Big Data Fundamentals and Applications

Statistical Analysis (II)

Descriptive Statistics – Graph

Asst. Prof. Chan, Chun-Hsiang

Master program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan
Undergraduate program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan
Undergraduate program in Applied Artificial Intelligence, Chung Yuan Christian University, Taoyuan, Taiwan

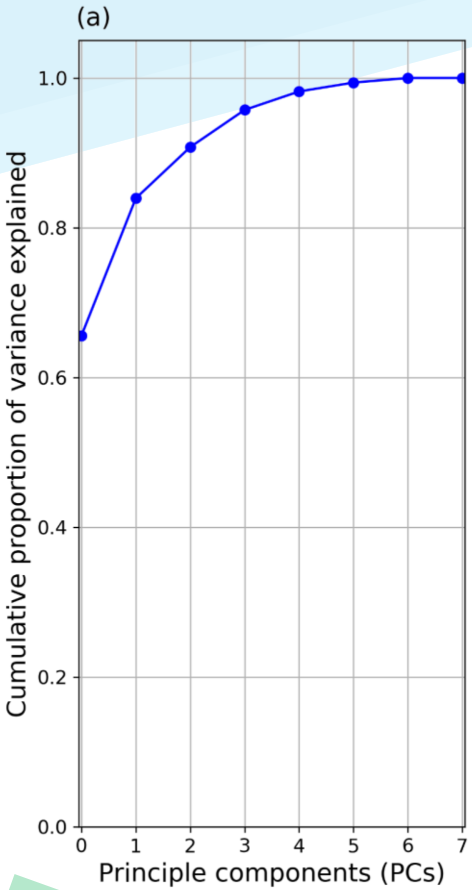
Outlines

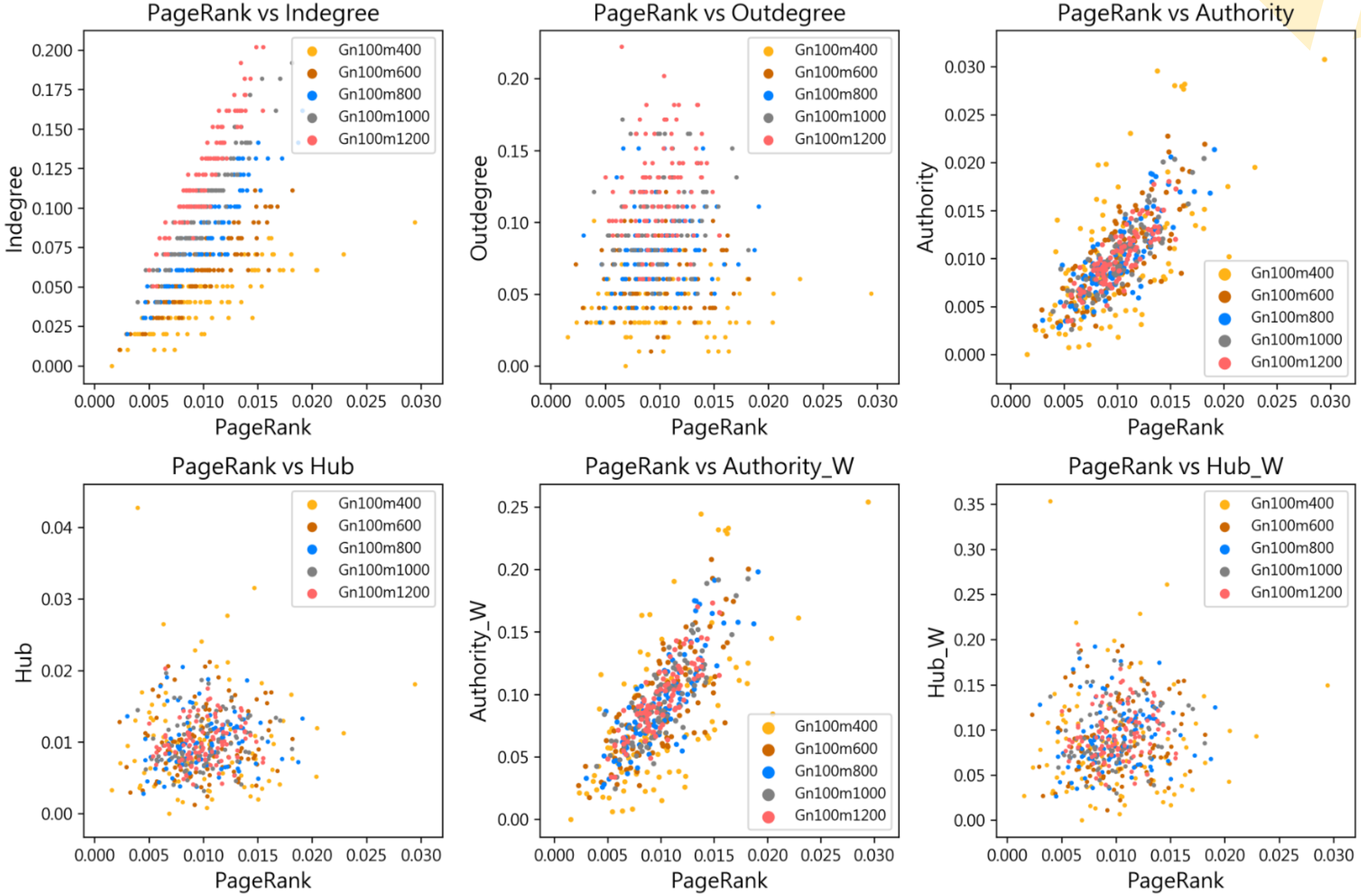
1. Introduction to Visualization
2. Line Plot & Scatter Plot
3. 1D Histogram
4. 2D Histogram
5. Area Plot
6. Stem Plot & Violin Plot
7. Box Plot
8. Stripe Plot & Swarm Plot
9. Error Bar Plot
10. Pie Chart
11. Bar Chart
12. Rose Plot & Radar Plot
13. Biplot & Control Chart
14. Forest Plot
15. Q-Q Plot
16. Combination Plot
17. Question Time

Visualization

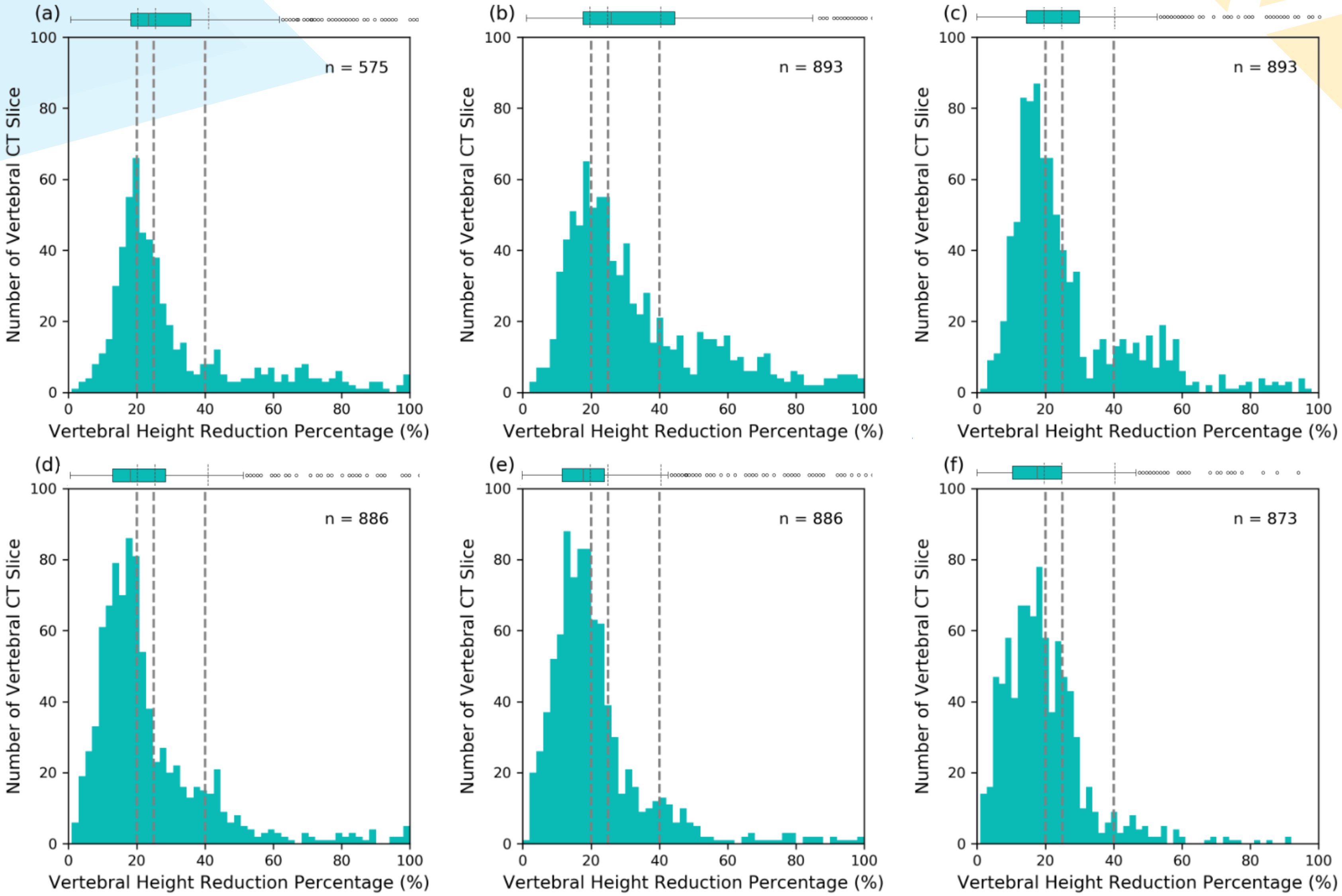
- As we mentioned before, data visualization is one of the most effective approach for data exploration to understand the characteristics of each feature or dataset.
- According to the attribute of data, we need to select the appropriate visualization methods for illustration.
- Here, we are going to introduce bar chart, biplot, box plot, control chart, forest plot, histogram, pie chart, Q-Q plot, scatter plot, stem plot, violin plot, radar plot, and line plot.

