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

HYPOTHESIS TESTING

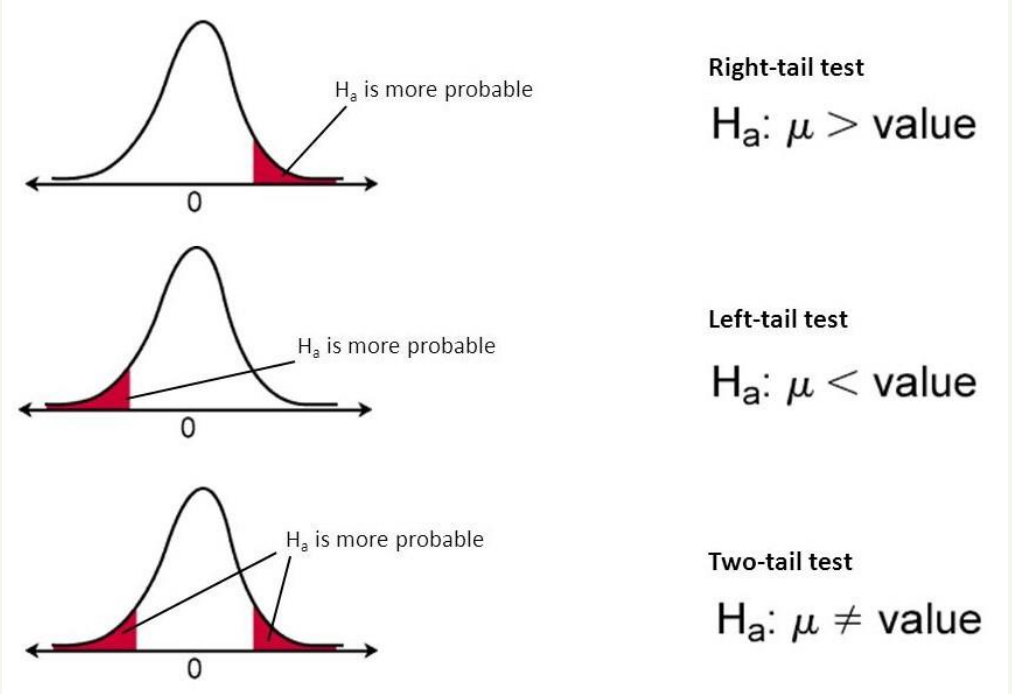
Data Science and Policy Studies Programme

E-Learning Space of Data Science for Public Policy

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

Agenda

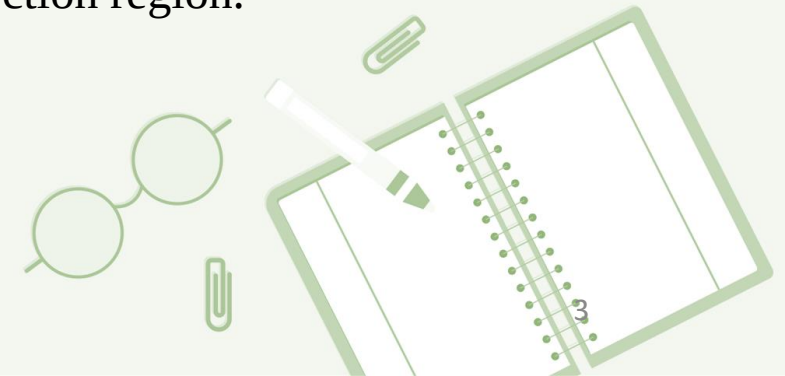
1. Hypothesis testing
2. Testing difference in mean
3. Application



1. Hypothesis Testing

- There are many problems where we need to decide whether to accept or to reject a statement about some population parameters.
  - The statement is called a ***hypothesis***, and the decision-making procedure regarding the hypothesis is called ***hypothesis testing***.

- Test Procedure
  (a) Set the significance level α, if not already given.
  (b) State the appropriate hypotheses, null and alternative.
  (c) Compute the appropriate test statistic based on sample information.
  (d) Sketch the acceptance and rejection regions.
  (e) Examine whether the calculated test statistic falls in the acceptance or rejection region.
  (f) Make suitable conclusion.



