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

CORRELATION

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. Correlation
2. Application

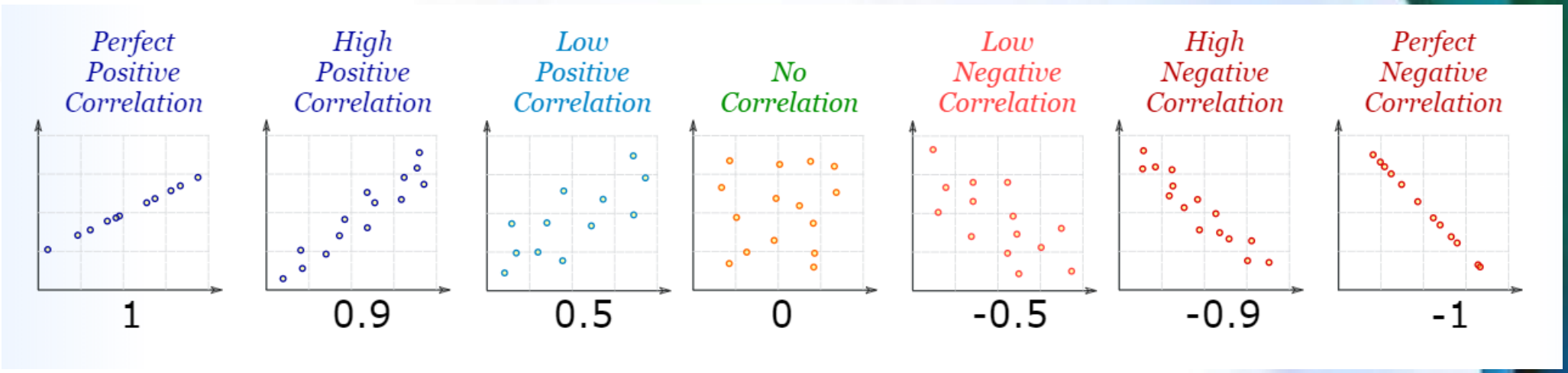
1. Correlation

- ***Correlation*** refers always to linear relationship.
- If X and Y are a pair of random variables with mean μ_x and μ_y and variances σ_x^2 and σ_y^2, population correlation coefficient

$$\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{E\{(X-\mu_x)(Y-\mu_y)\}}{\sqrt{E\{(X-\mu_x)^2\}E\{(Y-\mu_y)^2\}}}$$

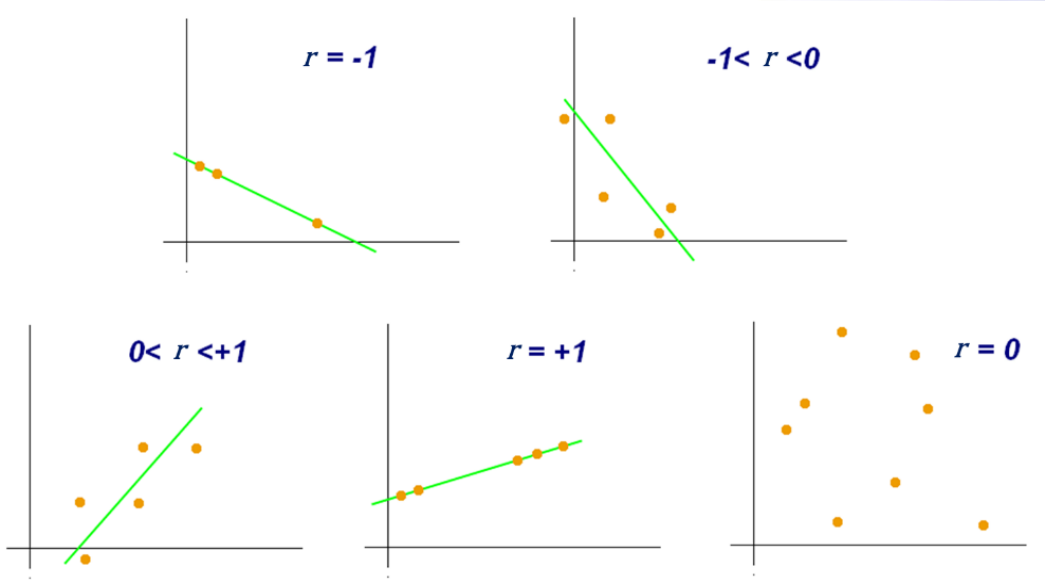
- **Sample correlation coefficient**

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \frac{\sum x_i y_i - n\,\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}}$$



1. Correlation

- Interpretation of ***correlation coefficient*** :
  - A *positive correlation* means that larger x are associated with larger y, and vice versa.
  - A *negative correlation* means that larger x are associated with smaller y, and vice versa.
  - $-1 \leq \rho,\ r \leq 1$



1. Correlation

- Test the hypothesis of no relationship between a pair of random variables in the population
  - Test if the population correlation coefficient is significantly different from zero:

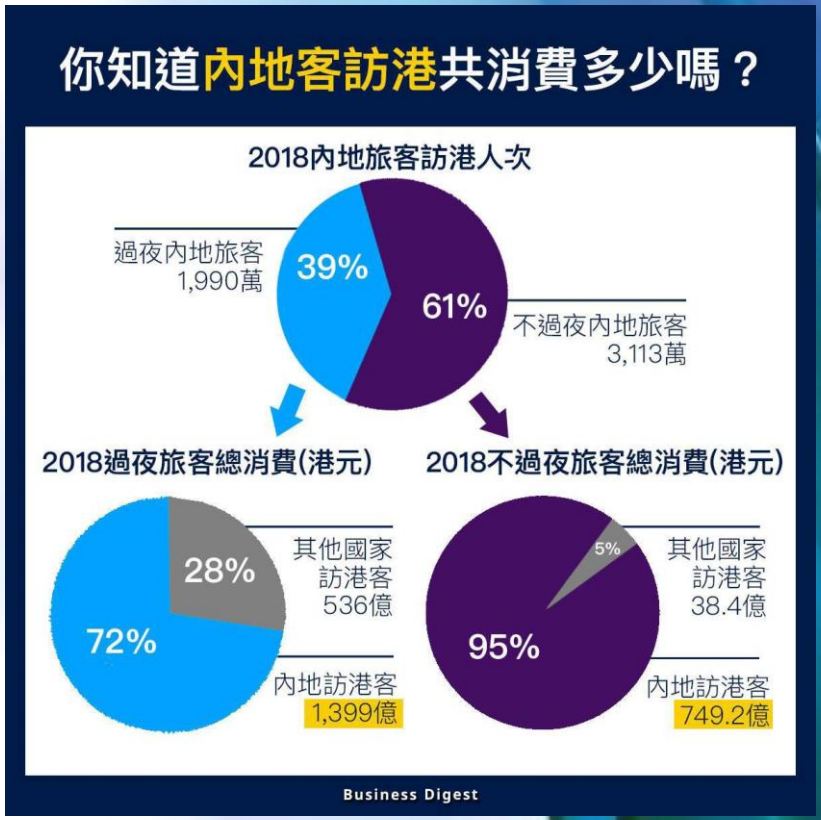
$$H_0: \rho = 0 \text{ vs. } H_a: \rho > 0; \rho < 0; \rho \neq 0$$

  - Test statistic with ***t* distribution with $n - 2$ degrees of freedom**:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

2. Application

- HK Visitor Arrivals sample data
  - Source: The Hong Kong Tourism Board.
  - **total_va_overnight**
    - Monthly Total Overnight Visitor Arrivals in Hong Kong
  - **total_va_sameday**
    - Monthly Total Sameday Visitor Arrivals in Hong Kong

