **hromadné** jevy.
- Zabývá se proměnlivými - **variabilními** - vlastnostmi.
- Pracuje s čísly a vyjadřuje se pomocí čísel - zajímá se především o **kvantitativní stránku** reality.
- Používá výpočetní techniku k vytváření a správě statistických **databází**, k provádění hromadného **zpracování a analýzy** dat a ke **komunikaci**.

### Významy pojmu STATISTIKA

#### I. Statistika jako praktická činnost

- Statistická evidence (např. sběr údajů, třídění, sumarizace apod.),
- Instrukce, která tuto evidenci provádí (např. ČSÚ, ministerstva aj.)
- Souhrn údajů o nějaké skutečnosti (statistika nezaměstnanosti, ročenka meteorologických pozorování atd.)

### Významy pojmu STATISTIKA

#### II. Statistika jako vědní disciplína

- **Popisná** statistika (výsledky nelze zobecnit)
- **Matematická** (induktivní) statistika - cílem je zobecnit výsledky (odhad a testování hypotéz) - použití **počtu pravděpodobnosti**
- Teorie výběrových zjišťování či měření
- Aplikované vědy („-metrie“ a „-grafie“): biometrie, dendrometrie, ekonometrie, chemometrie atd.
- Vědy se silným statistickým základem: klimatologie, hydrologie, sociologie, psychologie, demografie aj.

### Co statistika „umí“

- **Zjišťování** (počet domácností ČR, počet pracovníků v odvětví XY)
- **Shrnování** dílčích ukazatelů v čase a prostoru (průměrná nezaměstnanost v regionu)
- **Srovnávání** agregovaných ukazatelů v čase nebo prostoru (trend vývoje počtu obyvatelstva, teploty vzduchu dvou lokalit)
- Měření **závislosti** (závislost mezd na HDP, závislost meteorologického prvku na nadmořské výšce).
- Popis **struktury** (věková struktura obyvatel ČR, roční chod hodnot meteorologických prvků)
- **Předvídání** jejich budoucí úrovně (tržby v maloobchodě v příštím roce)

### ... a co statistika „neumí“:

Statistika selhává, pokud:

- Nemá k dispozici adekvátní číselné údaje
- Chybí-li představa o velikosti chyb měření a vlivu různých doprovodných činitelů
- Nemá-li k dispozici dostatečně rozsáhlý soubor případů
- Není-li v datech přítomna proměnlivost (variabilita).

### Vymezení základních pojmů I

**Hromadné jevy:** přírodní či společenské jevy, které jsou výsledkem působení velkého množství příčin, jejich vlastnosti se neprojevují v jednotlivých jevech, ale jen v souboru a to prostřednictvím řady náhod.

Hromadné jevy pozorujeme:

- jednoduché pozorování
- experiment

Některé jevy, které v geografii studujeme pomocí statistických metod mají povahu jevů náhodných – tzv. stochastických (hydrologické jevy či meteorologické jevy).

### Vymezení základních pojmů II

**Statistická jednotka:** je to určitý jev či prvek, který je předmětem statistického šetření a pro který se zjišťují údaje

Statistická jednotka musí být přesně vymezena na počátku vlastního šetření a to z hlediska **věcného, časového, prostorového**.

**Statistický znak:** je to určitá vlastnost statistické jednotky, kterou se snažíme postihnout. Tzv. **shodné (společné) znaky** vymezují příslušnost statistické jednotky k určitému statistickému souboru. Ostatní jsou znaky **proměnlivé (variabilní)**.

### Vymezení základních pojmů III

Statistické znaky lze dělit na znaky **prostorové, časové a věcné**.

Věcné znaky se dělí na znaky **kvantitativní a kvalitativní**

Kvalitativní znaky mohou být **alternativní a množné**

Kvantitativní znaky dělíme nejčastěji na znaky **spojité a diskrétní**.

Statistické znaky můžeme získat přímo – (např. **měřením**) a nebo **nepřím** (výpočtem). Tyto potom nazýváme znaky odvozenými.

