香港中文大學
The Chinese University of Hong Kong

DSPS CUHK
數據科學及政策研究
DATA SCIENCE & POLICY STUDIES

CONFIDENCE INTERVAL

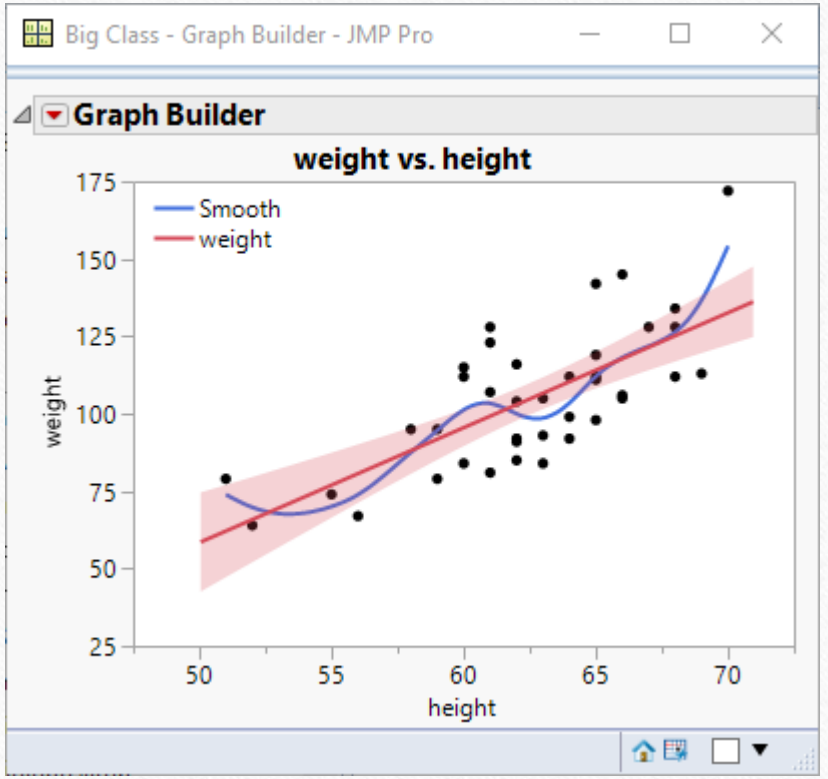
Data Science and Policy Studies Programme

E-Learning Space of Data Science for Public Policy

Supported by:
CUHK Courseware Development Grant Scheme (2019-22)

Agenda

1. Statistical Inference
2. Confidence Interval
3. Application



1. Statistical Inference

- **Statistical inference**: inferring value of **population parameter**
  - **Parameter**: number that describes a population distribution (e.g. mean, standard deviation)
- How do we make the statistical inference? We make the inference based on a number computed from the sample data. The number is called a **statistic** or a **sample statistic**.

- A random variable that is used to estimate the population parameter is called "**estimator**."
  - If a single number is given as the estimate, it is called a **point estimate**
  - The other way to report an estimate is to give an interval of values in which the population parameter is claimed to fall. This estimate is called an **interval estimate**

2. Confidence Interval

- Point estimate for the **population mean** μ: sample mean $\bar{x}$
- The interval estimate $L \leq \mu \leq U$ is called the $100(1-\alpha)\%$ **confidence interval** of the parameter μ.
- The $100(1-\alpha)\%$ confidence interval for the population proportion P is:

$$p - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}} \leq P \leq p + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}$$

3. Application

- Hong Kong Visitor Arrivals sample data
  - Source: The Hong Kong Tourism Board https://partnernet.hktb.com/en/research_statistics/research_publications/index.html
  - **total_va**
    - Monthly Total Visitor Arrivals in Hong Kong
  - **total_va_mainland_china**
    - Monthly Total Visitor Arrivals in Hong Kong from Mainland China
  - Time period covered: From Jan 2016 to Dec 2020

3. Application

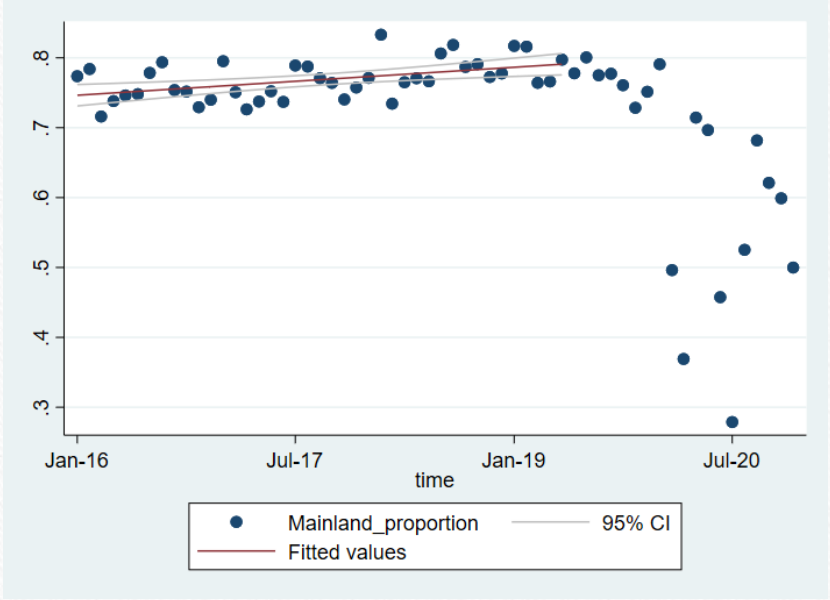
```
. import excel "F:\Users\admin\Desktop\CUHK (DSPS)\CUHK tourist data\Courseware grant\hk_visitors_sampl
> edata.xlsx", sheet("data") firstrow clear
(6 vars, 60 obs)

.
. tsset time
        time variable:  time, Jan-16 to Dec-20, but with gaps
                delta:  1 day
. gen period = "1. Normal_times" if inrange(time, td(1jan2016), td(31may2019))
(19 missing values generated)

. replace period = "2. Ebill_times" if inrange(time, td(1jun2019), td(31jan2020))
(8 real changes made)

. replace period = "3. COVID-19 times" if inrange(time, td(1feb2020), td(31dec2020))
variable period was str15 now str17
(11 real changes made)
```

