# My basic Stata companion

Niels Henrik Bruun

Research data and statistics, AaUH

1. Start

2. Example data and setting up a project

3. Important Stata commands

4. Graphing in Stata

5. Where to go now?

# Section 1

## Start

# Getting started

Start by seeing the following videos:

- The user interface
  - Stata youtube channel: Tour of the Stata user interface (3:39m)
  - Amy Penn: Introduction to Stata - Getting started (5:01m)
  - Amy Penn: Opening a dataset (1:11m)
- Using Stata
  - Amy Penn: Introduction to Stata - Thinking like Stata (14:58m)
    - The link for the example dataset does not work. Try the example dataset used later in this presentation
  - Amy Penn: Introduction to Stata - Data Cleaning using the Codebook and Sort Commands (3:10m)
  - Amy Penn: Introduction to Stata - Generating variables using the generate, replace, and label commands (8:30m)
- Saving commands and logging output
  - Amy Penn: Introduction to Stata - Do Files & Log Files (5:10m)

# `help`, the most important command

## `help|h [command|keywords]`

- Reading syntax
  - **first**|**second**|**third**: choose one of the three
  - **[]**: Within brackets means optional
  - **if**: a qualifier limiting the scope of the command, eg if male == 1
  - **in first/last**: select only rows with row numbers between first and last

# Section 2

## Example data and setting up a project

# Example project

Infant mortality rates and birth defect rates are very high for low birth weight babies.

Hence, low birth weight is an outcome that has been of concern to physicians for years.

**The aim is to see if a set of variables has an effect on birth weights**

# cd ["][*directory*]["]

cd changes the current working directory to the specified drive and directory.

Although optional, it is recommended always to use quotations marks (")

**Examples:**

- Getting (my) current directory in the Result window

```
cd
C:\Users\sttp\Documents\nhb\STATA\Presentations\2020-02 RN Course material
```

- Setting (my) current directory in the Result window

```
cd "C:\Users\sttp\Documents\nhb\STATA\Presentations\2020-02 RN Course material"
C:\Users\sttp\Documents\nhb\STATA\Presentations\2020-02 RN Course material
```

# mkdir and dir

**Examples:**

- Create a sub directory

```
mkdir "Smoking effect on low birth weight"
```

- Check if sub directory is created

```
dir
  <dir>   10/19/20 13:07  .
  <dir>   10/19/20 13:07  ..
 133.0k   10/19/20 13:05  bwt_by_m_age.png
  93.5k   10/19/20 13:05  bwt_hist.png
   9.5k    9/14/20  8:22  Course plan.xlsx
  <dir>    9/26/20 18:23  data
  <dir>   10/19/20 13:07  do-files
  <dir>   10/09/20 18:08  documents
   5.2k   10/19/20 13:05  my lbt.dta
  <dir>   10/19/20 13:05  output
  <dir>    9/26/20 18:23  RawData
  <dir>   10/19/20 13:07  Smoking effect on low birth weight
```

- Change to sub directory

```
cd "Smoking effect on low birth weight"
C:\Users\sttp\Documents\nhb\STATA\Presentations\2020-02 RN Course material\Smoking effect on low birth weight
```

Section 3

# Important Stata commands

# use

### use ["]*filename*["] [, clear]

Load dataset *filename* into the data editor.
Option **clear** is needed to empty the data editor.
Although optional, it is recommended always to use quotations marks (")

**Examples:**

- Retrieving a dataset for analysing low birth weight

```
use "https://www.stata-press.com/data/r16/lbw", clear
(Hosmer & Lemeshow data)
```

# keep|drop

## keep|drop [varlist|if]

keep or drop variables or rows satisfying if-expression

**Examples:**

- Keep the necessary variables *bwt*, *low*, *smoke age* and *race*

```
keep bwt low smoke age race
```

# save

## save ["]*filename*["] [, replace]

Saves dataset *filename* into current directory.
Option **replace** means replacing/overwriting a possible existing dataset.
Although optional, it is recommended always to use quotations marks (")

**Examples:**

- Saving my dataset (names always in quotes)

```
save "my lbt.dta", replace
file my lbt.dta saved
```

# codebook 1/2

## codebook [ *varlist* ] [, compact]

codebook examines the variable names, labels, and data to produce a codebook describing the dataset. Option **compact** makes the description short and in one table

**Examples:**

- Seing all variables and some of their characteristics

```
codebook, compact
Variable     Obs Unique      Mean  Min   Max  Label
------------------------------------------------------------------------------------------------------------------------------

low          189      2  .3121693    0     1  birthweight<2500g
age          189     24   23.2381   14    45  age of mother
race         189      3  1.846561    1     3  race
smoke        189      2  .3915344    0     1  smoked during pregnancy
bwt          189    133  2944.286  709  4990  birthweight (grams)
------------------------------------------------------------------------------------------------------------------------------
```