Podle škály, na které znaky zjišťujeme je dělíme na znaky **nominální, ordinální, poměrové, intervalové**

### Vymezení základních pojmů IV

**Statistický soubor:** skupina statistických jednotek stejného druhu (věcné, prostorové a časově vymezených)

Je to množina všech prvků, které jsou předmětem daného statistického zkoumání. Každý z prvků je statistickou jednotkou.

Prvky tvořící statistický soubor mají určité společné vlastnosti - tzv. **identifikační znaky** - umožňující určit, zda prvek do daného statistického souboru patří nebo nepatří. Identifikační znaky tedy statistický soubor vymezují.

Z hlediska cílů statistického zkoumání sledujeme na prvcích statistického souboru jednu nebo více vlastností - **sledované znaky**. Je-li vlastnost měřitelná v nějakých jednotkách, jde o kvantitativní znak, jinak jde o kvalitativní znak.

## Vymezení základních pojmů IV

Statistický soubor můžeme podle různých hledisek dále dělit:

- Statistický soubor **jednorozměrný, vícerozměrný**
- Statistický soubor **základní a výběrový**

**Základní soubor** - populace

**Výběrový soubor** je podmnožinou základního souboru. Je vytvořen ze statistických jednotek, vybraných podle určitého hlediska.

**Reprezentativní výběr:** Pokud zkoumaný výběr dobře odráží strukturu celého zkoumaného souboru, nazýváme jej reprezentativním výběrem.

**Rozsah statistického souboru:**

- N – rozsah základního souboru
- n – rozsah výběrového souboru

## Popisná statistika

*Popisná (deskriptivní) statistika* se zabývá uspořádáním souborů, jejich popisem a účelnou sumarizací.

Jak mohou být tyto jevy jednoduše popsány (charakterizovány, sumarizovány)?

Existují dvě základní možnosti, které se vzájemně doplňují:

- **Numerické metody** – jedním nebo několika málo čísly lze vystihnout určité vlastnosti jevu. Jsou přesnější a objektivnější
- **Grafické metody** – sestavení vhodného typu grafu. Jsou názornější a umožňují vystihnout vztahy.

## Induktivní statistika

*Induktivní (matematická) statistika* se vyvinula z popisné statistiky a jejím základem je **teorie pravděpodobnosti**.

Matematická statistika zkoumá soubory nepřímo prostřednictvím výběrů

*Induktivní statistika* se zabývá metodami jak poznatky **přenášet** a umožňuje z pozorovaných dat vytvářet **obecné závěry** s udáním *stupně jejich spolehlivosti*.

Výpočet stupně spolehlivosti závěrů je však objektivní, neboť je založen na poznacích teorie pravděpodobnosti a nezávisí na subjektivním názoru hodnotitele.

## Základní etapy statistického zpracování dat

- **Zjišťování** - shromáždění a zaznamenání údajů, jejich kontrola aj.,
- **Zpracování** - uspořádání, seskupení, shrnování, sumarizace,
- **Analýza** - výpočet charakteristik, měření závislosti, srovnávání, měření dynamiky
- **Prezentace** výsledků - tabulkové či grafické vyjádření a slovní zhodnocení výsledků předcházejících etap.

## Základní dělení statistických údajů

- podle zdroje — **primární** a **sekundární**,
- podle reálnosti situace — **skutečné** a **simulované**,
- podle periodicity zjišťování — **průběžné**, **periodické** a **jednorázové**,
- podle časového hlediska — **okamžikové** a **intervalové**.
- podle použité škály měření – **nominální**, **ordinální**, **intervalové**, **poměrové**

## Typy geografických dat

- **Nominální** (kategorie využití země)
- **Ordinální** (řád vodního toku, stupnice síly větru)
- **Intervalová** (teplota vzduchu) nula = data
- **Poměrová** (množství srážek, délka vodního toku) nula = neexistence jevu

## Typy geografických dat

**Nominální data** – hodnota představuje konkrétní kategorii či třídu a vyjadřuje její označení (jméno), kategorie se nesmějí překrývat – jsou disjunktní. Každý objekt je zařaditelný alespoň do jedné kategorie, žádný nespadá do více jak jedné. Čísla, která označují kategorie jsou pouze symboly a nelze s nimi provádět aritmetické operace. V nejjednodušší podobě mají binární charakter a lze je pouze porovnávat.