3. Application

```
. * Confidence interval
. gen Mainland_proportion = total_va_mainland_china/total_va

. ci means Mainland_proportion if period=="1. Normal_times"

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
Mainland_p~n |         41     .768663    .0043398        .7598919     .777434

. ci means Mainland_proportion if period=="2. Ebill_times"

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
Mainland_p~n |          8    .7702718    .0080859        .7511516     .789392

. ci means Mainland_proportion if period=="3. COVID-19 times"

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
Mainland_p~n |         11    .5399698    .0418968        .4466179    .6333217
```

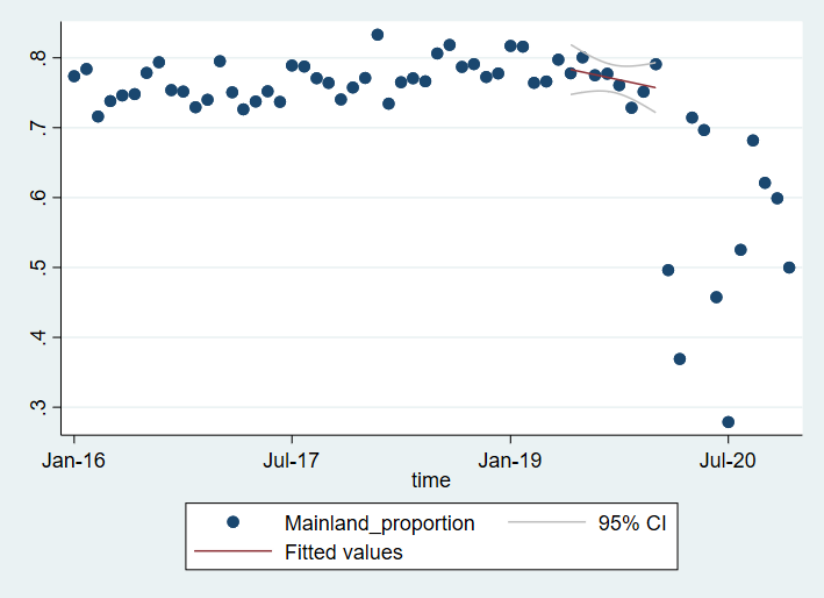
3. Application

```
. twoway (scatter Mainland_proportion time) (lfitci Mainland_proportion time if period=="1. Normal_time
> s", ciplot(rline)), ytitle("") name(graph3, replace)
```



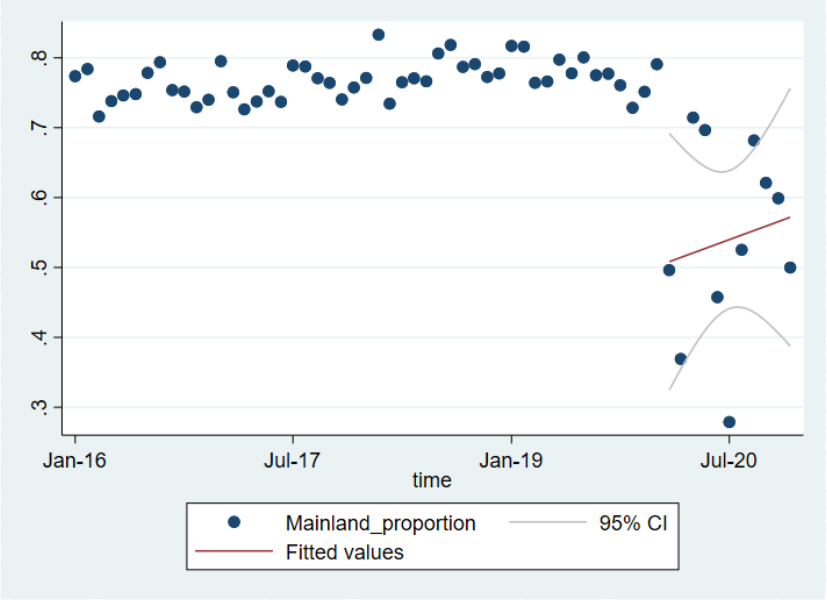
3. Application

```
. twoway (scatter Mainland_proportion time) (lfitci Mainland_proportion time if period=="2. Ebill_times
> ", ciplot(rline)), ytitle("") name(graph4, replace)
```



3. Application

```
. twoway (scatter Mainland_proportion time) (lfitci Mainland_proportion time if period=="3. COVID-19 ti
> mes", ciplot(rline)), ytitle("") name(graph5, replace)
```



4. Policy Implications

- COVID-19 is the most significant reason for bringing the tourist industry to a standstill. The impact on Mainland tourists' arrivals has been particularly large.
- Mainland inbound tourism will likely remain subdued in the near term, but may begin to recover later when vaccination programme in Hong Kong yield the intended results.
- After the COVID-19 situations are put under control, Mainland travelers' confidence could be restored by mass media campaigns.