

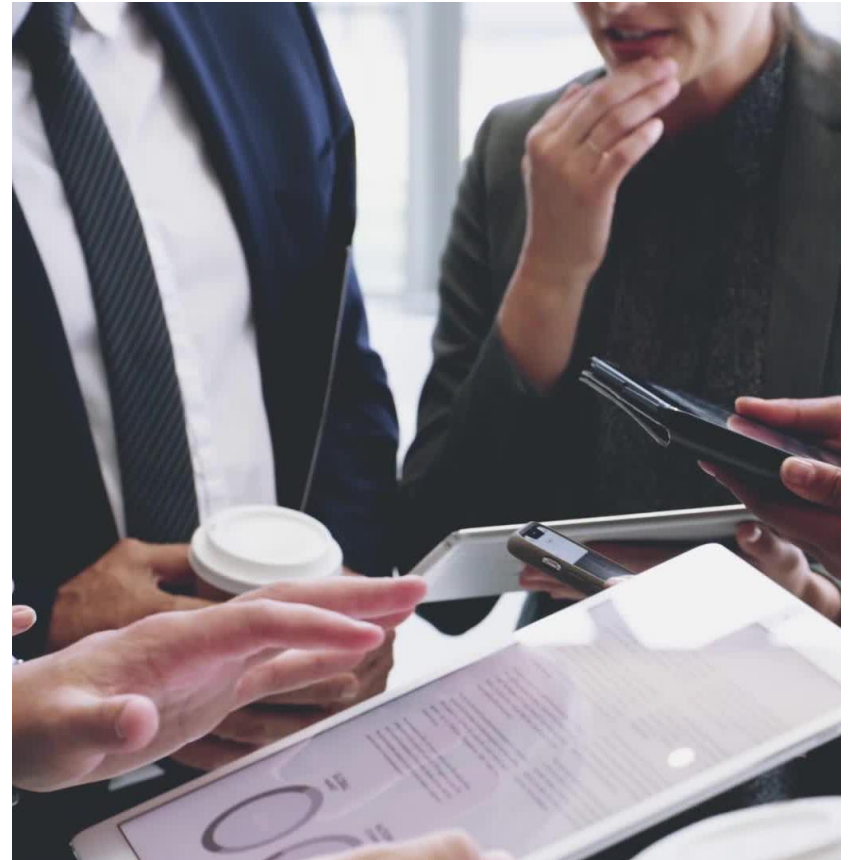
Session 8

Data Cleaning

Prepared by Associate prof. Prum Virak

What should be done for data examination?

- Before starting to do data analysis, it is absolutely essential that you have to examine the raw data first in order to find some types of error are included;
- **Illegal code:** values are not specified in code instruction
- **Omissions:** Do not follow correctly the SKIP instruction
- **Logical consistencies:** current age is less than marriage age
- **Improbabilities:** a woman aged 25 years old with 10 living children.



Where should you start

- First of all, take a look at all the variables you have identified for analysis, and produce simple tabulations for categorical variable or summary statistics for quantitative variable in order to search
Primarily for some errors.



Tabulating Categorical Variables

- Look at the distribution of categorical variable using the tabulate command.

```
. tab n201
```

Away from job	Freq.	Percent	Cum.
1	45	8.21	8.21
2	242	44.16	52.37
3	260	47.45	99.82
23	1	0.18	100.00
Total	548	100.00	

Tabulating Categorical Variables

- To obtain a two-way table by using command tab with the two variables.

```
. tab q414 q415, col mis
```

Key				
<i>frequency</i> <i>column percentage</i>				
Do your children usually wash their hand?	If yes, what do your children usually use in was..			Total
	water onl	water and	.	
Never	0 0.00	0 0.00	98 100.00	98 18.39
Rarely	74 30.71	12 6.19	0 0.00	86 16.14
Sometime	114 47.30	41 21.13	0 0.00	155 29.08
often	37 15.35	101 52.06	0 0.00	138 25.89
Always	16 6.64	40 20.62	0 0.00	56 10.51
Total	241 100.00	194 100.00	98 100.00	533 100.00

. sum n007

Variable	Obs	Mean	Std. Dev.	Min	Max
n007	550	49.00182	28.24454	4	589

Summarizing Quantitative

- For Quantitative variable, it is more efficient to look at summary measures (means, median and standard deviations and well as maximum and minimum values) rather than producing a table for each individual value of the variable. In doing that it can help you to identify incorrect values and outliers. E.g. I want to obtain a summary of distribution of weight (n007) of respondents
 - **sum n007**

Summarizing Quantitative (cont.)

- Stata can provide additional information by using the detail option with the summarize command. Stata would provide you the number observations, the mean, standard deviation, variance, ...etc for variable age. Ex, sum n007 (weight variable), detail

```
. sum n007, detail
```

weight				
Percentiles		Smallest		
1%	30	4		
5%	35	4		
10%	38	16	obs	550
25%	42	20	Sum of wgt.	550
50%	47	Largest	Mean	49.00182
75%	52		Std. Dev.	28.24454
90%	57		Variance	797.7541
95%	63		Skewness	15.22052
99%	80	589	Kurtosis	270.8839

Cross checking variables for error

- As well as tabulating and summarizing data to check for identify errors also cross check related variables.
 - Age at first sex cannot be older than current age. An easy way to check this is to create a check variable:
 - `gen checkage=currentage – agefirstsex`
 - `tab checkage...` if there are negative values, it means the respondent had reported an age at first sex that is older then his or her age now. These recodes need more investigation.
-



Correcting data error

- You should do some basic consistency checking before you think of starting analysis. E.g it is not possible for someone who has never had sex to report an age at first sex etc.
-

Practical exercise