**Ordinální data** – data, která lze seřadit do uspořádané posloupnosti podle určitého kritéria. Je známé pořadí kategorií, rozdíl však nemá smysl. Např. řád vodního toku, třída silnice, bonita půdy atd.

## Typy geografických dat

**Intervalová data** – umožňují provádět i odečítání mezi kategoriemi definovat rozdíl mezi kategoriemi. Teplota vzduchu. Stupnice většinou nezačíná nulou. Poměr dat závisí na zvolených jednotkách

**Poměrová data** – vedle rovnosti, uspořádání a odčítání umožňují také dělení. Nula vyjadřuje neexistenci jevu – objem, délka ...

## Konverze mezi jednotlivými typy dat

Čerpací stanice	Vzdálenost od středu místa	Čerpací stanice	Vzdálenosti utříděné vzestupně	Pořadí v utříděné posloupnosti	Příslušnost stanice do třídy
A	112,7	D	15,8	1	1
B	40,6	L	32,7	2	1
C	678,3	B	40,6	3	2
D	15,8	G	67,7	4	1
E	112,7	O	98,4	5	1
F	554,9	A	112,7	6	2
G	67,7	E	112,7	7	1
H	889,5	K	112,7	8	2
I	1006,5	J	445,1	9	2
J	445,1	F	554,9	10	1
K	112,7	M	654,5	11	1
L	32,7	C	678,3	12	1
M	654,5	H	889,5	13	2
N	1322,7	I	1006,5	14	2
O	98,4	N	1322,7	15	1

Třída příslušnosti: 1 – stanice blízka; 2 – stanice vzdálená  
kritérium: hodnota vzdálenosti 500 m  
Vzdálenost – poměrová data  
Pořadové číslo – ordinální data  
Třída – nominální data

## Statistika a výpočetní technika

- Výpočetní technika zasahuje do všech etap statistického zpracování dat.
- Exploze výpočetní techniky umožňuje provádět výpočty, které byly dříve nerealizovatelné (z důvodů velkého objemu dat, pracnosti, ...).
- Na druhou stranu však roste nebezpečí výběru nesprávného postupu.

## Výhody počítačového zpracování I.

**Přesnost a rychlost:** Dobré počítačové programy nám dají velmi rychle správné výsledky. Dřívější ruční zpracování dat bylo často zatíženo aritmetickými chybami a bylo časově velmi náročné.

**Univerzálnost:** Počítače zpřístupňují širokou škálu statistických metod a umožňují provést velmi rychle i rozsáhlé komplexní statistické analýzy.

**Grafika:** Počítače umožňují snadné grafické zobrazení pozorovaných dat a výsledků statistického zpracování.

**Flexibilita:** Velkou výhodou počítačů je, že umožňují rychle provést nové zpracování při změnách v datech či transformaci některých veličin.

## Výhody počítačového zpracování II.

**Nové veličiny:** Snadno lze vytvářet nové veličiny pomocí požadovaných transformací.

**Velikost datových souborů:** Počítače umožňují zpracování velmi rozsáhlých souborů dat pomocí vhodného softwaru, což bylo ještě před deseti lety velmi obtížné.

**Snadný přenos dat:** Jakmile se jednou data dostala do počítače, lze je snadno přenést elektronicky (například pomocí Internetu) na jiné místo.

...ale

## Nevýhody počítačového zpracování I.

### Kvalita, dostupnost, spolehlivost softwaru.

Ne všechny statistické programy jsou spolehlivé. Řada SW programů aplikaci statistických metod zjednodušuje

Je vhodné využívat ověřené postupy a programy - BMDP, SAS, SPSS, STATISTICA, S PLUS, STATGRAPHICS a další.

