3. POHLED NA DATA A PROMĚNNÉ

Data jsou ve Statě uložena v datové matici. Jednotlivé případy (respondenti) jsou uvedeny v řádcích, jejich charakteristiky (hodnoty proměnných) pak ve sloupcích. Datový soubor se otevírá pomocí příkazu USE, ke zobrazení datové matice slouží příkaz BROWSE, k úpravám dat pak příkaz EDIT.

3.1 Získání základních informací o proměnných

DESCRIBE – vypíše přehled všech proměnných nacházejících se v otevřeném datovém souboru včetně informací o jejich typu (*type*), popisu proměnné (*variable label*) a popisu jednotlivých hodnot (*value label).* Pokud vás zajímají informace o jedné konkrétní proměnné nebo více proměnných, napište její název za příkaz describe.

```
describe educ
```

Přehled datových typů:

- byte: celé číslo v rozsahu -127 – 100
- int: celé číslo v rozsahu -32 767 – 32 740
- long: celé číslo v rozsahu -2 147 483 647 – 2 147 483 620
- float: desetinné číslo v rozsahu $-1{,}70141173319*10^{38}$ – $1{,}70141173319*10^{38}$
- double: desetinné číslo v rozsahu $-8.9884656743*10^{307}$ – $8.9884656743*10^{307}$
- str: textový řetězec, číslo udává maximální délku řetězce, např. str7 může obsahovat 7 znaků

Tip: pokud si nejste dopředu jistí, jaký typ proměnné použít, zvolte nejvyšší možný (long pro celá čísla, float pro desetinná čísla). Před uložením datového souboru pak použijte příkaz COMPRESS. Stata provede analýzu obsahu proměnných a sama zvolí nejvhodnější a nejúspornější datový formát.

```
. describe

Contains data from D:\_FSS\phd\STATA predmet FRMU\2017\data\CHPS_w1.dta
  obs:        13,049
 vars:           666                          13 Feb 2017 10:38
 size:    34,436,311
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
hid             long   %12.0f                 Číslo domácnosti
pid             double %12.0g                 Číslo osoby
pno             byte   %12.0f                 Číslo člena domácnosti
pidhr           double %12.0g                 pid respondenta, který vyplnil dotazník pro domácnost
pidpro          double %12.0g                 pid proxy respondenta
hhsize          byte   %12.0f                 H1. Počet členů domácnosti
rstat           byte   %12.0f      rstat      H3. Rezidenční status člena domácnosti
sex             byte   %12.0f      sex        H4. Pohlaví člena domácnosti
birthm          byte   %12.0f      birthm     H5. Měsíc narození člena domácnosti
birthy          int    %12.0f      birthy     H6. Rok narození člena domácnosti
age             byte   %12.0f      age        Věk
agecat          byte   %12.0f      agecat     Věk kategorizovaný po 10 letech
hres            byte   %12.0f      hres       H7. Osoba vyplnila dotazník pro domácnost
mstat           byte   %12.0f      mstat      H8. Rodinný stav člena domácnosti
estat           byte   %12.0f      estat      H9. Ekonomické postavení člena domácnosti
educ            byte   %12.0f      educ       H10. Nejvyšší dosažené vzdělání člena domácnosti
std             byte   %12.0f      std        H11. Současné studium člena domácnosti
stdch           byte   %12.0f      stdch      H12. Škola navštěvovaná dětmi ve věku 10 až 17 let
rel1            byte   %12.0f      rel1       H13.1. Člen domácnosti pno 1 je pro osobu:
rel2            byte   %12.0f      rel2       H13.2. Člen domácnosti pno 2 je pro osobu:
rel3            byte   %12.0f      rel3       H13.3. Člen domácnosti pno 3 je pro osobu:
--more--
```

Obrázek 1 Ukázka použití příkazu describe

SUMMARIZE – zobrazí základní informace o proměnné. Parametr DETAIL přidá další statistické charakteristiky

```
summarize educ, detail
```

INSPECT – vypíše podrobnější informace o zvolené proměnné, konkrétně popisek proměnné, přehled rozložení hodnot, počet pozorování, počet chybějících hodnot a jednoduchý histogram.

```
inspect educ
```

CODEBOOK – vypíše informace o proměnných. Bez uvedení jména proměnné vypíše informace o všech proměnných v datovém souboru, při uvedení konkrétní proměnné nebo proměnných omezí výpis na informace pouze o zvolených proměnných. Výpis je možno upravit parametrem COMPACT pro zobrazení vybraných informací ve zkráceném formátu.

```
codebook
codebook educ, compact
```