Line Plot & Scatter Plot

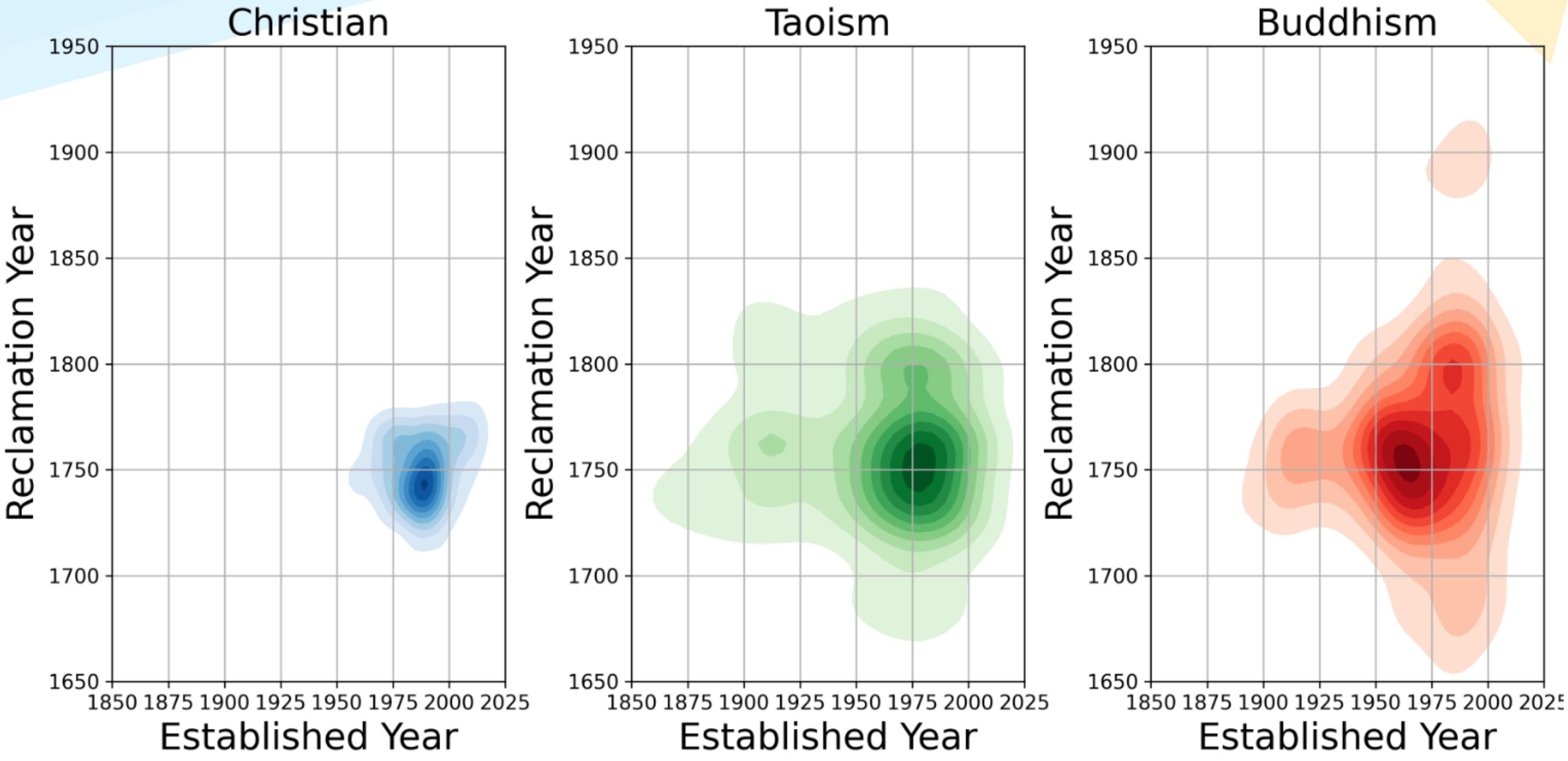




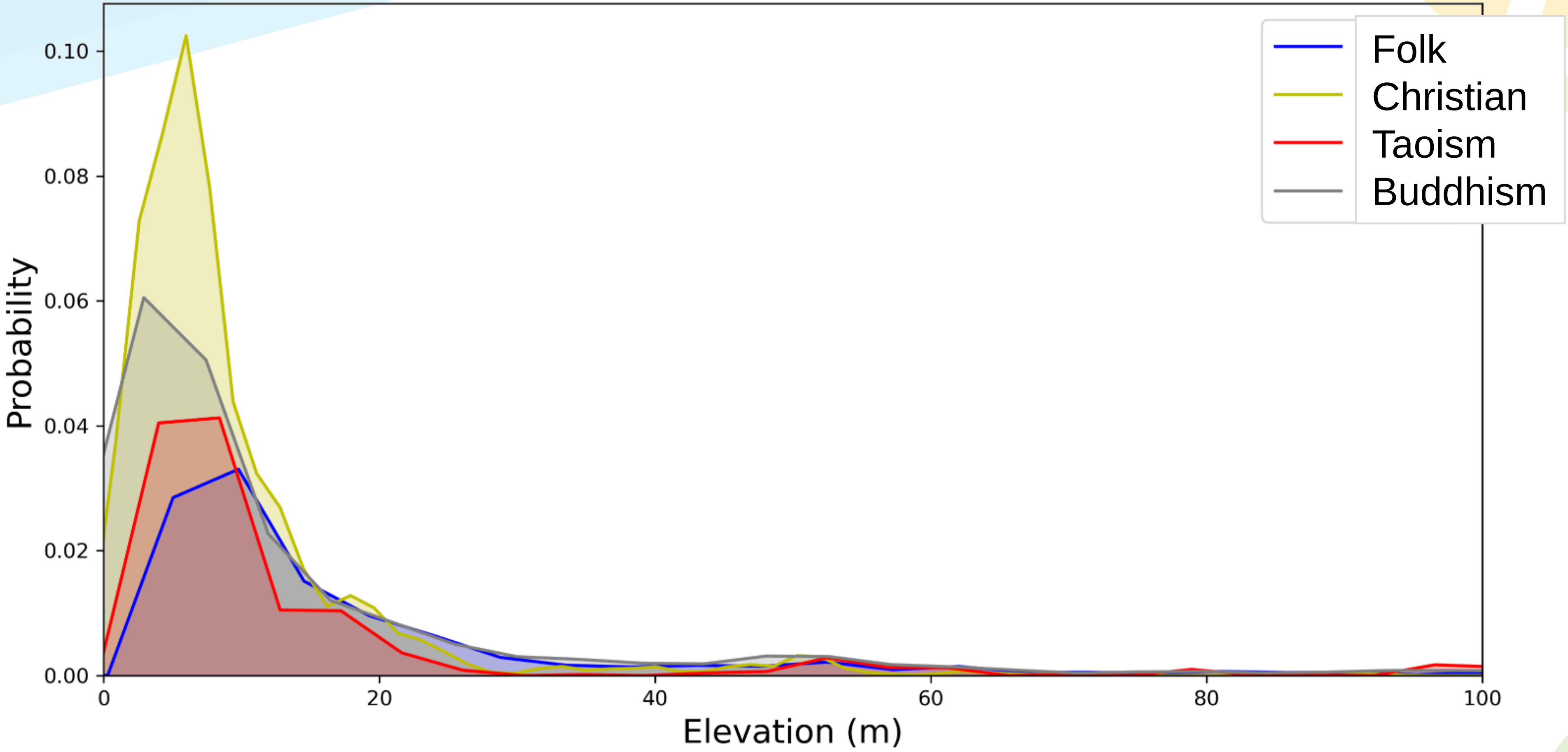
1D Histogram



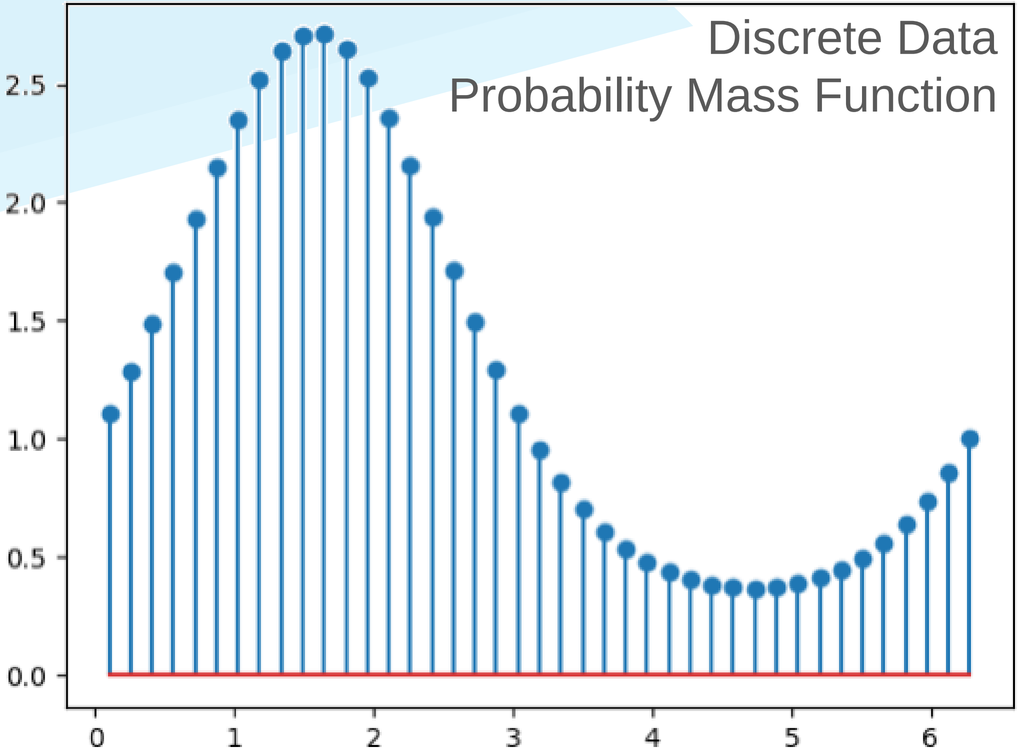
2D Histogram



Area Plot


Probability
0.00
0.02
0.04
0.06
0.08
0.10
0
20
40
60
80
100
Elevation (m)
Folk
Christian
Taoism
