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

DESCRIPTIVE STATISTICS

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

Data Science and Policy Studies
數據科學與政策研究
JUPAS Code: 4893

1. Measures of Central Tendency
2. Measures of Dispersion
3. Box Plot
4. Policy Implications

1. MEASURES OF CENTRAL TENDENCY

- For associated data,
  - Mean: $\bar{X} = \frac{\sum X_i}{N}$
    - where $\bar{X}$ is mean, X_i is score of i-th case, N is total number of cases.
  - Median: data arranged from low to high values,
    - If n is odd, median M_d is the value for the $(n+1)/2$-th observation.
    - If n is even, median M_d is the average of the $n/2$ and $(n+2)/2$-th observations.
  - Mode: the value which occurs most frequently.

- A warm-up example:

Ben	Susan	Peter
1 candy	2 candies	12 candies

- Mean = (1+2+12)/3 = 5 candies
- Median = 2 Candies

2. MEASURES OF DISPERSION

- For assorted data,
  - Range: Absolute difference between maximum value and minimum value.
  - Interquartile range:
    - Difference between location of the first quartile is $(\boldsymbol{n+1})/\boldsymbol{4}$ and the third quartile is $\boldsymbol{3}(\boldsymbol{n+1})/\boldsymbol{4}$.
  - Variance: average of squared deviations about mean, $\frac{\Sigma\{\boldsymbol{X_i}-\overline{\boldsymbol{X}}\}^{\boldsymbol{2}}}{\boldsymbol{n}}$
  - Standard deviation: square root of the variance

2. MEASURES OF DISPERSION

EXAMPLE 1

- {1,2,3,4,5,6,7,8,9,10}
- Range = 10-1 = 9
- Inter-quartile range:

  Step 1. Median is (5+6)/2 = 5.5

  Step 2. Q1 is 3 and Q3 is 8.

  Step 3. IQR = Q3-Q1 = 8-3 = 5

2. MEASURES OF DISPERSION

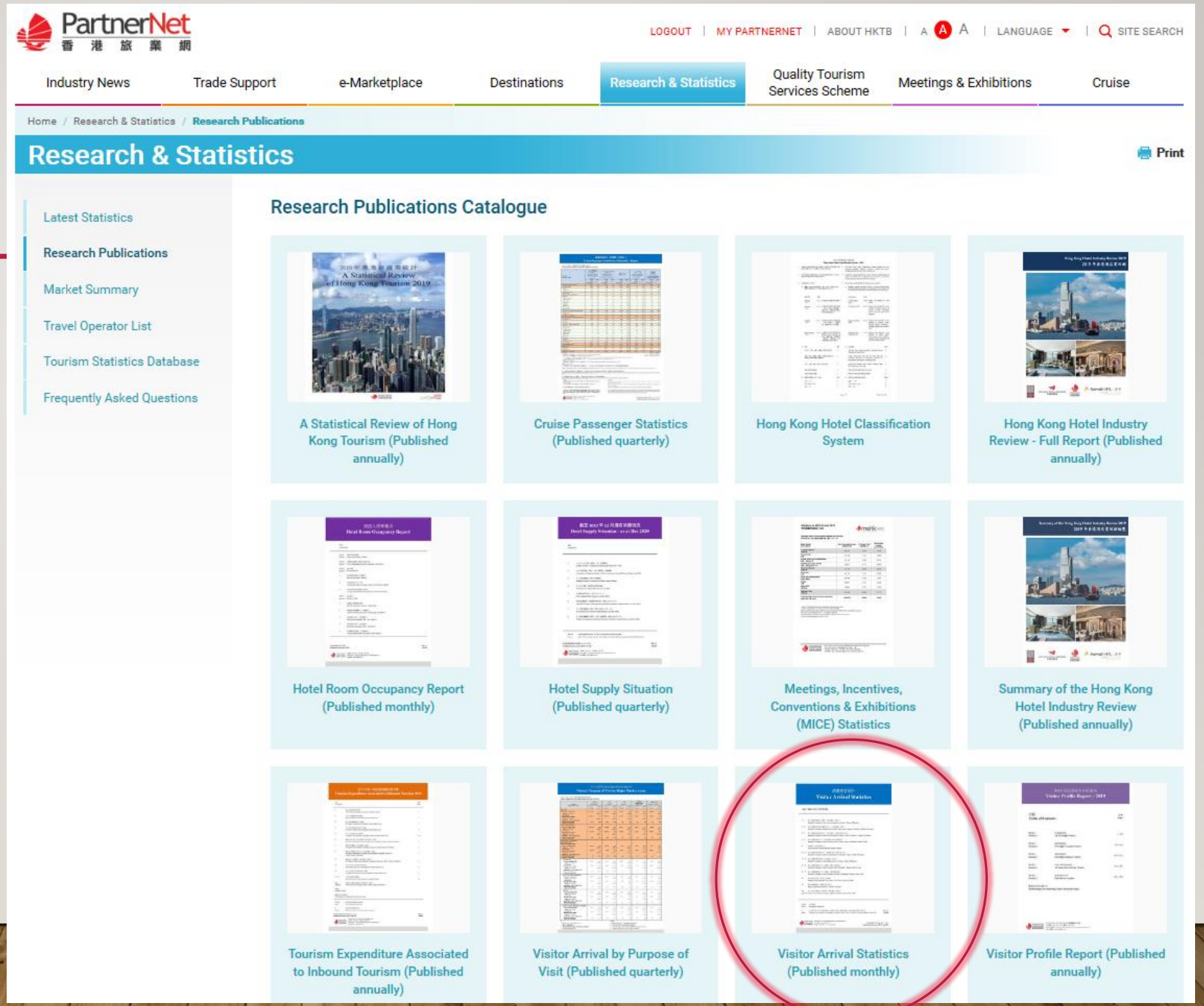
- EXAMPLE 2 – Finding variance and SD
- {1,2,3,4,5,6,7,8,9}
- Mean = (1+2+3+4+5+6+7+8+9)/9 = 45/9= 5
- Squared mean deviation =