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

```
label list educ
```

Tip: Stata je citlivá na velikost znaků (*case sensitive*), takže slova educ, Educ a EDUC označují naprosto odlišné věci. Obvykle se používají malá písmena pro názvy proměnných a velká písmena pro názvy popisků hodnot, není to ale pravidlem. Překlep ve velikosti písmen je navíc jednou z nejčastějších chyb, proč váš příkaz nefunguje.

```
. label list rstat
rstat:
           1 Resident (obvykle v bytě bydlí)
           2 Studium mimo domov
           3 Práce mimo domov
           4 Pobyt v instituci
          88 Nevím, nemohu posoudit
          99 Odmítl(a)
```

Obrázek 2 Ukázka využití příkazu label

3.2 Základní frekvenční tabulky

TABULATE – vypíše frekvenční tabulku zvolené proměnné. Podobně jako u všech příkazů je možno zkrátit název tak, aby Stata dokázala příkaz odlišit od všech jiných příkazů. V tomto případě lze použít **TAB,** ale nikoliv např. pouze T**.**

```
tab educ
```

V tabulce jsou popis proměnné (*label*), výčet všech možných hodnot (*values*) popsaných slovně (*values label*), počet výskytů jednotlivých možností (*Freq*), procentní podíl (*Percent*) a kumulativní procenta uvádějící součet aktuální procentní hodnoty se všemi předchozími řádky (*Cum.*). Ve spodním řádku pak vidíme celkový počet výskytů a celkový součet procent (100). Podobu tabulky je možno upravit řadou parametrů zapisovaných za čárku, mezi ty nejpoužívanější patří MISSING, NOFREQ a NOLABEL.

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

NOFREQ – v tabulce nebudou vypsány četnosti jednotlivých hodnot.

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

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 educ, nolabel plot
```

```
. tab educ
```

H10. Nejvyšší dosažené vzdělání člena domácnosti	Freq.	Percent	Cum.
Osoba se nikdy nevzdělávala	966	7.40	7.40
Neúplné základní	1,261	9.67	17.07
Základní	1,231	9.44	26.50
Vyučení bez maturity	2,397	18.37	44.88
Střední bez maturity	790	6.06	50.93
Vyučení s maturitou	236	1.81	52.74
Střední odborné s maturitou	2,993	22.94	75.68
Střední všeobecné s maturitou (gymnáziu	704	5.40	81.08
Vyšší odborné	236	1.81	82.88
Vysokoškolské bakalářské	405	3.10	85.99
Vysokoškolské magisterské (inženýrské,	1,656	12.69	98.68
Vysokoškolské doktorské (postgraduální,	172	1.32	100.00
Total	13,047	100.00	

Obrázek 3 Základní frekvenční tabulka

3.3 Kontingenční tabulky

Kontingenční tabulky ukazují souvislost dvou (a více) proměnných. Pomocí kontingenčních tabulek (*crosstabs*) můžeme například zjistit, jak se liší vzdělanostní rozložení mužů a žen. Ve Statě se kontingenční tabulky vypisují stejně jako tabulky jednoduché, jen se za příkaz TAB uvedou dvě proměnné. I v tomto případě lze používat většinu parametrů.

```
tab educ sex, missing
```

```
. tab educ sex, col nofreq
```

H10. Nejvyšší dosažené vzdělání člena domácnosti	H4. Pohlaví člena domácnosti: Muž	Žena	Total
Osoba se nikdy nevzdě	7.87	6.97	7.40
Neúplné základní	10.51	8.89	9.67
Základní	7.66	11.07	9.44
Vyučení bez maturity	22.09	14.95	18.37
Střední bez maturity	5.92	6.18	6.06
Vyučení s maturitou	2.24	1.41	1.81
Střední odborné s mat	21.21	24.53	22.94
Střední všeobecné s m	4.03	6.65	5.40
Vyšší odborné	1.33	2.25	1.81
Vysokoškolské bakalář	2.69	3.49	3.10
Vysokoškolské magiste	12.81	12.58	12.69
Vysokoškolské doktors	1.63	1.03	1.32
Total	100.00	100.00	100.00

Obrázek 4 Kontingenční tabulka se sloupcovými procenty

Protože v každé skupině (muži, ženy) je odlišný počet respondentů, nejsou hodnoty získané touto analýzou srovnatelné. 100 mužů se základním vzděláním vypovídá o jiné situaci než 100 žen se základním vzděláním, protože počet mužů a žen není stejný. Z tohoto důvodu je výhodnější uvádět procentní hodnoty. Ty jsou vždy vztaženy k celku a lze tak snadno porovnávat podíly jednotlivých vzdělanostních skupin.

V kontingenčních tabulkách se rozlišují dva typy procent. Sloupcová procenta získáte zadáním parametru COL. Stata k jednotlivým hodnotám vypočte jejich podíl z celkového součtu sloupce. V našem případě sloupcová procenta říkají, **jaký podíl ze všech mužů má daný stupeň vzdělání** a **jaký podíl ze všech žen má daný stupeň vzdělání.** Oproti tomu lze využít i řádková procenta pomocí parametru ROW. Stata k jednotlivým hodnotám vypočte jejich podíl z celkového součtu řádku. V našem případě řádková procenta říkají, **jaký podíl lidí se základním vzděláním tvoří muži a jaký ženy.**