Buddhism

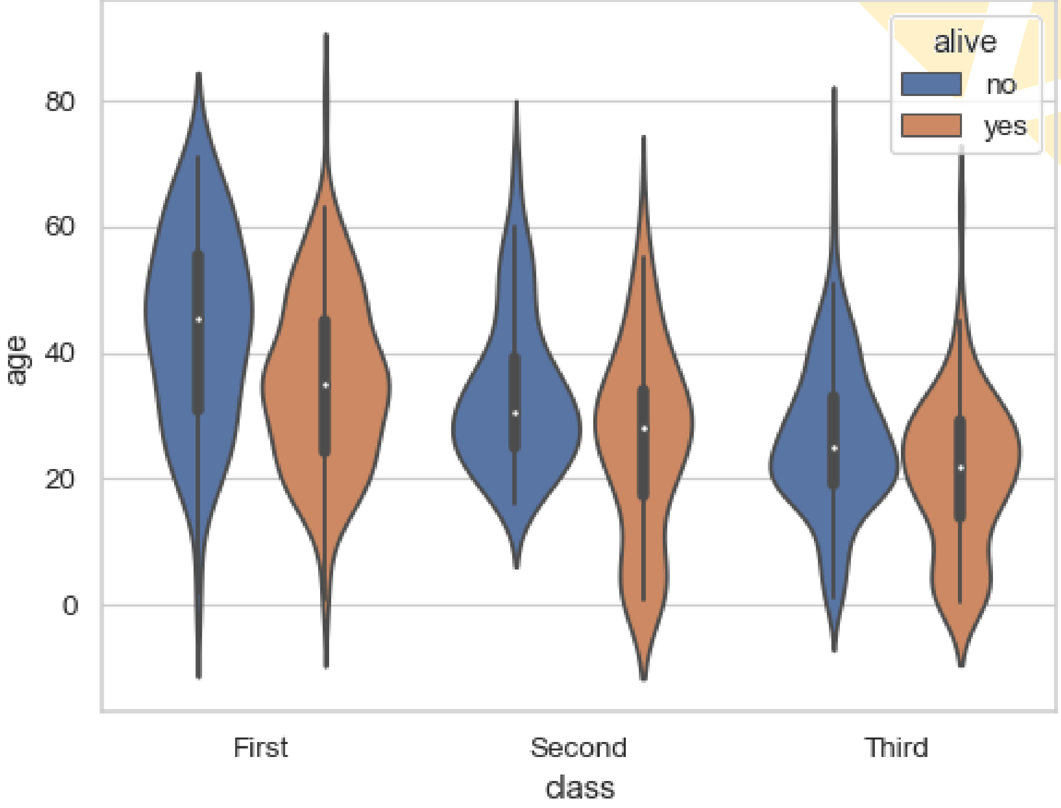
Stem Plot & Violin Plot



Discrete Data
Probability Mass Function

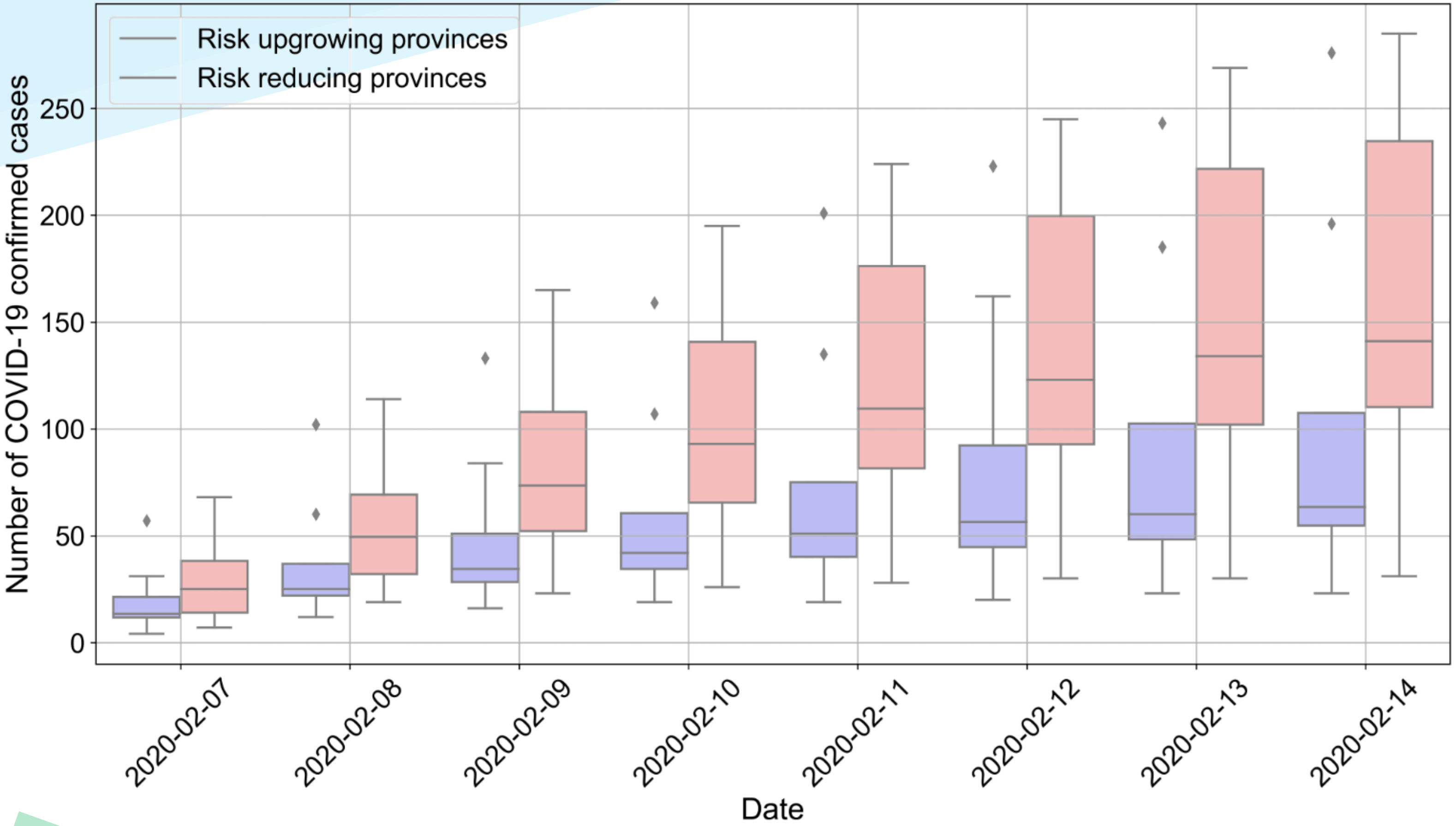
Source: https://matplotlib.org/stable/gallery/lines_bars_and_markers/stem_plot.html#sphx-glr-gallery-lines-bars-and-markers-stem-plot-py