### Univerzálnost.

Může vést k výběru nevhodné metody zpracování. Je velmi důležité, aby každý, kdo používá statistický software, si byl vědom úrovně svých statistických znalostí a užíval pouze ty metody, kterým rozumí. Pozor na používání neznámých statistických metod.

## Nevýhody počítačového zpracování II

### Černá skříňka.

Počítač vzdaluje uživatele od dat i metody zpracování. Statistická analýza se provádí automaticky, nová data se zpracovávají a výsledky se ukládají, aniž by byly posouzeny člověkem. Protože většinou výsledky zachycují jen průměrné efekty, může se zcela ztratit citlivost k individuálním pozorováním.

### Špatná data plodí špatné závěry.

Jestliže data jsou nasbírána či naměřena špatně (například jsou špatně kladené otázky v dotazníku), nelze očekávat, že závěry z takových dat budou správné. Sem náleží i nesprávné zpracování datových souborů, chybějící či ovlivněné (tzv. nehomogenní) údaje.

## Statistický software

1. Programové vybavení založené na využití vlastního programovacího jazyka (R, Splus, SAS)
2. Interaktivní zpracování v „oknech“ MINITAB, SPSS, STATGRAPHICS, Statistica
3. Programové vybavení s knihovnou statistických, matematických a grafických funkcí (EXCEL)

## Statistické metody a zpracování dat

### II. Vyjadřovací prostředky ve statistice

Petr Dobrovolný



## Základní vyjadřovací prostředky ve statistice

- Statistické tabulky
- Statistické grafy

Tabulky – složené z buněk, přehledné, nezávislé na textu

Tab. 1 Základní statistické charakteristiky teploty vzduchu [°C] na vybraných stanicích za období 1961-2000

Charakteristika	Stanice	
	Praha-Klementinum	Strání
průměr [°C]	9,205	7,673
rozptyl <sup>1)</sup>	0,679	0,403
rozsah souborů	120	30
směrodatná odchylka [°C]	0,821	0,624
F	1,687	-
F (kritické)	1,699	-

Pramen: CHMÚ  
Vysvětlivky: 1) ....

- nadpis
- záhlaví
- legenda
- pramen
- poznámky
- vysvětlivky

## Statistické tabulky

- Záhlaví a legenda mají obsahovat měrné jednotky
- Tabulka má vyplněna všechna políčka
- Smluvené znaky pro políčka bez číselného údaje
  - – údaj se nevyskytuje
  - x – údaj není možný z logických důvodů
  - 0 – hodnota je menší než polovina nejmenší měrné jednotky
  - . – údaj nelze spolehlivě zjistit

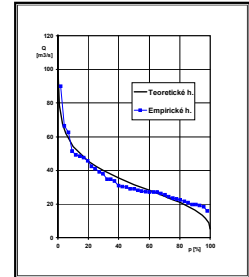
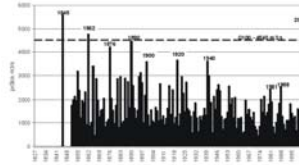
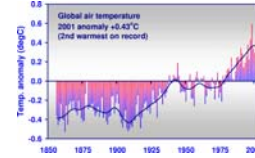
## Druhy statistických tabulek

- Podle účelu – pracovní, koncentrační, publikační
- Podle obsahu – jednoduché, kombinační
- Korelační, asociační, kontingenční

## Metody grafického znázornění geografických jevů

Motto:

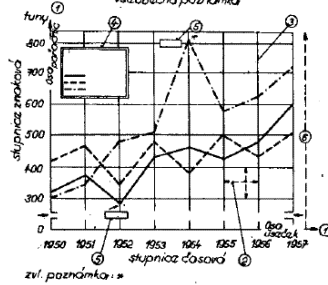
Jeden obrázek je za tisíce slov



Cílem grafického znázornění je podat rychlou a srozumitelnou informaci o studovaném jevu či o vzájemném vztahu více jevů.