```
tab educ sex, col
```

Řádková a sloupcová procenta se často pletou. Vždy je vhodné podívat se do sloupce/řádku na hodnotu Total. Pokud je 100 % uvedeno v řádku, jedná se o řádková procenta, pokud ve sloupci, jedná se o sloupcová procenta. Pak postupujeme následovně. Za větu „řádková/sloupcová procenta vyjadřují, jaký podíl“ doplníme název proměnné, která se vyskytuje v řádcích (pro řádková procenta) či ve sloupcích (pro sloupcová procenta) a pokračujeme dále: „spadá do skupiny“ s doplněním jednotlivých hodnot druhé proměnné.

```
. tab educ sex, row nofreq
```

H10. Nejvyšší dosažené vzdělání člena domácnosti	H4. Pohlaví člena domácnosti		
	Muž	Žena	Total
Osoba se nikdy nevzdě	50.93	49.07	100.00
Neúplné základní	52.10	47.90	100.00
Základní	38.91	61.09	100.00
Vyučení bez maturity	57.61	42.39	100.00
Střední bez maturity	46.84	53.16	100.00
Vyučení s maturitou	59.32	40.68	100.00
Střední odborné s mat	44.30	55.70	100.00
Střední všeobecné s m	35.80	64.20	100.00
Vyšší odborné	35.17	64.83	100.00
Vysokoškolské bakalář	41.48	58.52	100.00
Vysokoškolské magiste	48.37	51.63	100.00
Vysokoškolské doktors	59.30	40.70	100.00
Total	47.91	52.09	100.00

Obrázek 5 Kontingenční tabulka s řádkovými procenty

Pro náš případ tedy nejprve zjistíme, že 100 % je uvedeno ve sloupcích, proto se jedná o sloupcová procenta. Ve sloupcích je uvedena proměnná pohlaví. Říkáme proto, že procentní hodnota vyjadřuje, jaký podíl z jednotlivých skupin pohlaví tvoří lidé s určitým stupněm vzdělání.

Tip: Stata na začátku kontingenční tabulky vypíše schematicky strukturu buňky, snadno tak zjistíte význam jednotlivých řádků. Jelikož uvádět absolutní hodnoty nemá v podstatě žádný smysl, je vhodné použít parametr NOFREQ.

3.4 Filtrování případů a dělení do skupin

Někdy se ale může hodit, když vypíšeme kontingenční tabulku jen pro určité skupiny respondentů, například jen pro respondenty starší 30 let. Ve Statě existuje několik postupů, jak toho lze dosáhnout: příkaz IF a příkaz BY (resp. BYSORT).

IF – podmínkový příkaz, který vybere určité případy splňující definovanou podmínku. Podmínka se uvádí na konec seznamu proměnných, za ni je pak možno přiřadit libovolné parametry. Jednotlivé případy jsou vybrány, pokud je podmínka pro jejich případ platná. Např. následující podmínka vybere všechny případy, při nichž je respondent ženatý či respondentka vdaná:

```
tab educ sex if (mstat==1), missing
```

Podmínka je konstruována pomocí matematických a logických operátorů. Uvedeme si jen nejdůležitější z nich:

== rovná se je vyjádřeno pomocí dvou rovnítek. Podmínka (sex==1) platí, pokud je proměnná v291 rovna hodnotě 1. Pozor, častá chyba je porovnávání pomocí jednoho rovnítka!

!= nerovná se je vyjádřeno kombinací vykřičníku a rovnítka. Podmínka (sex!=1) platí, pokud se proměnná v291 nerovná 1 (tedy pro všechny ostatní hodnoty)

>, <, >=, <= znamená (v pořadí zleva doprava) větší než, menší než, větší nebo rovno než, menší nebo rovno než

& logická spojka a (and). Podmínka platí, pokud platí všechny podmínky oddělené spojkou &. Např. (mstat>1)&(mstat<3) platí tehdy, pokud je proměnná mstat větší než 1 A SOUČASNĚ je proměnná mstat menší než 3.

| logická spojka nebo (or). Podmínka platí, pokud platí alespoň jedna podmínka oddělená spojkou |. Např. (mstat>10)|(mstat<5) platí tehdy, pokud je proměnná mstat větší než 10 NEBO je proměnná v291 menší než 5.

Podmínky je možné libovolně řetězit, je vhodné jednotlivé části podmínek uzavírat do závorek, které mají stejný význam jako v matematice. Stata nejprve vyhodnocuje podmínku uzavřenou v závorce jako celek. Poté postupuje zleva doprava. Např. můžeme vytvořit podmínku (((mstat==1) | (sex==2)) & (educ>5)).

BY – většina příkazů ve Statě umožňuje použití příkazu nebo parametru BY, který způsobí, že je daný příkaz zopakován pro jednotlivé hodnoty zvolené proměnné. Pokud např. uvažujeme výše uvedený příklad s pohlavím respondenta, za pomocí příkazu BY zajistíme, že Stata vypíše kontingenční tabulku zvlášť pro muže a zvlášť pro ženy. Máme tedy následující dvě možnosti:

```
tab educ mstat if (sex==1)
tab educ mstat if (sex==2)
```

nebo

```
by sex, sort: tab educ mstat
```

Parametr SORT u příkazu BY způsobí, že Stata jednotlivé případy nejprve seřadí podle proměnné sex. Bez předchozího seřazení není Stata schopna vypsat kontingenční tabulku pro jednotlivé skupiny proměnných. V praxi můžete mít štěstí na soubor, který už je podle dané proměnné setříděn, nelze s tím ale dopředu počítat, proto je vhodnější rovnou si zapamatovat syntaxi včetně parametru SORT.

BYSORT – novější verze Staty nabízí příkaz, který v sobě přímo obsahuje pokyn k seřazení souboru. Používá se stejně jako příkaz BY:

```
bysort sex: tab educ mstat
```

3.5 Souhrnné statistické charakteristiky

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 age, stat (mean median min max)
```

```
. tabstat age, stat (mean median min max)
```

variable	mean	p50	min	max
age	40.52633	41	0	94

Obrázek 6 Ukázka využití příkazu tabstat

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

```
tabstat age payn, stat (mean median min max)
```

I pro příkaz TABSTAT je možno využít již známé doplňkové příkazy IF, BY nebo BYSORT.

```
. bysort sex: tabstat age, stat (mean median min max)
```

-> sex = Muž

variable	mean	p50	min	max
age	39.4864	40	0	94

-> sex = Žena

variable	mean	p50	min	max
age	41.48257	42	0	94

Obrázek 7 Ukázka využití příkazu tabstat ve spojení s bysort

3.6 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 educ sex [aw=W_indi], row
tab educ sex [iw=W_indi], row
```

Všimněte si, že po zapnutí vážení už nedávají absolutní hodnoty vůbec žádný reálný smysl. O to důležitější je nyní striktně využívat správně zvolené procentní podíly.

3.7 Grafická prezentace výsledků

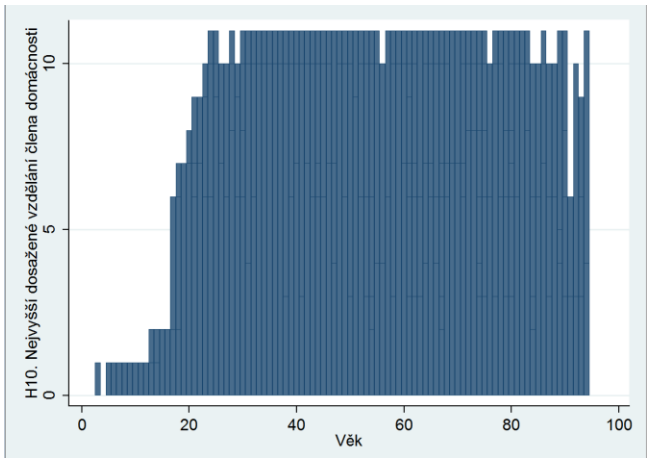
PLOT – zobrazí jednoduchý textový graf, od Staty verze 8 už se dál nevyvíjí