2. Testing difference in mean

- Assume that two normal populations have equal variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
- Estimate the population variance σ^2 by the pooled sample variance s_p^2

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- The statistic $\boldsymbol{t = \frac{(x_1 - x_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}$ is a ***t* distribution with $(\boldsymbol{n_1 + n_2 - 2})$ degrees of freedom.**

3. Application

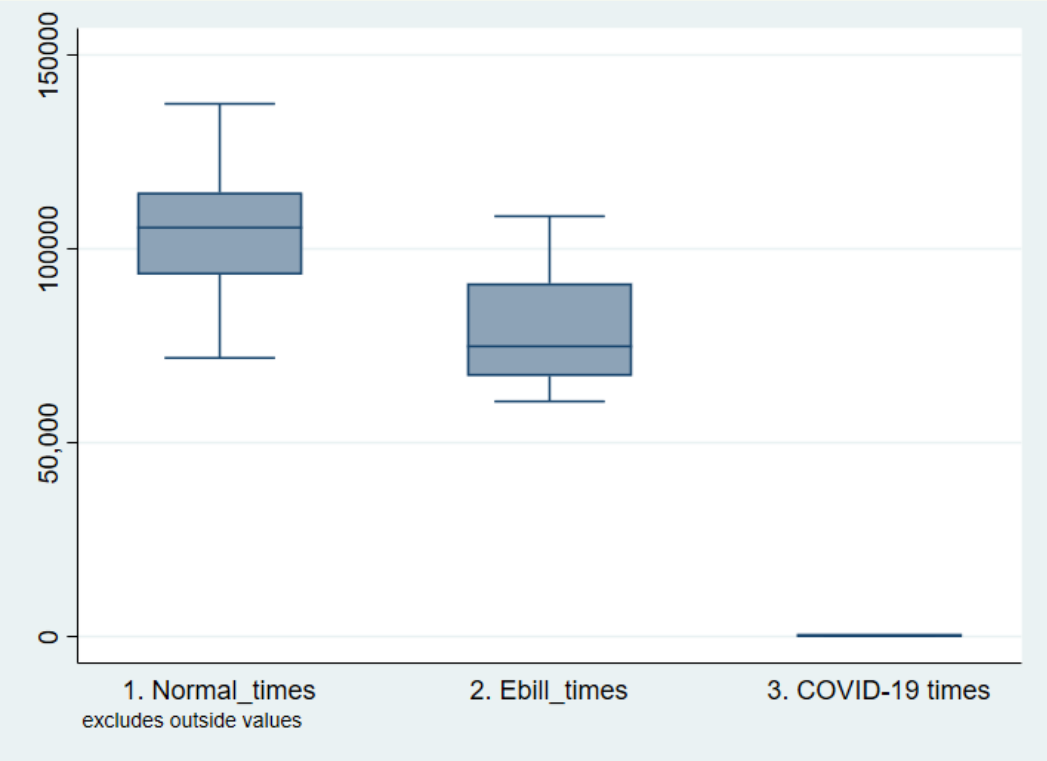
- HK Visitor Arrivals sample data
  - Source: The Hong Kong Tourism Board.
  - **total_va**
    - Monthly Total Visitor Arrivals in Hong Kong
  - **total_va_usa**
    - Monthly Total Visitor Arrivals in Hong Kong from the USA

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

```
. graph box total_va_usa, over(period) noout ytitle("") name(graph4, replace)
```