# codebook 2/2

**Examples:**

- Seing some of their characteristics for variable *race*

```
codebook race
-------------------------------------------------------------------------------------------------------------------------
race
-------------------------------------------------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  race

                 range:  [1,3]                        units:  1
         unique values:  3                        missing .:  0/189

            tabulation:  Freq.   Numeric  Label
                            96         1  white
                            26         2  black
                            67         3  other
```

# An overview of logical expressions

| Symbol | Meaning |
|---|---|
| `>` | Greater than |
| `<` | Lesser than |
| `>=` | Greater than or equal |
| `<=` | Lesser than or equal |
| `==` | Equal to |
| `!=` | Not equal to |
| `&` | Logical **and** |
| `\|` | Logical **or** |
| `!` | Logical **not** |
| `inlist(value1, value2, ...)` | Value1 is equal to one of the following values |

| Example code | Meaning |
|---|---|
| `... if age >= 20` | Mothers of age 20 and above |
| `... if inlist(race, 1, 2)` | Mothers of race white (1) or black (2) |

# generate

## generate|egen *new_varname* =exp [if]

generate|replace creates a new variable. The values of the variable are specified by =exp and possibly [if].

**Examples:**

- Generate variable *bwlt1500* being 1 if children have a birth weight less than 1500 and zero otherwise if a value for birth weight exists

```
generate bwlt1500 = bwt < 1500 if !missing(bwt)
```

# rename

### rename

rename renames one or more variables.

**Examples:**

- Rename variable *low* to bwlt2500 (Birth Weight Less Than 2500)

```
rename low bwlt2500
```

# labels

## labels define|values|variable

labels adds or modifies variable and value labels.
Labels can have different values at different times. In datesets they contain information on content, but they can be modified when variables are used in tables or graphs.

**Examples:**

- Adding labels to variables *bwlt1500* and *bwlt2500*

```
label variable bwlt2500 "Birth weight < 2500g"
label variable bwlt1500 "Birth weight < 1500g"
```

- Defining a value label and attach it to variables *smoke*, *bwlt1500* and *bwlt2500*
  - bwlt* means all variables starting with “bwlt”

```
label define no_yes 0 "no" 1 "yes"
label values smoke bwlt* no_yes
```

# summarize

## summarize [*varlist*] [, detail]

summarize calculates and displays a variety of univariate summary statistics.
Option **detail** gives a detailed summary of the variables in *varlist*

**Examples:**

- A detailed summary over variable *bwt*

```
summarize bwt, detail
                     birthweight (grams)
-------------------------------------------------------------
      Percentiles      Smallest
 1%         1021            709
 5%         1790           1021
10%         1970           1135       Obs                 189
25%         2414           1330       Sum of Wgt.         189

50%         2977                      Mean           2944.286
                        Largest       Std. Dev.       729.016
75%         3475           4174
90%         3884           4238       Variance       531464.4
95%         3997           4593       Skewness      -.2069782
99%         4593           4990       Kurtosis       2.888821
```

# tabulate 1/3

```
tabulate|tab1 varname [if] [, sort]
```

One-way count table for variable
Option **sort** sort rows in descending order based on frequency

```
tabulate|tab2 varname1 varname2 [if] [, chi2 exact]
```

Two-way table for pairwise count combinations
Option **chi2** adds Pearson's chisquare test for row and column dependency
Option **exact** adds Fisher's exact test for row and column dependency

# tabulate 2/3

**Examples:**

- Oneway sorted (highest frequency first) count table for variable *race*

```
tabulate race, sort
       race |      Freq.     Percent        Cum.
------------+-----------------------------------
      white |         96       50.79       50.79
      other |         67       35.45       86.24
      black |         26       13.76      100.00
------------+-----------------------------------
      Total |        189      100.00
```

# tabulate 3/3

**Examples:**

- Pearson's chisquare and Fisher's exact test for dependence between variables *smoke* and *race*

```
tabulate race smoke, chi2 exact
Enumerating sample-space combinations:
stage 3:  enumerations = 1
stage 2:  enumerations = 21
stage 1:  enumerations = 0

           |     smoked during
           |       pregnancy
      race |        no        yes |     Total
-----------+----------------------+----------
     white |        44         52 |        96
     black |        16         10 |        26
     other |        55         12 |        67
-----------+----------------------+----------
     Total |       115         74 |       189

          Pearson chi2(2) =  21.7790   Pr = 0.000
           Fisher's exact =                 0.000
```