```
plot educ age
```

GRAPH – moderní příkaz pro vykreslování grafů ve Statě. Nabízí široké možnosti nastavení, jak má graf vypadat a co má obsahovat za informace. Všechny možnosti lze najít v nabídce Graphics.

GRAPH TWOWAY BAR – sloupcový graf

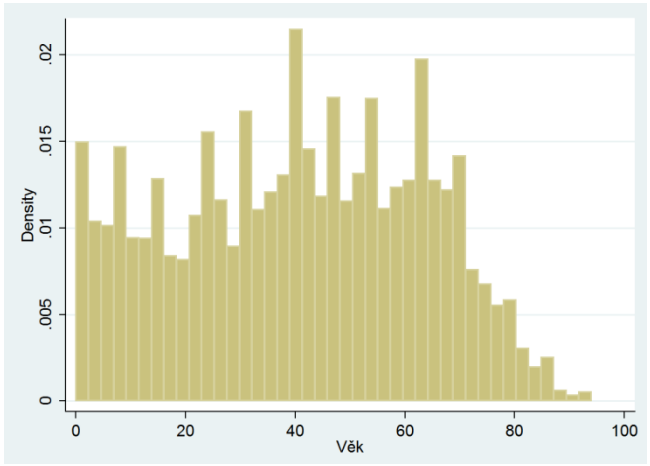
```
graph twoway bar educ age
```



Obrázek 8 Sloupcový graf

GRAPH TWOWAY HISTOGRAM – histogram jedné proměnné

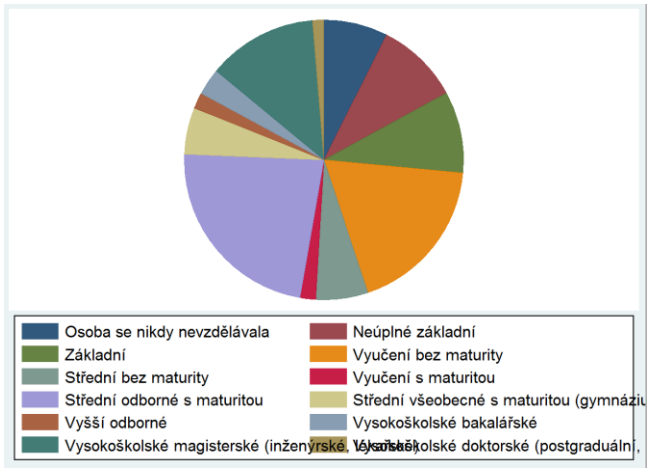
```
graph twoway histogram age
```



Obrázek 9 Histogram

GRAPH TWOWAY PIE – koláčový graf jedné proměnné

```
graph pie, over (educ)
```

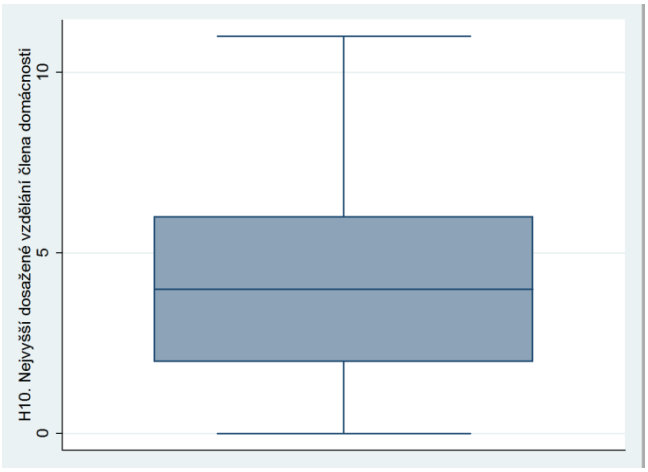


Obrázek 10 Koláčový graf

GRAPH TWOWAY BOX – boxplot

```
graph box educ
```

```
graph hbox educ
```



Obrázek 11 Box plot

Krabice je zdola ohraničená prvním kvartilem, shora třetím kvartilem, v krabici tak leží 50 % hodnot. Čára v krabici označuje hodnotu mediánu. Vousy pak označují minimum a maximum, případné mimolehlé hodnoty (*outliers*) jsou označeny kolečky mimo naznačený rozsah.