## Graf a jeho základní prvky

Název grafu  
Podtitul  
Všeobecná poznámka

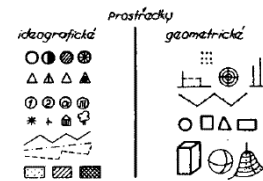


- stupnice
- grafický interval
- sít
- klíč
- vysvětlivka
- délka stupnice

**Graf** – kresba provedená podle předem dohodnutých pravidel, která znázorňuje kvalitativní či kvantitativní znaky.

**Grafický obraz** – soubor grafických prostředků, pomocí kterých na základě dohodnutého výkladu jejich smlouveného významu sestrojujeme i čteme graf

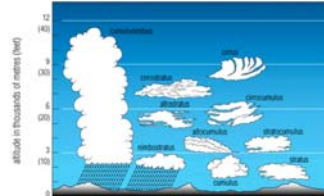
## Dělení grafických prostředků podle významu



**Ideografické** – mají kvalitativní význam a v grafu fungují jako znaky (klasifikační, identifikační). Jejich tvar a rozměry slouží pouze k jejich odlišení, nemají kvantitativní význam (písmena, číslice, symboly, geometrické obrazce, šrafura, barva, druhy čar apod.).

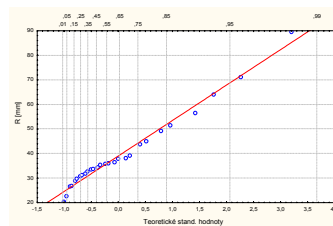
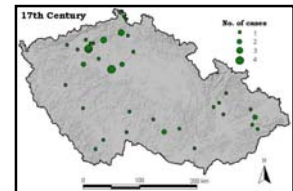
**Geometrické** – mají vždy kvantitativní význam, často však také slouží ke kvalitativnímu odlišení statistických jednotek (body, úsečky, obrazce).

## Ideografické prostředky



- windbreakage
- damage on buildings of lesser extent
- destroyed buildings
- damage without specification

## Geometrické prostředky



## Grafický výklad

Soubor zásad, podle kterých interpretujeme (čteme) příslušný graf).

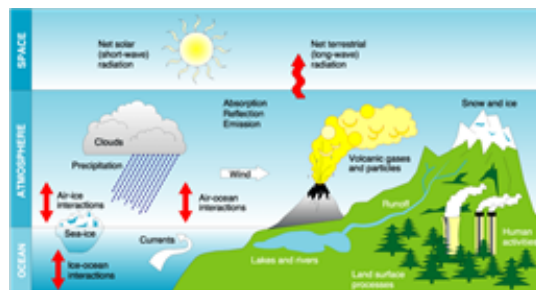
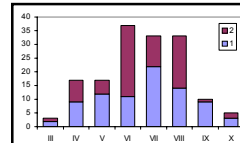
Základní prvky grafického výkladu:

- Klíč
- Stupnice
- Souřadná soustava
- Zonální soustava
- Název a podtitul
- Poznámky a vysvětlivky

## Základní typy grafů

Z hlediska předmětu grafu:

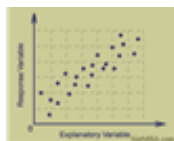
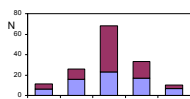
- schémata
- diagramy



## Základní typy grafů

Z hlediska způsobu použití geometrických prostředků:

- rozměrové grafy
- souřadnicové grafy



Speciální typy grafů využívané v geografii:

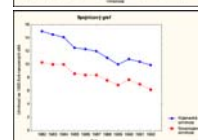
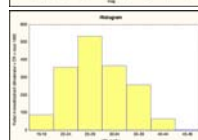
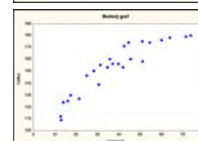
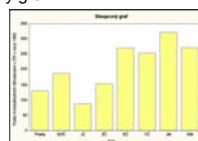
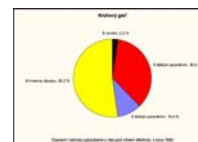
- ternární graf
- větrná růžice, klimadiagram, ...

