2. ZÁKLADNÍ POHLED NA DATA A PROMĚNNÉ

2.1 Získání základních informací o proměnných

DESCRIBE – vypíše přehled všech proměnných nacházejících se v otevřeném datovém souboru včetně informací o jejich typu (byte, int, long – celá čísla, double, float – desetinná čísla, str – textový řetězec), popisu proměnné (*variable label*) a popisu jednotlivých hodnot (*value label).* Pokud vás zajímají informace o jedné konkrétní proměnné, napište její název za příkaz desribe.

```
describe year
```

LABEL LIST – vypíše popis jednotlivých hodnot dané proměnné. Nejprve je potřeba pomocí příkazu describe zjistit, jak se popisek hodnot jmenuje.

```
label list YEAR
```

Tip: Stata je citlivá na velikost znaků (*case sensitive*), takže slova year, Year a YEAR označují naprosto odlišné věci. Obvykle se používají malá písmena pro názvy proměnných a velká písmena pro názvy popisků hodnot. Překlep ve velikosti písmen je navíc jedním z nejčastějších chyb, proč váš příkaz nefunguje.

2.2 Základní tabulky

TABULATE – vypíše frekvenční tabulku zvolené proměnné. Je možno tento příkaz zkrátit jako **TAB.**

```
tab year
```

Podobu tabulky je možno upravit řadou parametrů zapisovaných za čárku, mezi ty nejpoužívanější patří MISSING, NOFREQ a NOLABEL.

TABULATE MISSING – do výpočtu procent ve frekvenční tabulce budou zahrnuty i chybějící hodnoty.

TABULATE NOFREQ – v tabulce nebudou vypsány četnosti jednotlivých hodnot.

TABULATE NOLABEL – v tabulce nebudou zobrazeny popisky jednotlivých hodnot, ale pouze jejich číselné kódy.

TABULATE PLOT – součástí tabulky bude jednoduchý histogram.

Tip: Parametry je možno libovolně kombinovat, vždy ale platí, že se píší až na konec příkazu (za seznam proměnných) a jsou od proměnných odděleny právě jednou čárkou. Jednotlivé parametry už se pak od sebe oddělují pouze mezerou.

```
tab v1, nolabel plot
```

TABSTAT – vypíše vybrané statistické charakteristiky zvolené proměnné. Standardně (bez doplňujícího parametru) vypisuje statistický průměr (*mean*). Pokud potřebujete jinou charakteristiku, zařaďte ji do závorky u parametru **STAT.** Vybrané použitelné charakteristiky: průměr (*mean*), počet (*count* nebo *n*), součet *(sum*), minimum (*min*), maximum *(max*), standardní odchylka (*sd*), jednotlivé percentily (*p1, p5, p10, p25, p50, p75, p90, p95, p99*), kvartily a medián (*q1, q2, q3, q4, median*). Všechny charakteristiky jsou k nalezení v manuálových stránkách příkazu **TABSTAT**.

Jednotlivé charakteristiky je možno řadit za sebe v libovolném pořadí, oddělují se mezerami.

```
tabstat v1, stat (mean median min max)
```

Pokud potřebujete charakteristiky více proměnných současně, napište je vedle sebe oddělené mezerou.

```
tabstat v1 v2 v3, stat (mean median min max)
```

2.3 Kontingenční tabulky

TABULATE – kontingenční tabulky (*crosstabs*) lze snadno vytvořit pomocí již známého příkazu **TABULATE** tak, že se za něj uvedou dvě proměnné. Stata pak první proměnnou použije pro označení sloupců a druhou proměnnou pro označení řádků. Fungují zmíněné parametry **MISSING, NOLABEL, NOFREQ** atd.

```
tab v1 v2
tab v2 v1
```

Dalšími důležitými parametry jsou parametry pro řádková a sloupcová procenta. Sloupcová procenta Stata vypočte po zadání parametru **COL** (*column*) a vyznačují se tím, že součet všech procent v každém sloupci je roven 100 %. Řádková procenta Stata vypočte po zadání parametru **ROW** (*row*) a vyznačují se tím, že součet všech procent v každém řádku je roven 100 %.

```
tab v1 v2, row
tab v1 v2, nofreq row
```

2.4 Váhy a vážení

Reálná data získaná kvótním výběrem se často vyznačují nedokonalou reprezentativností. Abychom tento nedostatek co nejvíce odstranili, používá se tzv. vážení (*weighting*). Každému případu je přiřazena váha (*weight*), která označuje, jak velkou váhu má Stata tomuto případu přikládat.

Příklad: Dejme tomu, že v základní populaci, na které provádíme dotazníkové šetření, je přesně vyrovnaný poměr žen a mužů. Podaří se nám ale získat data jen od 400 žen a 500 mužů. Aby statistická analýza lépe odpovídala sociální realitě, musí Stata přiřadit každému muži nižší váhu, v tomto případě je každá odpověď muže vynásobena hodnotou 0,8 (která se vypočte jako podíl 400/500). Reálný výpočet vah je samozřejmě mnohem komplikovanější, protože se počítá podle více charakteristik, např. podle věku, vzdělání, bydliště apod. Vypočtené váhy bývají obvykle uloženy v některé proměnné, která se jmenuje weight nebo podobně.

Stata rozeznává několik typů vah, jejichž použití můžete u většiny příkazů explicitně stanovit. Existuje ale také obecný příkaz, který nechá Statu, aby sama vybrala, který typ vah je podle ní pro konkrétní situaci nejlepší.

aw – analytické váhy, určují, kolik osob by mělo mít podobnou charakteristiku jako příslušný případ.

fw – frekvenční váhy, vyjadřují frekvenci, kolikrát má být konkrétní případ zopakován. Musí být celočíselné.

iw – váhy stanovující důležitosti jednotlivých případů (*importance*).

pw – vzorkovací váhy, které upravují chybu způsobenou nesprávnou konstrukcí vzorku.

w – Stata sama rozhodne, který typ vah je nejvhodnější. Ne vždy ale musí rozhodnout správně.

Příkaz k použití vah se píše do hranatých závorek za seznam proměnných. Za typem vah (aw, fw, iw, pw, w) následuje rovnítko a název proměnné, která obsahuje informaci o váze jednotlivých případů.

```
tab v1 v2 [aw=weight], row
tab v1 v2 [iw=weight], row
```