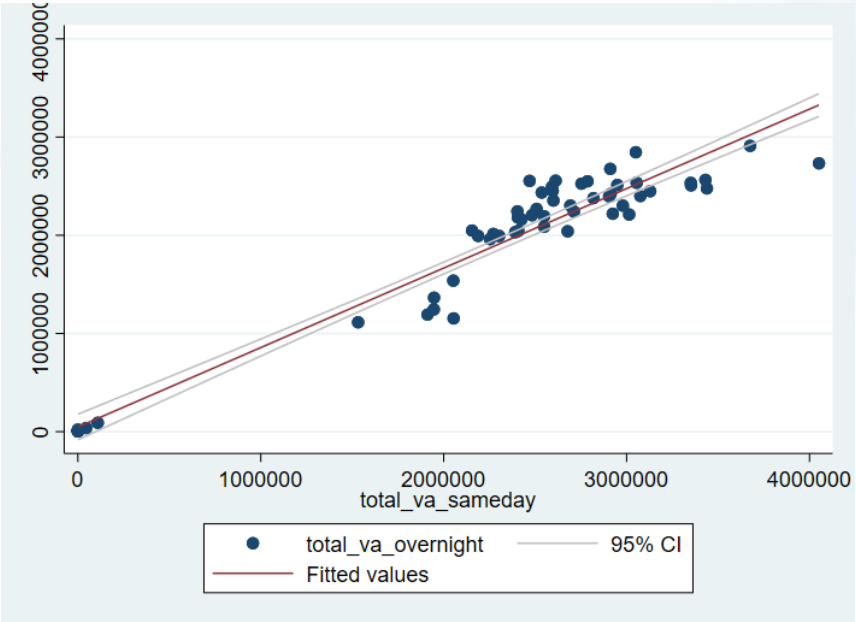


2. Application

```
. 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)
```

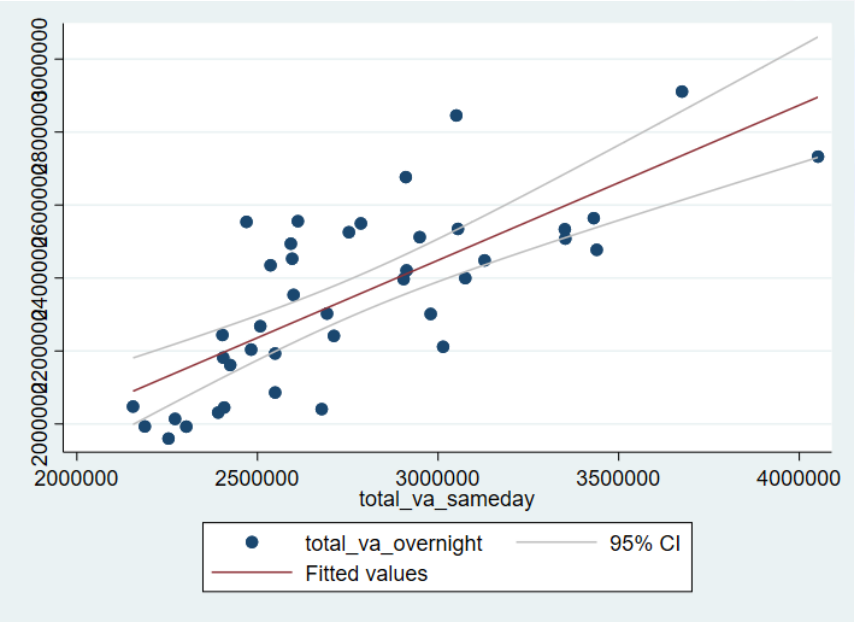
2. Application

```
. * Correlation
. pwcorr total_va_overnight total_va_sameday, star(.01)

             | total_~t total_~y
-------------+------------------
total_va_o~t |   1.0000
total_va_s~y |   0.9702*  1.0000

. twoway (scatter total_va_overnight total_va_sameday) (lfitci total_va_overnight total_va_sameday, cip
> lot(rline)), ytitle("") name(graph10, replace)
```



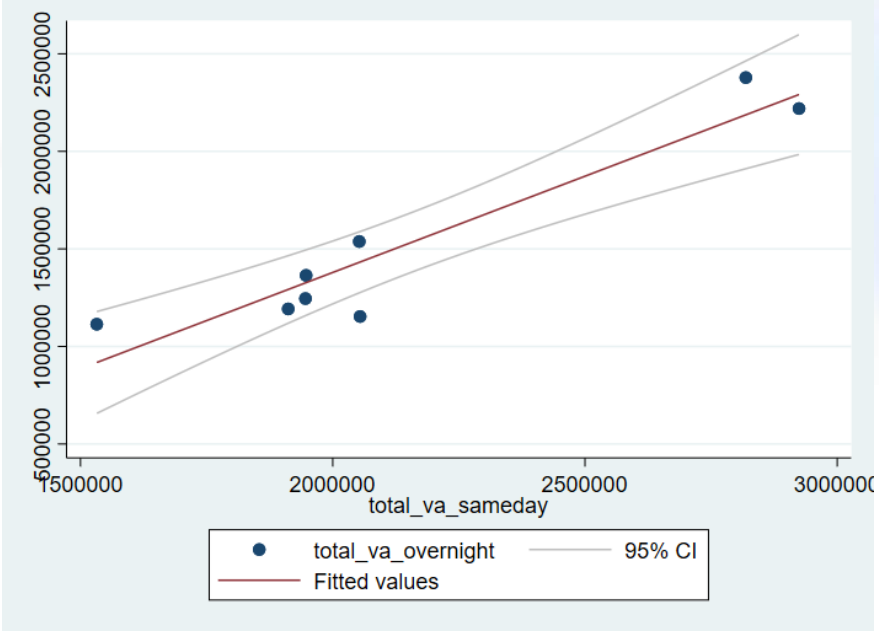
2. Application

```
. pwcorr total_va_overnight total_va_sameday if period=="1. Normal_times", star(.01)

             | total_~t total_~y
-------------+------------------
total_va_o~t |   1.0000
total_va_s~y |   0.7473*  1.0000

. twoway (scatter total_va_overnight total_va_sameday if period=="1. Normal_times") (lfitci total_va_ov
> ernight total_va_sameday if period=="1. Normal_times", ciplot(rline)), ytitle("") name(graph11, repla
> ce)
```



2. Application

```
. pwcorr total_va_overnight total_va_sameday if period=="2. Ebill_times", star(.01)

             | total_~t total_~y
-------------+------------------
total_va_o~t |   1.0000
total_va_s~y |   0.9444*  1.0000

. twoway (scatter total_va_overnight total_va_sameday if period=="2. Ebill_times") (lfitci total_va_ove
> rnight total_va_sameday if period=="2. Ebill_times", ciplot(rline)), ytitle("") name(graph12, replace
> )
```



2. Application

```
. pwcorr total_va_overnight total_va_sameday if period=="3. COVID-19 times", star(.01)

             | total_~t total_~y
-------------+------------------
total_va_o~t |   1.0000
total_va_s~y |   0.9794*  1.0000

. twoway (scatter total_va_overnight total_va_sameday if period=="3. COVID-19 times") (lfitci total_va_
> overnight total_va_sameday if period=="3. COVID-19 times", ciplot(rline)), ytitle("") name(graph13, r
> eplace)
```