- Statistické mapy
- kartogramy
- kartodiagramy

## Základní typy grafů

Grafy pro vyjádření jedné proměnné

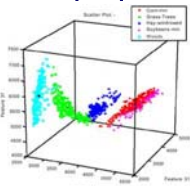
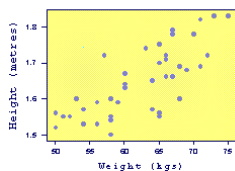
- sloupcový diagram
- histogram
- kruhový diagram, výsečový graf
- bodový graf
- spojnicový graf



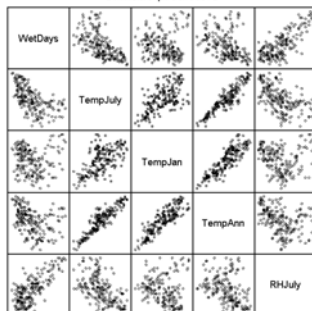
## Základní typy grafů

Grafy pro vyjádření dvou a více proměnných - korelogram

Scatterplot

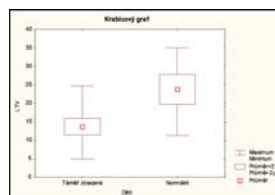


Climatic predictors



## Speciální typy grafů

- krabíčkový graf
- graf stonku a listů (stem-and-leaf-plot)
- piktogram



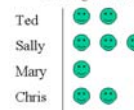
Stem-and-leaf plot of 'Skinfold Thickness'

N = 40  
Leaf Unit = 0.10

```

5 02468
6 246898
7 0000226
8 2444666
9 44666
10 488
11 088
12
13 22
14 0
15
16
17
18 2
    
```

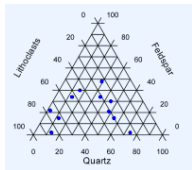
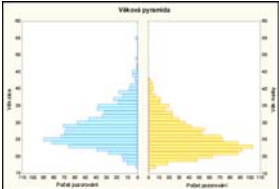
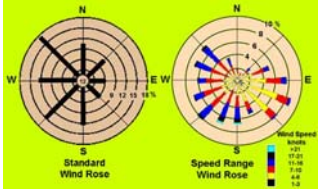
Good grades on spelling test



KEY: ● Represents a month of 80%+ scores

**Speciální typy grafů využívané v geografii:**

- ternární graf
- „strom života“
- větrná růžice
- klimadiagram

**Analýza grafů**

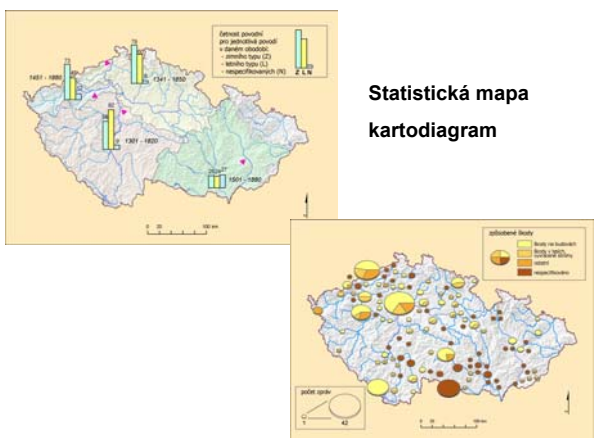
Všimáme si základního tvaru a také odchylek od něho

U tvaru grafu hodnotíme:

- zhuštění – místa největší četnosti hodnot
- shluky – existence jednoho či více shluků hodnot
- mezery – existence intervalů či oblastí bez hodnot
- odlehle hodnoty – existence údajů podstatně rozdílných od ostatních hodnot
- extrémní hodnoty – poloha min a max hodnot v grafu
- tvar rozdělení – jak ho lze popsat – symetrie, počet vrcholů

