SEMINÁŘ VÝPOČETNÍ STATISTIKY P1

2008-02-18

- ✓ Samostatná práce, navazuje na předchozí statistické předměty.
- ✓ Zkouška – písemná, část praktická

SAS:

SAS – Statistical Analysis System

Moduly – nabídkově řízené režimy:
- ✓ SAS Insight – náhledy, umožňuje prohlédnutí zpracovaných materiálů (datových souborů)
- ✓ SASRAB – v systému pod tímto názvem není, v SASu je jako Guided Data Analysis
- ✓ SAS Analyst
- ✓ SAS Assisst

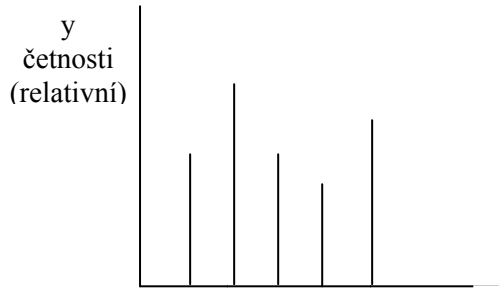
Programovací postup – „program“:
- ✓ Speciální programovací jazyk systému SAS.
- ✓ Moduly:
  - Base SAS
  - SAS STAT
  - Syntaxe je stejná, není nutné rozlišovat.

SEMMA:
- ✓ Technologie SASu, využívaná u dataminingových postupů
- ✓ Doporučení, jak by měla statistická analýza vypadat
- ✓ **S – Sample** – výběr, určení velikosti výběrového souboru.
- ✓ **E – Explore** – prozkoumat, průzkumová (explorační) analýza dat, základní seznámení s datovým souborem, zjištění zvláštností, které mohou výpočet zkreslit.
- ✓ **M – Modify** – uprav, cílem je upravit data, o kterých v Explore zjistíme, že nejsou zcela v pořádku, např. vyřazením.
- ✓ **M – Model** – vlastní analytická část, sestavení statistického modelu, pomocí kterého se pokusíme data analyzovat (regresní a korelační analýza, analýza časových řad, testovací postupy, zpracování kontingenčních tabulek atd.).
- ✓ **A – Assess** – vyhodnoť, oceň, hodnocení, posuzování a zkoumání výsledků, zda souhlasí s daty a postupem, jsou-li korektní a v souladu se zpracovávanými daty, lze je považovat za přijatelné. Pokud korektní nejsou, je celý proces opakován.

Zápis procedur v SASu:

`proc data=jmeno;`	Klíčové slovo, data, jméno, zakončit středníkem.
`var ...;`	Slovo var a seznam proměnných, které chceme zpracovávat.
`run;`	Spuštění.

Zobrazování datového souboru:
- ✓ Graf – **sloupcový diagram**, v SASu Bar Chart, umožňuje vizualizovat datový soubor, připomíná histogram
  - **Typy diagramu:**
    - Vertikální graf:



  - Horizontální graf
- **Procedura pro vyvolání sloupcového diagramu:**

```
proc chart data=jmeno;
vbar vyska;
run;
```

Vertikální diagram, pokud chceme horizontální, tak `hbar`

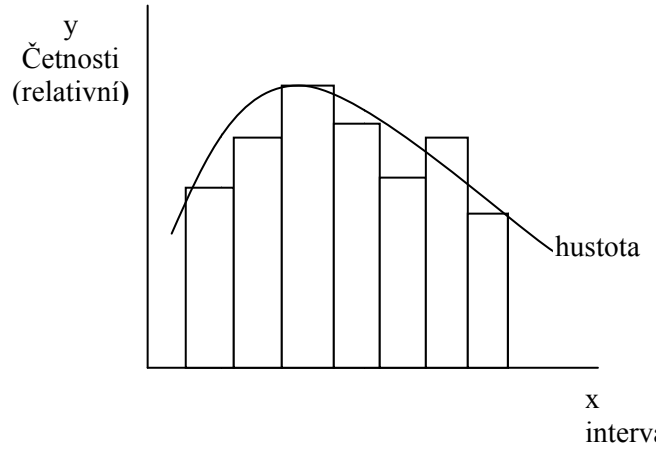
- **Zdařilejší grafické výstupy** – lze je dále graficky upravovat:

```
proc gchart data=jmeno;
vbar vyska;
run;
```