Original number	1	2	3	4	5	6	7	8	9
Number - mean 5	-4	-3	-2	-1	0	1	2	3	4
Square of (number - mean)	16	9	4	1	0	1	4	9	16

- Sum of squared mean deviation = 16+9+4+1+0+1+4+9+16=60
- Variance = $\frac{\sum\{X_i - \bar{X}\}^2}{n}$, n=9
- Variance = 60/9 = 7.5
- SD = square root of variance = √7.5 = 2.74 (correct to 3 sig. fig.)

3. BOX PLOT

- Data source: Visitor Arrival Statistics (Published monthly) from Research Publications on PartnerNet by the Hong Kong Tourism board
- https://partnernet.hktb.com/en/research_statistics/research_publications/index.html



3. BOX PLOT

- Hong Kong Visitor Arrivals sample data
- **total_va**
  - Monthly Total Visitor Arrivals in Hong Kong
  - Time period covered: From Jan 2016 to Dec 2020

```
. 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

. tsline total_va, name(graph1, replace) title("{bf: Total visitor arrivals in HK, Jan 2016-Dec 2020}")
>  ytitle("") ttitle("")
```

3. BOX PLOT

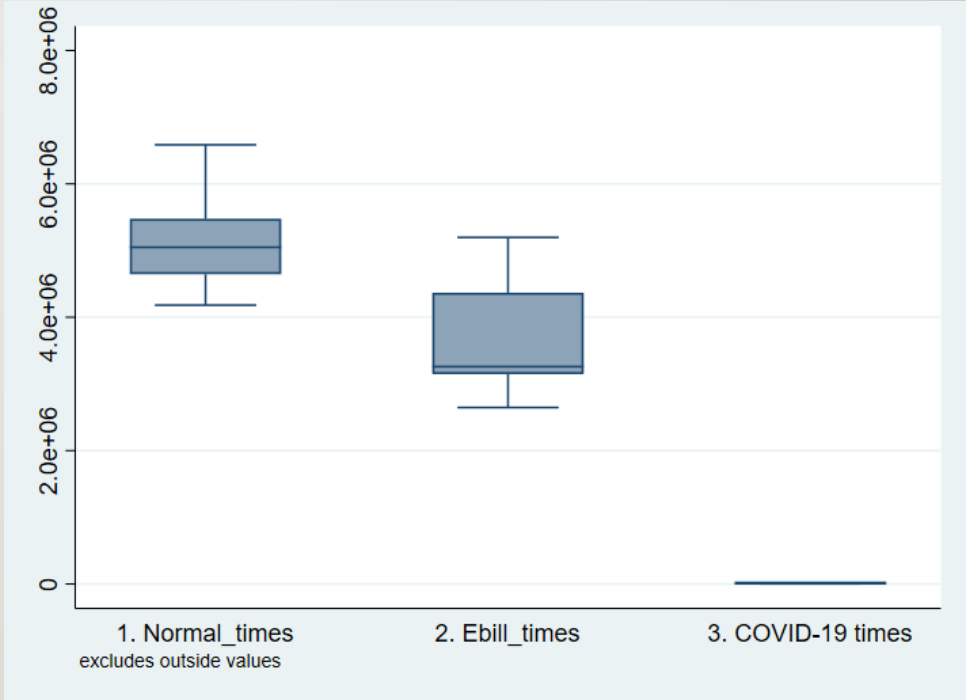
```
. * Descriptive statistics
. gen period = "1. Normal_times" if inrange(time, td(1jan2016), td(31may2019))
(19 missing values generated)

. 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)

.
. sum total_va if period=="1. Normal_times", d
```

total_va

	Percentiles	Smallest		
1%	4181417	4181417		
5%	4213801	4203256		
10%	4295731	4213801	Obs	41
25%	4646938	4285730	Sum of Wgt.	41
50%	5049022		Mean	5122018
		Largest	Std. Dev.	627132.2
75%	5475176	5916541		
90%	5895951	5995027	Variance	3.93e+11
95%	5995027	6586268	Skewness	.5834509
99%	6784406	6784406	Kurtosis	3.035397