# ttest

## ttest *varname* [if] [, by(groupvar)]

Two-sample t test by a binary group variable

### Examples:

- Testing equal mean birth weights between smokers and non-smokers

```
ttest bwt, by(smoke)
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      no |     115    3054.957     70.1625     752.409    2915.965    3193.948
     yes |      74    2772.297    76.70106    659.8075    2619.432    2925.162
---------+--------------------------------------------------------------------
combined |     189    2944.286    53.02811     729.016    2839.679    3048.892
---------+--------------------------------------------------------------------
    diff |            282.6592    106.9544                71.66693    493.6515
------------------------------------------------------------------------------
    diff = mean(no) - mean(yes)                                   t =   2.6428
Ho: diff = 0                                     degrees of freedom =      187

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9955         Pr(|T| > |t|) = 0.0089          Pr(T > t) = 0.0045
```

**We reject the Ho-hypothesis of no evidence of a weight difference against the two-sided alternative, P-value = 0.01.**
**The expected difference in birthweight is 282.659kg and the 95% confidence interval is [71.667; 493.652] (kg).**

# `cs` 1/2

`cs` *outcomevar exposurevar* `[if] [, or exact]`

`cs` is used with cohort study data with equal follow-up time per subject and sometimes with cross-sectional data. Risk is then the proportion of subjects who become cases.
Option **or** makes the command return odds ratio instead of relative risk
Option **exact** adds Fisher's exact test for equal outcome rates between the exposure groups.

# cs 1/2

**Examples:**

- Estimation of relative risk of birthweight below 2500g for smokers vs non-smokers with 95% confidence interval.

```
cs bwlt2500 smoke
                 | smoked during          |
                 | pregnancy              |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+-----------
           Cases |        30          29  |         59
        Noncases |        44          86  |        130
-----------------+------------------------+-----------
           Total |        74         115  |        189
                 |                        |
            Risk |  .4054054    .2521739  |   .3121693
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .1532315       |    .0160718    .2903912
      Risk ratio |         1.607642       |    1.057812    2.443262
 Attr. frac. ex. |          .377971       |    .0546528    .5907112
 Attr. frac. pop |         .1921887       |
                 +-------------------------------------------------
                               chi2(1) =     4.92  Pr>chi2 = 0.0265
```

**We reject the Ho-hypothesis of same risk of birth weight below 2500g for smokers and non-smokers against the two-sided alternative, P-value = 0.03.**
**The expected risk of birth weight below 2500g is 1.61 times higher for smokers than for non-smokers. The 95% confidence interval for the relative risk is [ 1.06; 2.44].**

Section 4

# Graphing in Stata

# histogram|hist

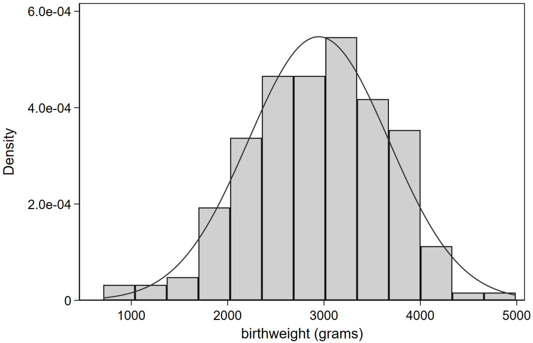
## histogram *varname* [, norm]

histogram generates a histogram for a variable named *varname*.
Option **norm** adds best fitting normal curve. For testing if data is normal distributed

**Examples:**

- Check if birthweight normally distributed

```
histogram bwt, norm
(bin=13, start=709, width=329.30769)
```



# scatter

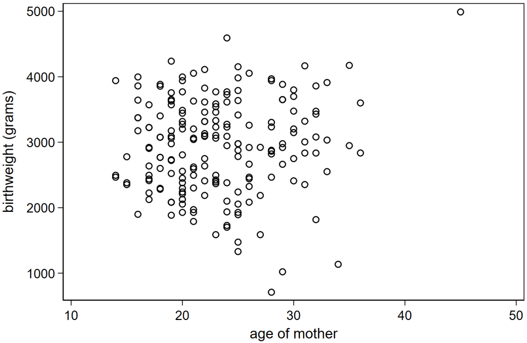
scatter *y-varname x-varname* ‘

scatter plots pairs of x and y values.

**Examples:**

- See dependency how mothers age affect birthweight

```
scatter bwt age
```



Section 5

# Where to go now?

# Useful stuff

Statistics

- seanharrisonblog.com: Series on Evaluation of Scientific Publications

Stata

- Stata Youtube Channel - Videos on usage
- UCLA IDRE (Institute for Digital Research and Education)
- Survey Design and Analysis Services: Tips on graphics
- Stata cheat sheets

Books

- Kirkwood and Sterne (2003)
- Peacock, Kerry, and Balise (2017)

# References

Kirkwood, B., and J. Sterne. 2003. *Essential Medical Statistics*. Wiley.

Peacock, Janet L., Sally M. Kerry, and Raymond R. Balise. 2017. *Presenting Medical Statistics from Proposal to Publication*. Oxford, UK: Oxford University Press. https://oxfordmedicine.com/view/10.1093/med/9780198779100.001.0001/med-9780198779100.