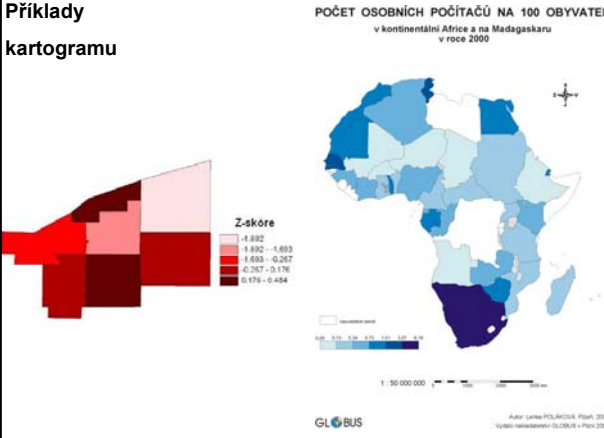
Volba vhodného typu grafu musí zohledňovat typ zobrazované proměnné (spojitá či diskrétní)

**Statistická mapa kartodiagram**



**Příklady kartogramu**

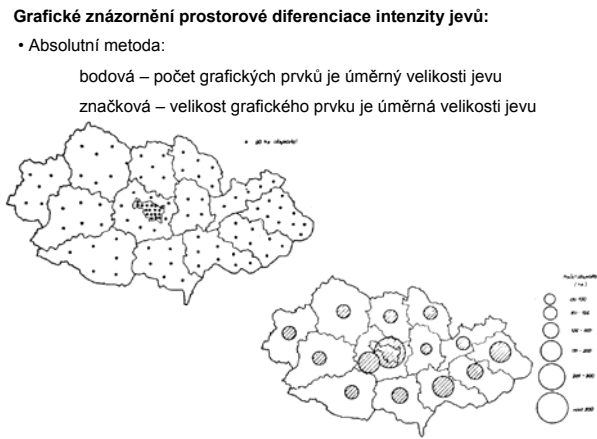
**POČET OSOBNÍCH POČÍTAČŮ NA 100 OBYVATEL**  
v kontinentální Africe a na Madagaskaru v roce 2000



**Grafické znázornění prostorové diferenciace intenzity jevů:**

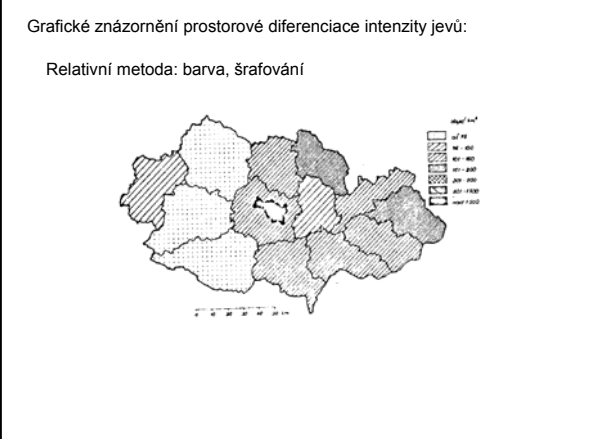
• Absolutní metoda:

- bodová – počet grafických prvků je úměrný velikosti jevu
- značková – velikost grafického prvku je úměrná velikosti jevu



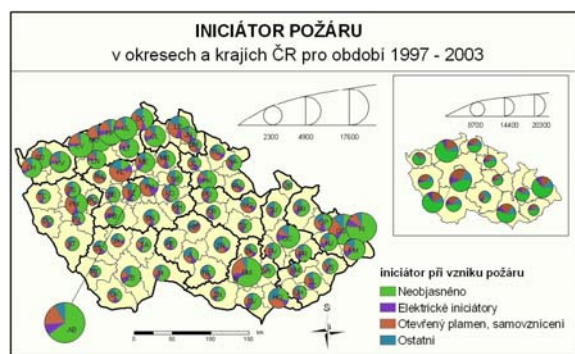
Grafické znázornění prostorové diferenciace intenzity jevů:

Relativní metoda: barva, šrafování





Grafické znázornění prostorové diferenciace intenzity a struktury jevů



**Interpolace:**

Konstrukce spojitých polí a izolinií  
(viz. Geostatistika)