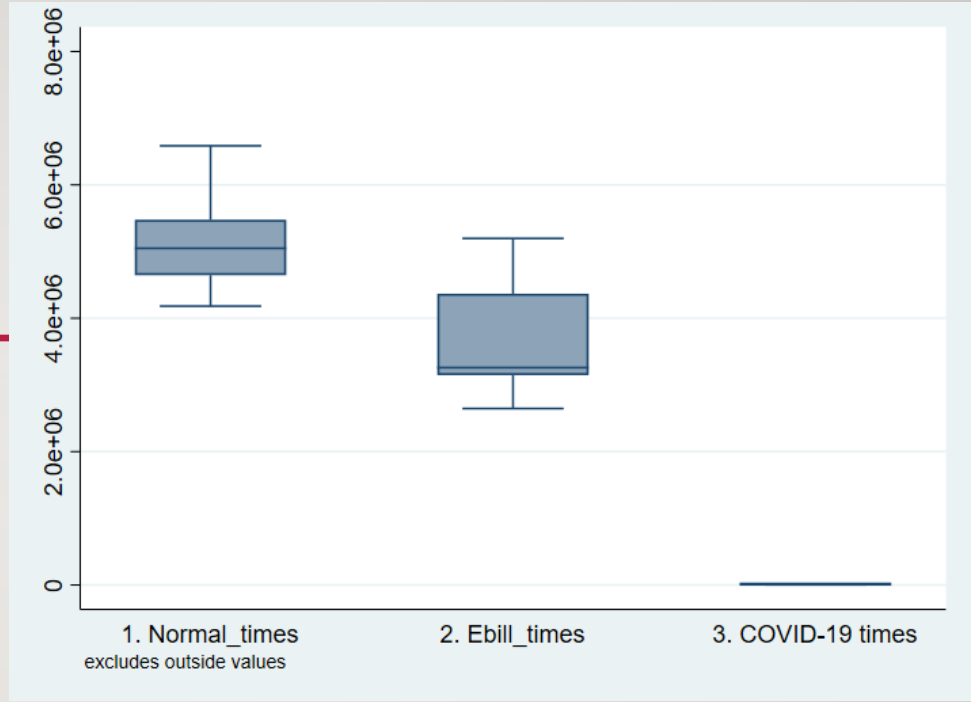
3. BOX PLOT

```
. sum total_va if period=="2. Ebill_times", d

                          total_va
-------------------------------------------------------------
      Percentiles      Smallest
 1%      2646127       2646127
 5%      2646127       3104049
10%      2646127       3191466       Obs                   8
25%      3147758       3207802       Sum of Wgt.           8

50%      3259687                     Mean           3674036
                       Largest       Std. Dev.      959867.1
75%      4367153       3311571
90%      5196969       3590571       Variance       9.21e+11
95%      5196969       5143734       Skewness       .8934143
99%      5196969       5196969       Kurtosis       2.189815

. sum total_va if period=="3. COVID-19 times", d

                          total_va
-------------------------------------------------------------
      Percentiles      Smallest
 1%         4125          4125
 5%         4125          4449
10%         4449          4867       Obs                  11
25%         4867          5962       Sum of Wgt.          11

50%         8139                     Mean           32824.82
                       Largest       Std. Dev.      59591.78
75%        20568         14606
90%        82285         20568       Variance       3.55e+09
95%       199123         82285       Skewness       2.270073
99%       199123        199123       Kurtosis         6.7553

. graph box total_va, over(period) noout ytitle("") name(graph2, replace)
```



