



SEMINÁŘ VÝPOČETNÍ STATISTIKY

P4

2008-03-10

VÍCEROZMĚRNÉ STATISTICKÉ SOUBORY:

SASUser.Fitness:

- ✓ **Procedura univariate – průzkumová analýza rozdělení četnosti** proměnné OXYGEN uložené v souboru SASUSER.FITNESS.

- **Procedura:**

```
proc univariate data=sasuser.fitness mu0=50 cibasic normal plot trimmed=2
winsorized=2;
var oxygen;
run;
```

- **Příkazy:**

- Cibasic – výpočet intervalů spolehlivosti pro základní statistické charakteristiky.
- Normal – výpočet testů normality rozdělení.
- Plot – konstrukce jednoduchých vizualizačních prostředků (stem-and-leaf display, boxplot, qqplot)
- Trimmed= – výpočet useknutého (cenzorovaného) průměru spolu s výpočtem intervalu spolehlivosti pro tento průměr a testem hypotézy, že průměr základního souboru je 50.
- Winsor= – výpočet winsorizovaného průměru spolu s příslušným intervalem spolehlivosti pro průměr a jednovýběrovým testem hypotézy o hodnotě průměru.

- **Výstupy:**

- Testy polohy – standardně tři testy, Studentovo t je jednovýběrový t-test, který řadíme mezi parametrické testy, tyto testy požadují, aby soubor měl normální rozdělení
- Testy normality – standardně čtyři testy, je ponecháno na analytikovi, aby vybral test, ke kterému má důvěru, pro malé soubory (do 2000 pozorování) je vhodný test Shapiro-Wilk. P-hodnota je 0,1968, nulová hypotéza říká, že data mají normální rozdělení, jestliže p-hodnota je téměř 20%, nemáme důvod, abychom hypotézu zamítli, a konstatujeme, že soubor má normální rozdělení a můžeme použít jednovýběrový t-test. Přesto musíme být opatrní, výpočet je vhodné doplnit graficky, v proceduře proto byl příkaz plot.
- Graf pravděpodobnosti normálního rozdělení – zobrazuje se jako přímka
- Pokud se neprokáže normální rozdělení, využijeme znaménkový test (znaménko) nebo jednovýběrový Wilcoxonův test (znam. pořadí). Patří mezi neparametrické testy, oproti parametrickým nepožadují normální rozdělení. Wilcoxonův test se kvalitou blíží jednovýběrovému t-testu, znaménkový je relativně jednoduchý a má horší vlastnosti. Všechny testy dávají závěr, že nulovou hypotézu zamítneme, protože všechny hodnoty jsou menší než 5%. Rozhodování o volbě testu – aby boxplot byl symetrický, medián, aby ležel zhruba uprostřed. Hodnotám, které překročí 1,5 násobek kvantilového rozpětí, je nutné věnovat zvýšenou pozornost.
- Winsorizace, zapíšeme-li winsor=2, jsou dvě největší hodnoty nahrazeny třetí největší hodnotou, která již nebyla označena jako odlehlá. Stejná operace se provede na opačném konci souboru. Soubor se nezmenšuje, pouze se jedná o nahrazení hodnot.
- Useknutý (cenzorovaný) průměr – vznikl odseknutím dvou největších a dvou nejmenších hodnot

- ✓ **Procedura means:**

- **Procedura:**

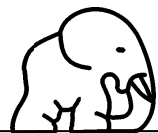
```
proc means data=sasuser.fitness;
var oxygen;
run;
```

- **Rozšíření stručného výstupu:**

```
proc means data=sasuser.fitness n mean median min max q1 q3 range qrange std
cv skewness kurtosis maxdec=3;
var oxygen age weight runtime repulse repulse;
run;
```

- **Argumenty:**

- n – počet prvků
- mean – průměr



- stddev – směrodatná odchylka
- min – minimum
- max – maximum
- median – medián
- q1, q3 – dolní a horní kvartil
- range – variační rozpětí
- qrange – mezikvartilové rozpětí
- cv – variační koeficient
- skewness – šikmost
- kurtosis – špičatost
- maxdec=3 – maximální počet desetinných míst

✓ **Příklad:**

- Množství podkožního tuku, muži a ženy
- **Procedura** – možnost prohlédnutí datového souboru:

```
proc print data=svs;  
var fat gender;  
run;
```

- **Procedura:**

```
proc means data=svs;  
class gender;  
var fat;  
run;
```

Výstup bude tříděn podle pohlaví

- **Procedura** – t-test s možností stanovení titulku:

```
proc ttest data=svs;  
class gender;  
var fat;  
title „Porovnání skupin“;  
run;
```

T-testy – oba soubory mají mít stejný rozptyl, pouze tehdy t-test dobře funguje. Rovnost variancí – prověřování, zda mají oba soubory stejnou variabilitu z hlediska hodnoty podkožního tuku. Pokud je p-hodnota větší než 5%, nezamítáme.