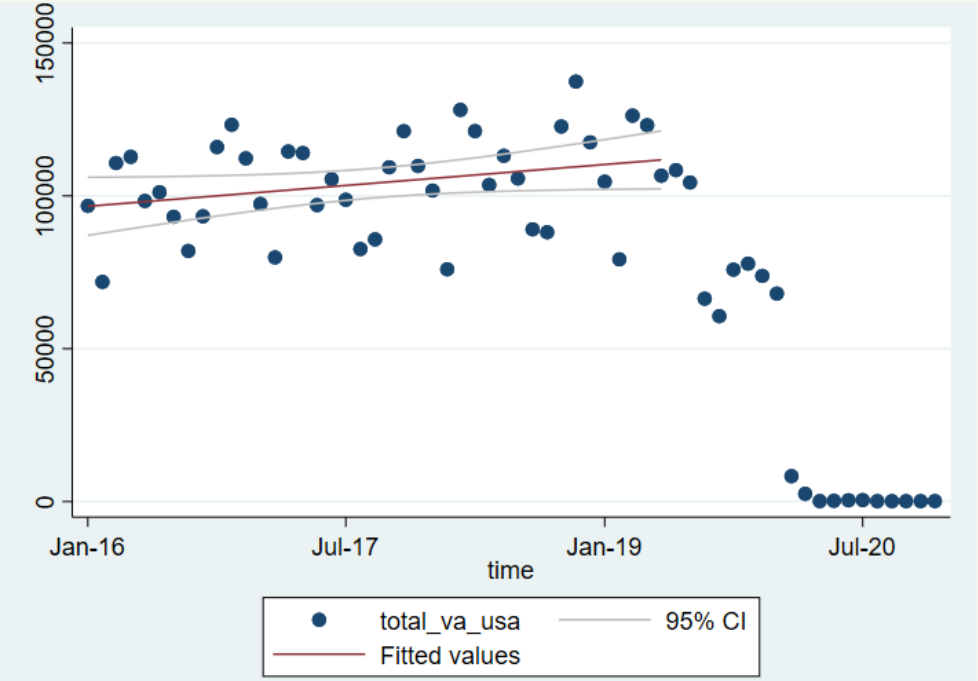


3. Application

```
. gen USA_va1 = total_va_usa if period=="1. Normal_times"
(19 missing values generated)

. gen USA_va2 = total_va_usa if period=="2. Ebill_times"
(52 missing values generated)

. gen USA_va3 = total_va_usa if period=="3. COVID-19 times"
(49 missing values generated)

. ttest USA_va1 = USA_va2, unpaired

Two-sample t test with equal variances
```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
USA_va1	41	104177.3	2465.574	15787.38	99194.18	109160.4
USA_va2	8	79415.13	6214.686	17577.79	64719.73	94110.52
combined	49	100134.5	2627.445	18392.11	94851.66	105417.3
diff		24762.17	6209.94		12269.38	37254.95

```
    diff = mean(USA_va1) - mean(USA_va2)                      t =   3.9875
Ho: diff = 0                                     degrees of freedom =       47

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9999         Pr(|T| > |t|) = 0.0002          Pr(T > t) = 0.0001
```



(diff – 0)/ standard error of diff
= {[mean(USA_va1)-mean(USA_va2)] – 0} / standard error of diff
= (24762.17 – 0) / 6209.94 = 3.9875

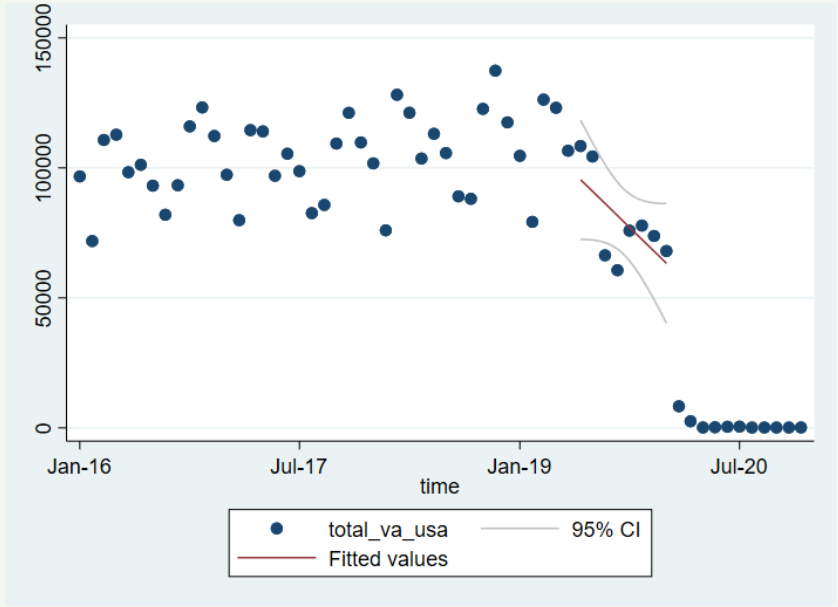
```
. twoway (scatter total_va_usa time) (lfitci total_va_usa time if period=="1. Normal_times", ciplot(rli
> ne)), ytitle("") name(graph6, replace)
```

3. Application

```
. ttest USA_va2 = USA_va3, unpaired

Two-sample t test with equal variances
```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
USA_va2	8	79415.13	6214.686	17577.79	64719.73	94110.52
USA_va3	11	1163.636	744.3671	2468.786	-494.9169	2822.19
combined	19	34111.63	9456.63	41220.49	14243.99	53979.27
diff		78251.49	5314.456		67038.97	89464.01

```
    diff = mean(USA_va2) - mean(USA_va3)                      t =  14.7243
Ho: diff = 0                                     degrees of freedom =       17

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```




```
. twoway (scatter total_va_usa time) (lfitci total_va_usa time if period=="2. Ebill_times", ciplot(rlin
> e)), ytitle("") name(graph7, replace)
```

4. Policy Implications

- COVID-19 is the most significant reason for bringing the tourist industry to a standstill. The impact on U.S. tourist has been large.

- U.S. inbound tourism will likely remain subdued in the near term, but may begin to recover later when vaccination programme in Hong Kong further improves the current situation.

- After the COVID-19 situations are put under control, it is also suggested that tourism promotional campaigns might restore U.S. travelers' confidence.