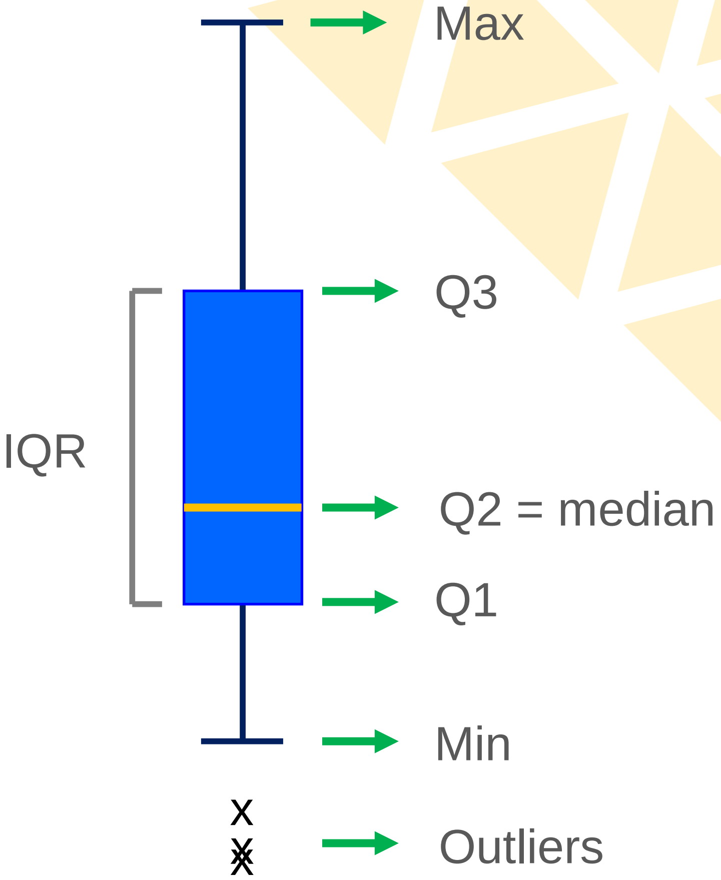
Continuous Data



Source: https://seaborn.pydata.org/generated/seaborn.violinplot.html

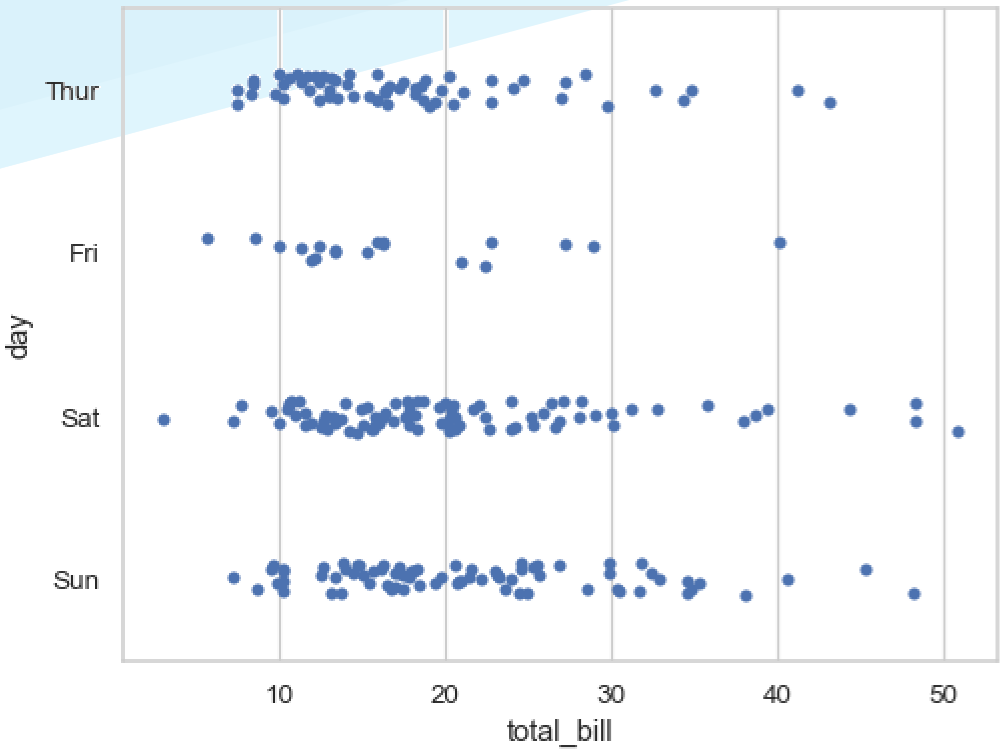
Box Plot


Risk upgrowing provinces
Risk reducing provinces
Number of COVID-19 confirmed cases
0
50
100
150
200
250
2020-02-07
2020-02-08
2020-02-09
2020-02-10
2020-02-11
2020-02-12
2020-02-13
2020-02-14
Date