- **Sturgesovo pravidlo** – vypočítání doporučeného počtu intervalů: $k_1 = 1 + 3{,}3 \log n$
- **Možnosti**:
  - Lze zvolit, kolik intervalů má být
  - Jak mají vypadat středy intervalů (nemusí být stejně vzdálené)
  - Každá jednotlivá hodnota může být uvedena a mít vlastní úsečku
  - Doplňkové volby se dělají v druhém řádku procedury, píší se za lomítko

✓ **Histogram:**

- **Graf:**



- **Procedura univariate** – pro grafické i numerické výstupy:
  - Pouze grafické výstupy:

```
proc univariate data= jmeno noprint;
histogram <jmeno_promenne>;
run;
```

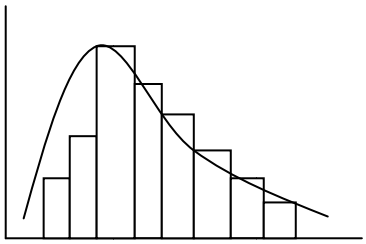
noprint – netisknou se numerické výstupy

  - Chceme-li také Gaussovu křivku:

```
proc univariate data= jmeno noprint;
histogram <jmeno_promenne> / normal;
run;
```

normal – normální rozdělení, exponential – exponenciální

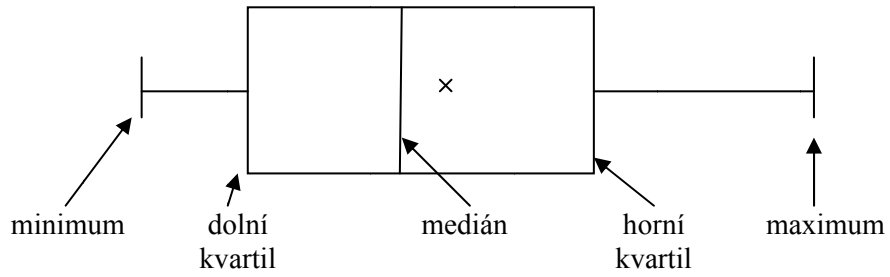
- **Z histogramu lze odvodit:**
  - Středovou, centrální, nejdůležitější hodnotu datového souboru
  - Variabilita, rozptýlenost hodnot okolo středové hodnoty
  - Orientační odhad, zda údaje jsou soustředěny okolo průměru symetricky nebo asymetricky
    - Pozitivní asymetrie, zešikmení doprava:



- Negativní asymetrie, zešikmení doleva – obráceně

✓ **BoxPlot**, Box-and-whisker plot:

- Grafické zobrazení takzvaného **pětičíselného souhrnu**, tedy five-number summary – minimum, dolní kvartil, medián, horní kvartil, maximum: x_{min}, $\tilde{x}_{0,25}$, $\tilde{x}$, $\tilde{x}_{0,75}$, x_{max}. Dolní kvartil, medián a horní kvartil rozdělují soubor seřazený podle velikosti na čtyři stejně početné části. Jako hvězdička či křížek se zobrazuje také průměr
- **Zobrazení:**



- **Kvartilové rozpětí:** IQR = $x_{0,75} - x_{0,25}$
  Zobrazuje variabilitu souboru
- **Skeletal** – úsečky jsou protaženy k minimu a maximu
- **Schematic** – zobrazuje také odlehlá pozorování, tzv. outliers, mohou být také chybné