4. POLICY IMPLICATIONS

- COVID-19 is the most significant reason for bringing the tourist industry to a standstill
- Inbound tourism will likely remain subdued in the near term, but may begin to recover later when the vaccination programme yield the intended results
- After the COVID-19 situations are put under control, suggested mass media campaign could restore travelers' confidence.



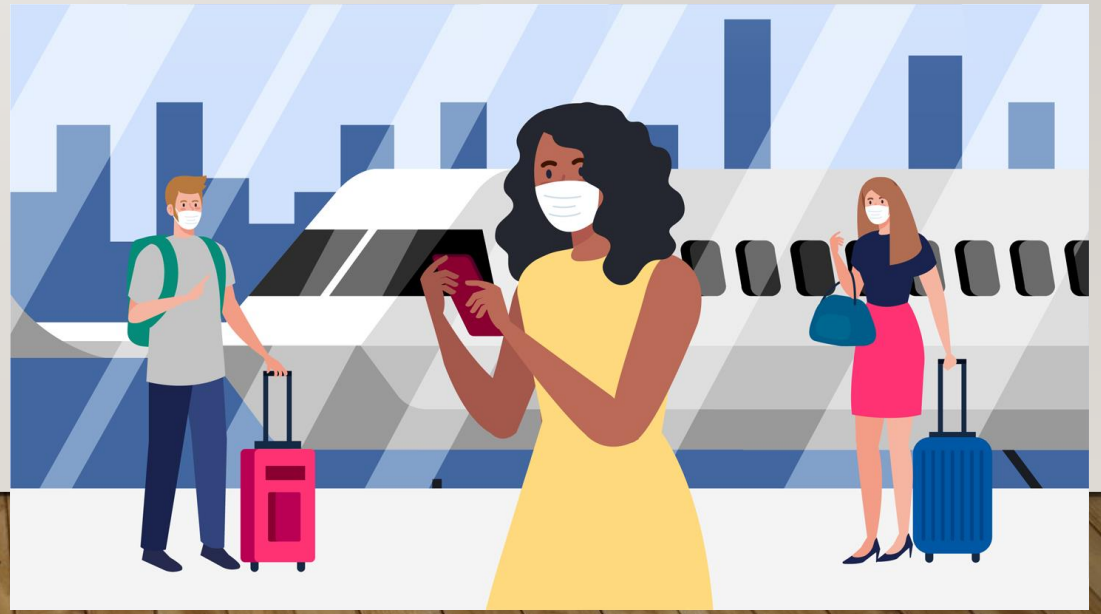


Image credit: https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.travelperk.com%2Fguides%2Fcorporate-travel-safety%2F&psig=AOvVaw3I_gdvxYU7t_dju2Ts9IEx&ust=1618386942624000&source=images&cd=vfe&ved=0CA0QjhxqFwoTCOjD6qnf-u8CFQAAAAAdAAAAABAT