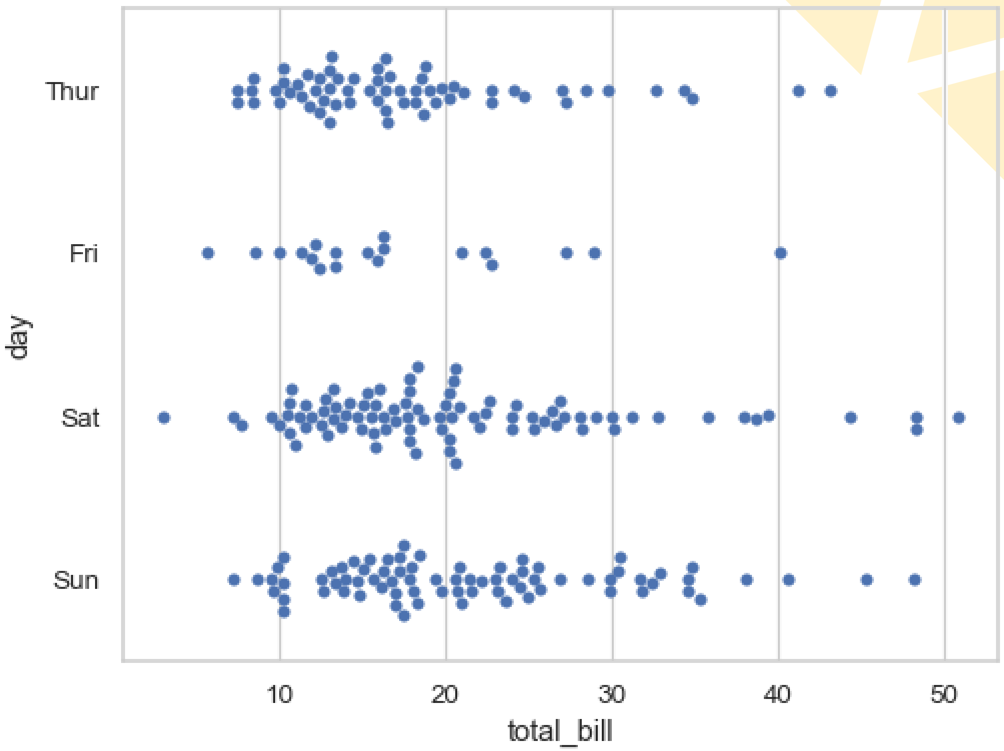

Max
Q3
IQR
Q2 = median
Q1
Min
X
X
Outliers

Strip Plot & Swarm Plot

Observe the d

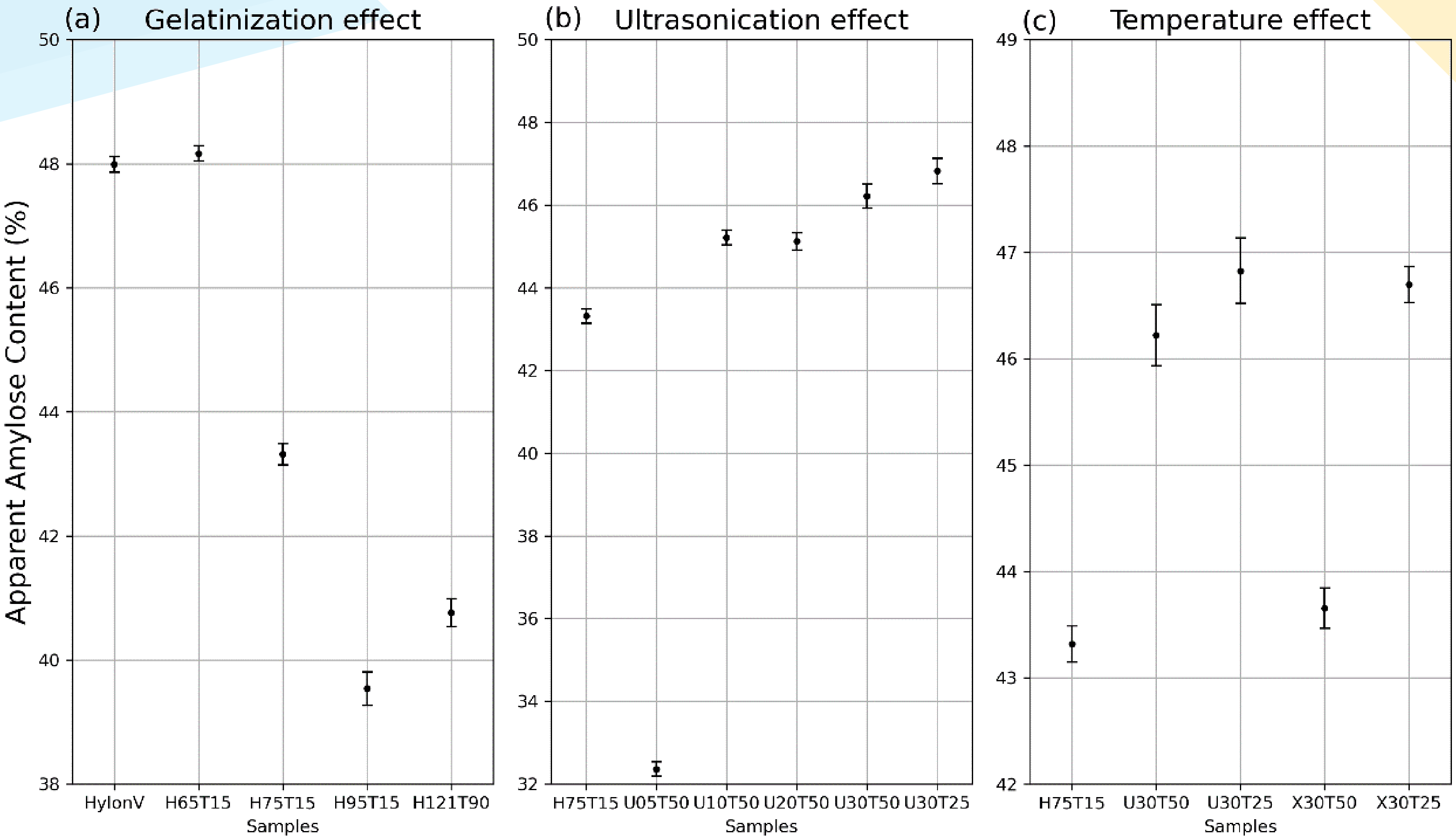


Source: https://seaborn.pydata.org/generated/seaborn.stripplot.html

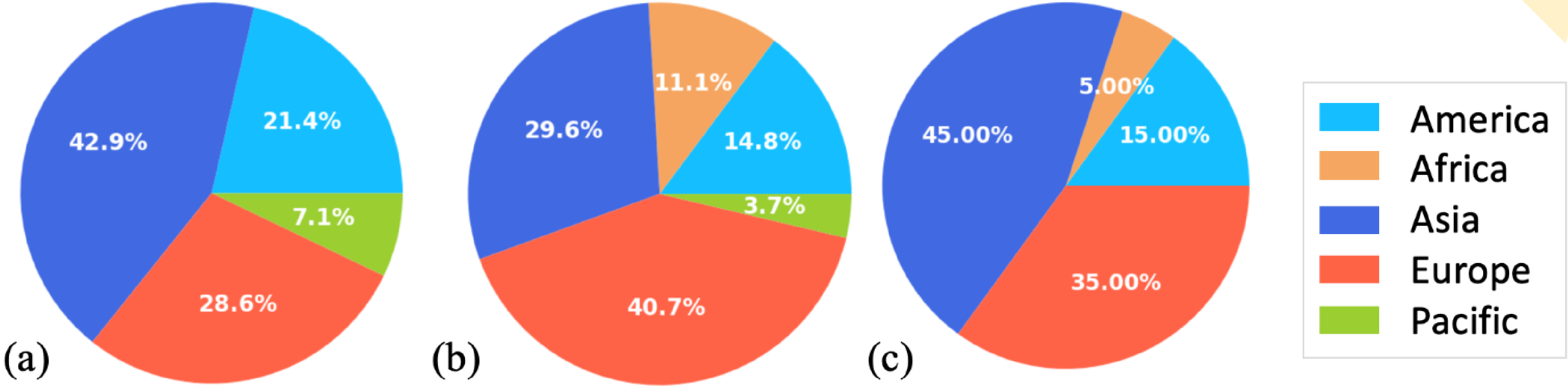


Source: https://seaborn.pydata.org/generated/seaborn.swarmplot.html

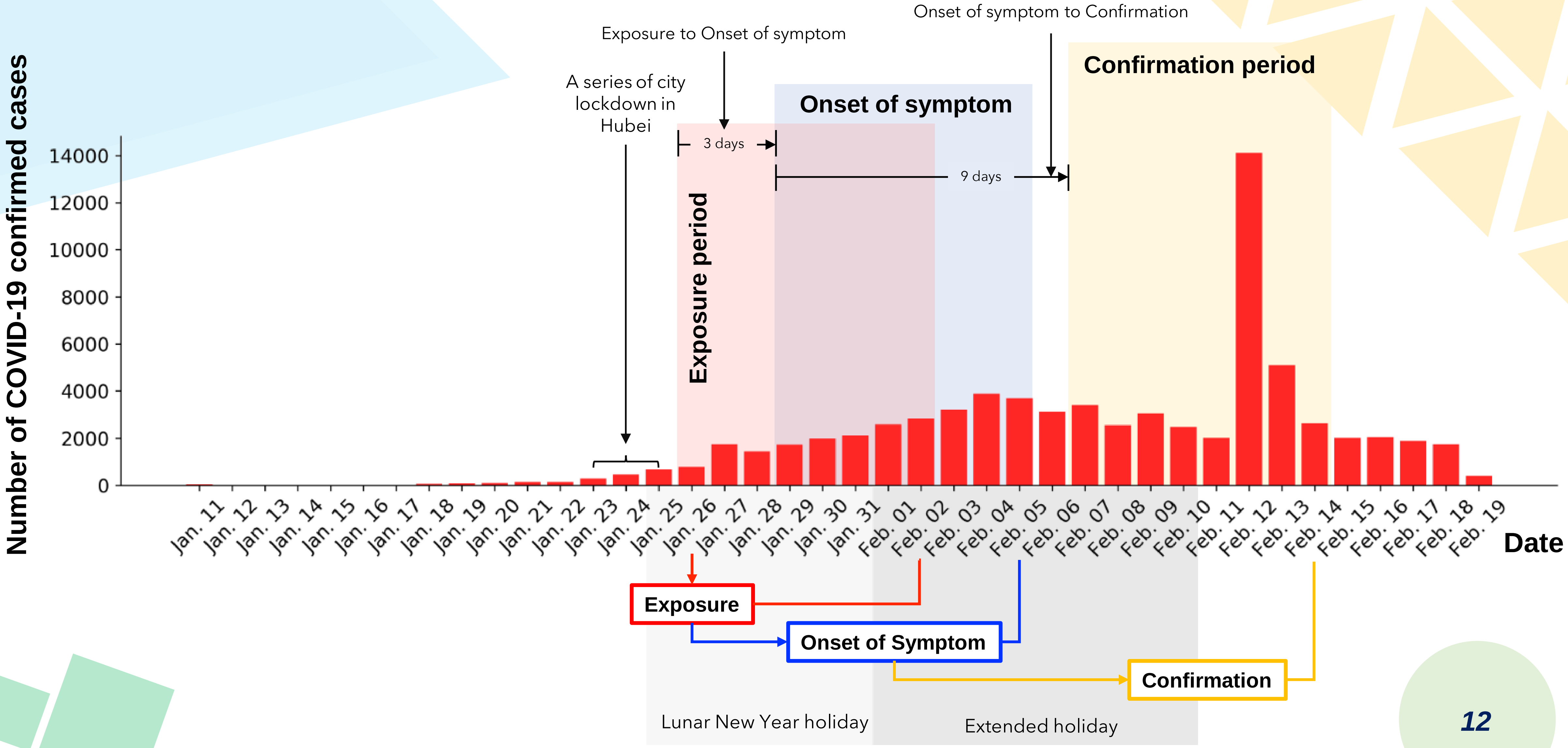
Error Bar Plot



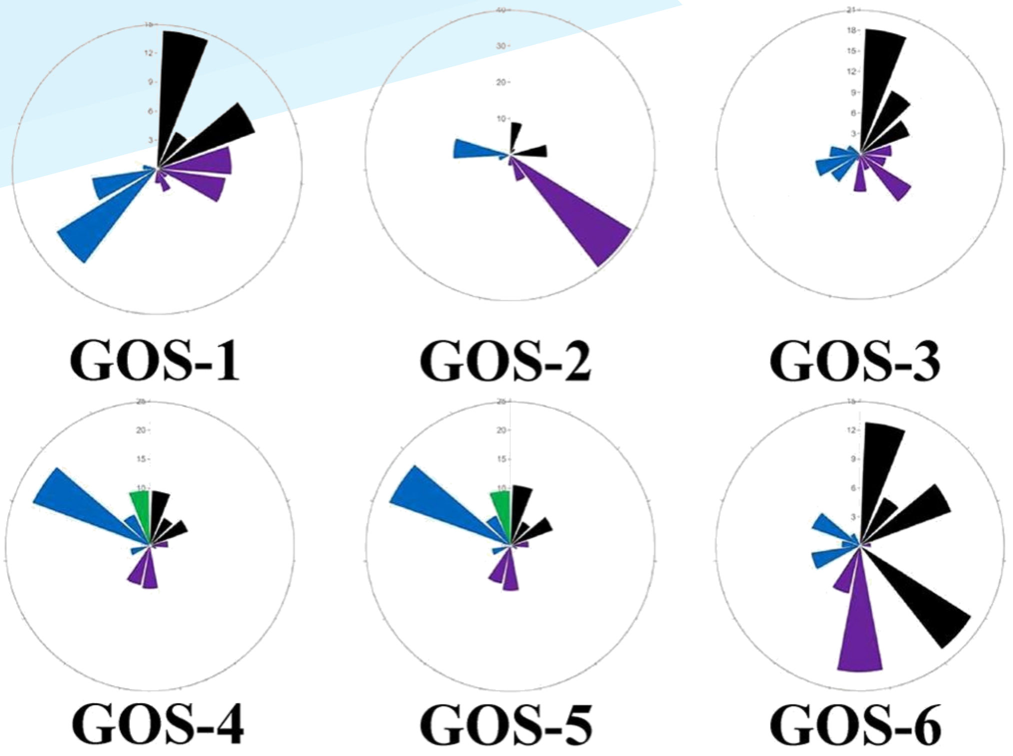
Pie Chart


42.9%
21.4%
7.1%
28.6%
(a)
29.6%
11.1%
14.8%
3.7%
40.7%
(b)
45.00%
5.00%
15.00%
35.00%
(c)
America
Africa
Asia
Europe
Pacific

Bar Chart

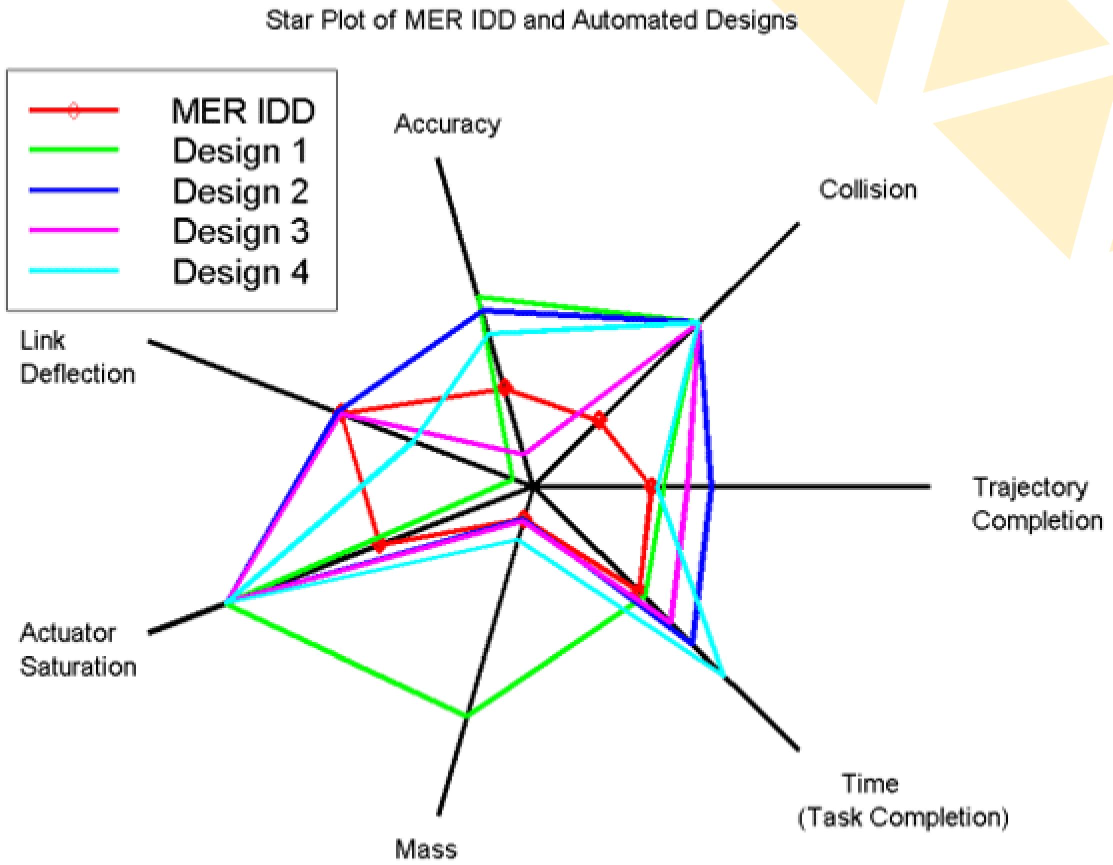

Onset of symptom to Confirmation
Exposure to Onset of symptom
