



SEMINÁŘ VÝPOČETNÍ STATISTIKY

P2
2008-02-25

Zobrazování datového souboru:

✓ Semigrafická technika (napůl grafická) – **Stem-and-leaf display** (stemplot):

- **Příklad:**

- Číselný soubor: 35, 36, 38, 40, 42, 42, 44, 45, 45, 47, 48, 49, 50, 50, 50
- Konstruování stemplotu: čísla rozdělíme do dvou částí:
 - Odsekne desítkovou část, tzv. lodyhu nebo stonek
 - Listy je to, co zbude

Stem	Leaves
3	5, 6, 8
4	0, 2, 2, 4, 5, 5, 7, 8, 9
5	0, 0, 0

- Pokud otočíme o 90°, připomínají „řádky“ sloupce histogramu => vizuální interpretace, ale na rozdíl od histogramu jsou vidět hodnoty. Lze objevit mezery, asymetrie atd.

- **Příklad:**

- Výsledky:

Fat	Gender
13	m
19	m
20	m
8	m
18	m
22	m
20	m
31	m
21	m
12	m
16	m
12	m
24	m

Fat	Gender
22	f
26	f
16	f
12	f
22	f
23	f
21	f
28	f
30	f
23	f

- Stemplot:

F	Stem	M
	0	8
2, 6	1	3, 9, 8, 2, 6, 2
3, 8, 1, 3, 2, 6, 2	2	0, 2, 0, 1, 4
0	3	1

V případě žen jsou hodnoty více koncentrované.

- Vhodné u souborů s menším rozsahem, u větších bychom dali přednost klasickému histogramu.
- V SASu se stemplot zobrazí automaticky, ale řádky se ještě půlí, v jednom řádku se píší hodnoty 0 – 4 a ve druhém 5 – 9. Důvodem je délka řádků.

Stem	
2	0, 2, 3, 3, 4
2	5, 6, 9

- Procedura:

```
proc univariate data=jmeno plot;  
...  
run;
```

Příkaz plot vyvolá data ve stemplotu.

Kvalitativní znaky:

- ✓ Jejich jednotlivé varianty můžeme popsat pouze slovně, nelze vyjádřit číselně. Je možné spočítat, kolikrát se jednotlivá varianta objeví nebo procentuální zastoupení.

✓ **Vizualizace údajů o kvalitativních znacích** – příklad:

- Výsledky:

Strana	%
A	38,5
B	38,5
C	7,4
D	8,6
E	4,0
F	3,0

- Lze využít **sloupcový diagram** (Bar Chart)

- Jednobarevný graf, jednotlivé skupiny nejsou uspořádány podle velikosti.

```
proc chart data=svs;
```

```
hbar strana/sumvar=podil; Proměnná „strana“ je kvalitativní, příkaz sumvar sečte hodnoty.
```

```
run;
```

Výsledný graf – horizontální, po levé straně jednotlivá označení (strana A, B...), na pravé straně konkrétní procentuální hodnoty. Chceme-li vertikálními, píšeme vbar místo hbar.

- Modifikace barev:

```
proc chart data=svs;
```

```
hbar strana/sumvar=podil subgroup=strana;
```

```
run;
```

Výsledný graf – každá strana vlastní barva

- Uspořádání sloupečků sestupně:

```
proc chart data=svs;
```

```
hbar strana/sumvar=podil subgroup=strana descending;
```

```
run;
```

- Uspořádání od nejkratší – místo descending se píše ascending.

- **Výsečový (koláčový) graf**:

- Procedura:

```
proc gchart data=svs;
```

```
pie strana/sumvar=podil;
```

```
run;
```

- **Koblihový graf** – místo pie se píše donut:

- Procedura:

```
proc gchart data=svs;
```

```
donut strana/sumvar=podil;
```

```
run;
```

- Uprostřed koláče je díra.

- **Trojrozměrné výsečové (koláčové) grafy**:

- Procedura:

```
proc gchart data=svs;
```

```
pie3d strana/sumvar=podil slice=arrob explode="A" „B“;
```

```
run;
```

- Příkazy:

- Sumvar=<variable> – počítá součet hodnot dané proměnné.

- Slice=arrow/inside/none/outside – ovlivňuje popis zvoleného segmentu zobrazované proměnné.

- Arrow – ke každému segmentu povede čára s informací o názvu strany a procentech.

- Outside – bez čáry, pouze se vypíše jméno a procenta - ven

- Insight – informace se vypíše dovnitř

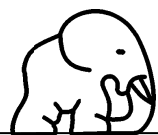
- None – nebude žádný popis

- Explode=<seznam> – uvádí seznam oddělených segmentů, tyto části grafu se zobrazí odděleně (mezerou) od ostatních.

- Psaní poznámky – na začátku hvězdička a na konci středník, správná poznámka je zapsána zeleně, lze ji vepsat kamkoli.

Například: *Grafické zobrazení pomocí trojrozměrného „koláčového grafu“;

- Segmenty, které jsou zastoupeny méně než 5%, jsou v grafu automaticky spojeny a označeny jako „other“.

✓ **Kvalitativní znaky** – příklad:• **Výsledky:**

Akcie	Počet
A	2.020
B	1.200
C	890
D	630
E	1500

• **Grafické zobrazení pomocí trojrozměrného koláčového grafu:**▪ Procedura:

```
proc chart data=svs;  
pie3d akcie/sumvar=počet noheading percent=arrow value=inside slice=arrow  
explode="C";  
run;
```

▪ Příkazy:

- Noheading – potlačuje tisk hlavičky (nadpisu).
- Percent=arrow/inside/none/outside – připisuje jednotlivým segmentům jejich procentuální vyjádření.
- Value=arrow/inside/none/outside – připisuje jednotlivým segmentům jejich absolutní hodnoty.
- Slice=arrow/inside/none/outside – ovlivňuje popis zvoleného segmentu zobrazované proměnné

Textový výstup ze SASu:

✓ Stemplot rozepsán v tabulce – kmen a list, po otočení obdoba histogramu.

✓ **Odlehlé či extrémní hodnoty:**

- **Vybočující** hodnoty jsou zobrazeny v jednoduchém „box plotu“ na pravé straně tabulky. Písmeno O znamená outlier, tedy odlehlé pozorování. Další hodnoty označené pouze hvězdičkami znamenají extrémní pozorování.
- U odlehlých a extrémních hodnot nutno **ověřit správnost** hodnot. Pokud jsou hodnoty správné, můžeme hodnoty v souboru nechat (zkreslí se průměr), nebo je vyřadit (cenzorování).
- SAS nabízí **cenzorování**, které se však provádí symetricky, takže pokud se vyřadí tři nejhorší, tak se vyřadí automaticky také tři nejlepší. Nevýhoda u malých souborů, ztráta informace může být závažná.
- **Winsorizace** – metoda, která hodnoty potlačuje či koriguje, ale nevyřazuje. Problematická hodnota se nahrazuje hodnotou těsně před těmito hodnotami. Totéž se provádí z druhé strany.
- **Výběr techniky** záleží na analytikovi.