Confirmation period
Onset of symptom
A series of city lockdown in Hubei
3 days
9 days
Exposure period
Number of COVID-19 confirmed cases
14000
12000
10000
8000
6000
4000
2000
0
Jan. 11
Jan. 12
Jan. 13
Jan. 14
Jan. 15
Jan. 16
Jan. 17
Jan. 18
Jan. 19
Jan. 20
Jan. 21
Jan. 22
Jan. 23
Jan. 24
Jan. 25
Jan. 26
Jan. 27
Jan. 28
Jan. 29
Jan. 30
Jan. 31
Feb. 01
Feb. 02
Feb. 03
Feb. 04
Feb. 05
Feb. 06
Feb. 07
Feb. 08
Feb. 09
Feb. 10
Feb. 11
Feb. 12
Feb. 13
Feb. 14
Feb. 15
Feb. 16
Feb. 17
Feb. 18
Feb. 19
Date
Exposure
Onset of Symptom
Confirmation
Lunar New Year holiday
Extended holiday

Rose Plot & Radar Plot



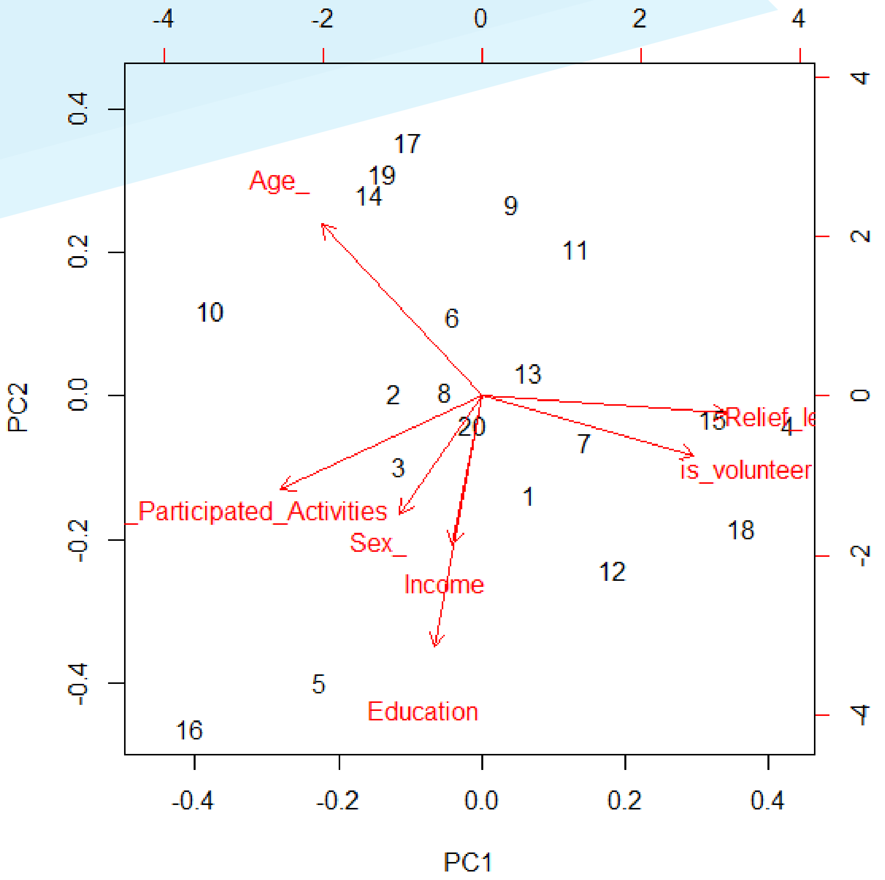
Profile diversity in various GOS samples

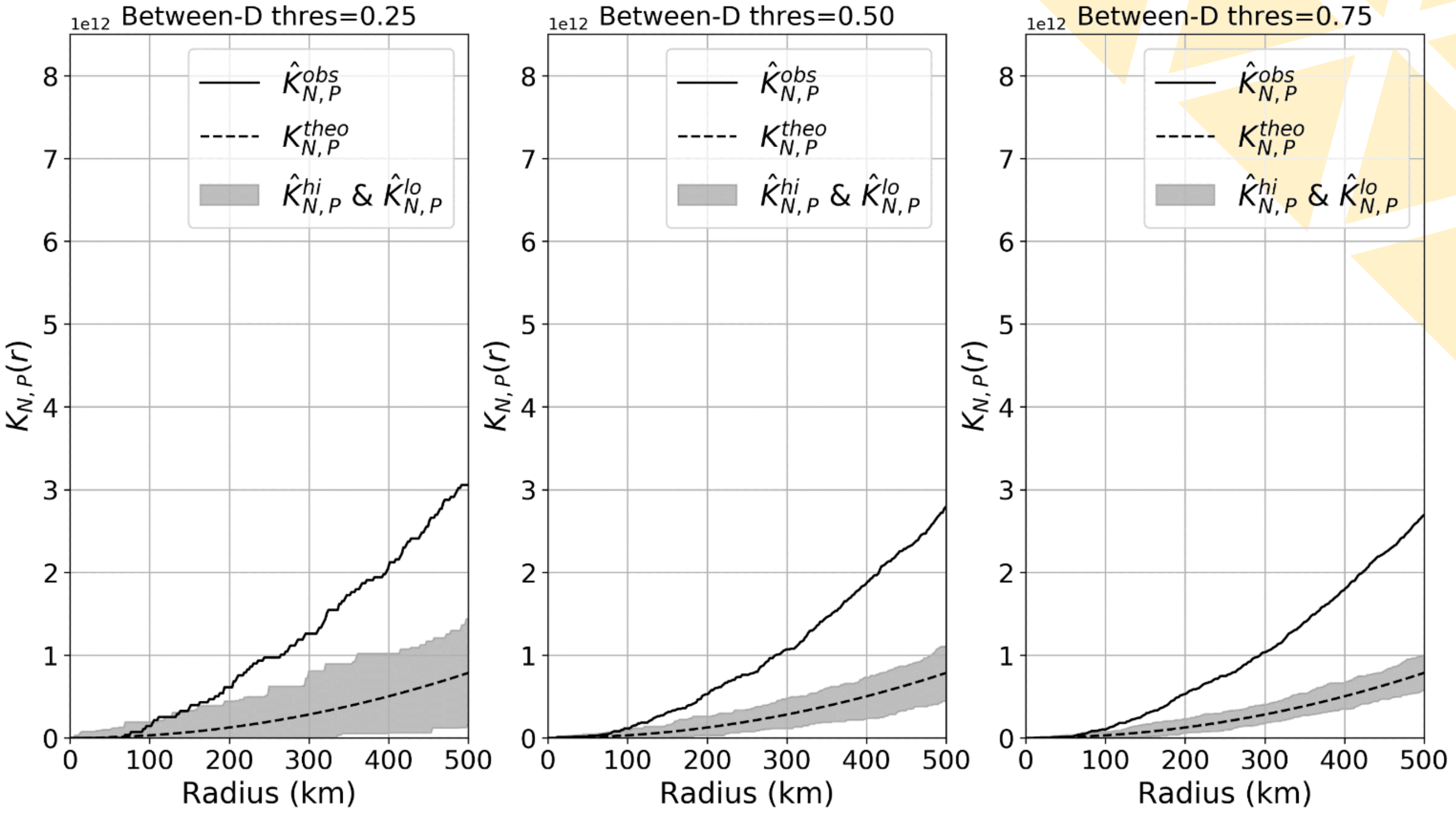
Lin et al. (2022) Profile diversity of galacto-oligosaccharides from disaccharides to hexasaccharides by porous graphitic carbon liquid chromatography-orbitrap tandem mass spectrometry. Food Chem. Vol. 390. 133151.



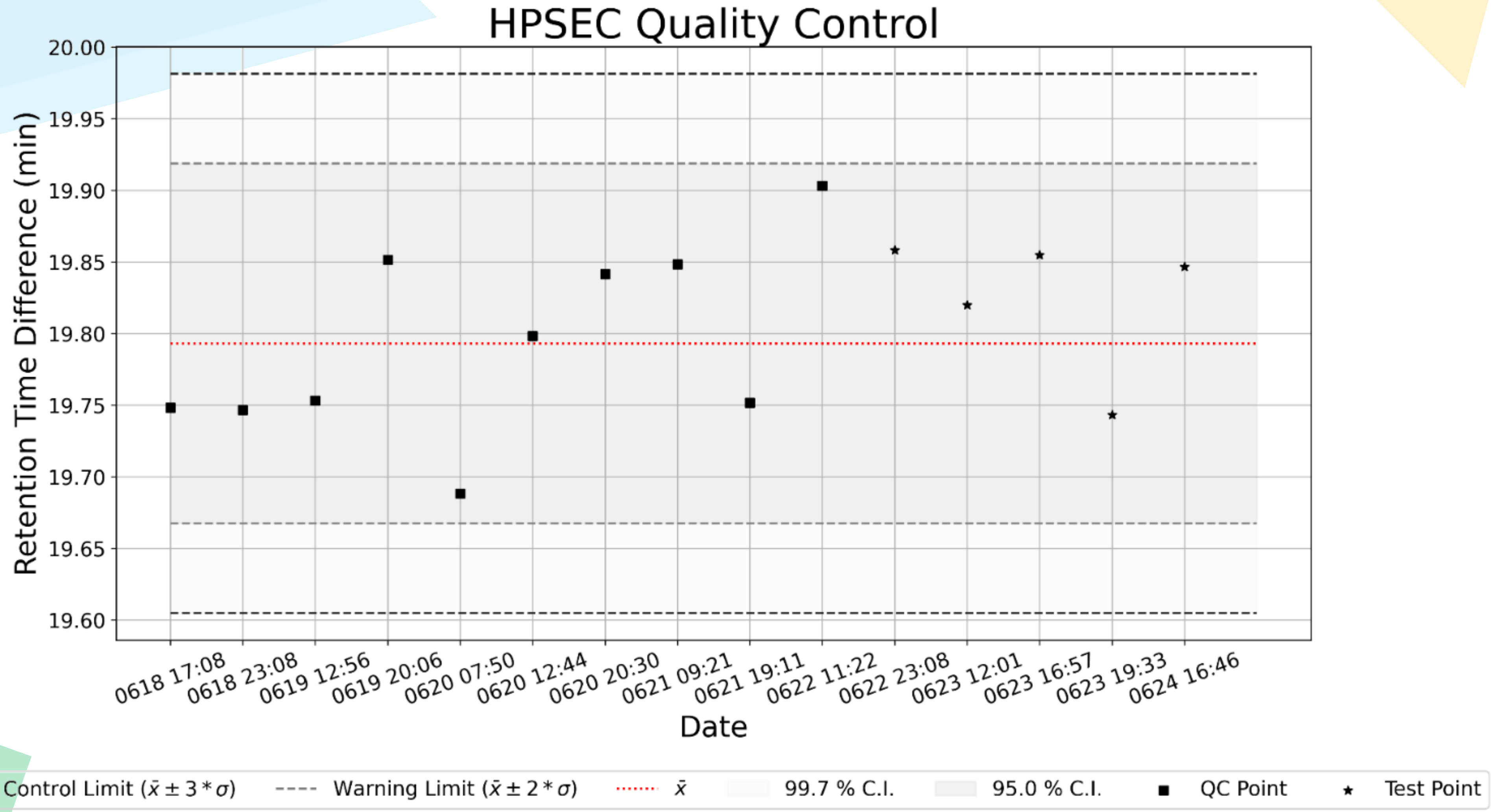
Source: https://en.wikipedia.org/wiki/Radar_chart#/media/File:MER_Star_Plot.gif

Biplot & Control Chart (I)

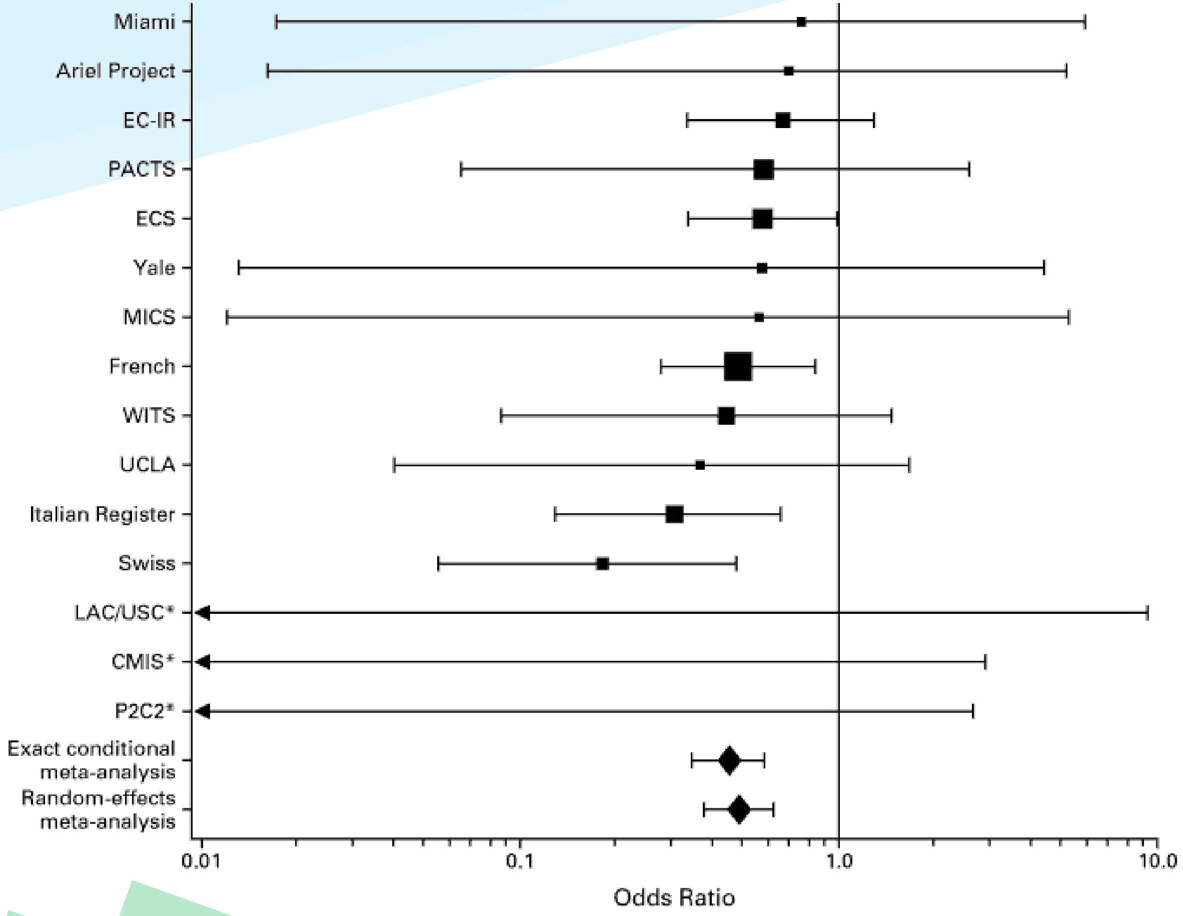




Control Chart (II)



Forest Plot



	Diseased	Healthy
Exposed	20	380
Not Exposed	10	490

$$risk\ of\ developing\ the\ disease\ given\ exposure = \frac{DE}{VE} = \frac{20}{400}$$

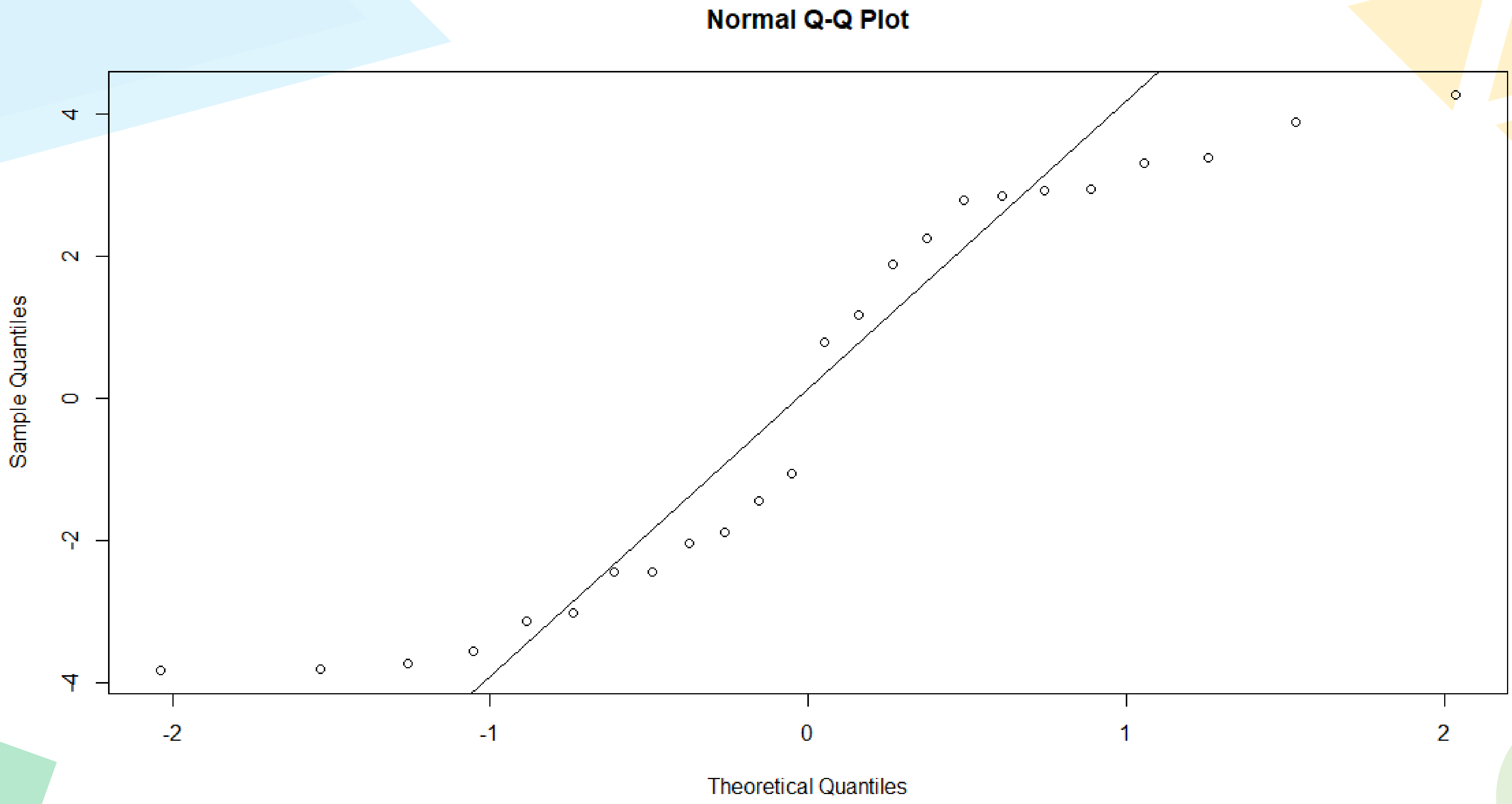
$$risk\ of\ developing\ the\ disease\ given\ non - exposure = \frac{DN}{VN} = \frac{10}{500}$$

$$relative\ risk = \frac{\frac{DE}{(DE+HE)}}{\frac{DN}{(DN+HN)}} = \frac{DE/VE}{DN/VN} = \frac{20/400}{10/500}$$

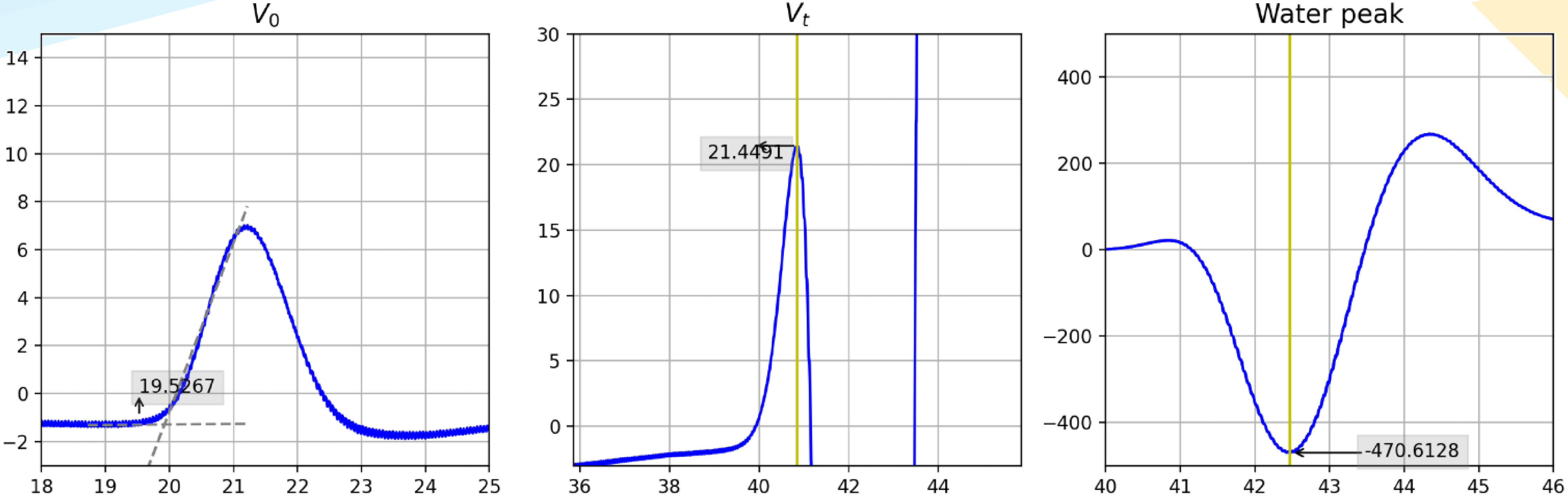
$$odds\ ratio = \frac{DE/HE}{DN/HN} = \frac{20/380}{10/500}$$

Source: https://www.nejm.org/doi/full/10.1056/NEJM199904013401301

Quantile-quantile Plot (Q-Q Plot)



Combination Chart



Question Time

If you have any questions, please do not hesitate to ask me.

The End

Thank you for your attention))