Central Tendency (Mean)

- Given a sample of n pieces of data $(y_1\ y_2, y_3, .., y_n)$ taken from a given population of size N, the arithmetic mean of the sample, denoted by $\bar{\boldsymbol{y}}$ is

$$\bar{y} = \sum_{k=1}^{n} \frac{y_i}{n}$$

- Population mean $\boldsymbol{\mu}$ given as,

$$\mu = \sum_{k=1}^{N} \frac{y_i}{N}$$

- **Note:** True mean ($\boldsymbol{\mu}$) of the population of size N could be different than sample mean $\bar{\boldsymbol{y}}$.

DIY

```
Can you show that as n -> N, the
sample mean -> population mean?
```

Dispersion/Variability

- **Range (R):** Difference between the largest value and the smallest value of the data

$$R = \max(y_i) - \min(y_i)$$

- **Variance (s^2):** Sample variance is given by

$$s^2 = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}$$

DIY

```
Why do we use (n-1) in the denominator for sample variance?
```

- Note, that the true variance (σ^2) of the population of size N could be different

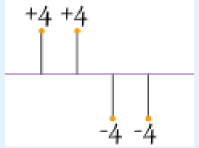
$$\sigma^2 = \sum_{i=1}^{N} \frac{(y_i - \mu)^2}{N}$$

- Notice (n-1) in the denominator for sample variance, while N for true variance
- Standard Deviation = Square root of Variance

Dispersion/Variability

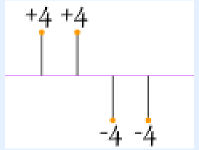
Why is variance defined as follows is a GOOD way to assess variability?

- What if we just add the deviations from the mean and take the average?

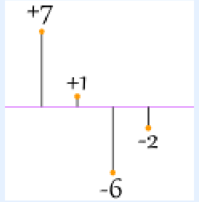


$$\frac{4+4-4-4}{4} = 0$$

- What if we just take absolute values of deviations from the mean?

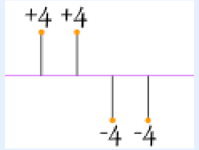


$$\frac{|4|+|4|+|-4|+|-4|}{4} = \frac{4+4+4+4}{4} = 4$$

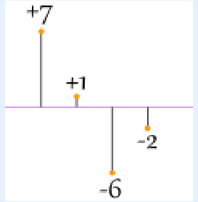


$$\frac{|7|+|1|+|-6|+|-2|}{4} = \frac{7+1+6+2}{4} = 4$$

- But, when we square..



$$\sqrt{\left(\frac{4^2+4^2+(-4)^2+(-4)^2}{4}\right)} = \sqrt{\left(\frac{64}{4}\right)} = 4$$



$$\sqrt{\left(\frac{7^2+1^2+(-6)^2+(-2)^2}{4}\right)} = \sqrt{\left(\frac{90}{4}\right)} = 4.74\ldots$$

Ref: https://www.mathsisfun.com/data/standard-deviation.html

Median and Mode

- The median of a finite list of numbers is the "middle" number when those numbers are listed in order from smallest to greatest.
- Mode is the most frequent value in the data set

Examples:

- Test Scores out of 20: (0, 11, 15, 8, 18, 19, 7, 8, 9, 12)
- Test Scores out of 20: (0, 11, 15, 8, 18, 7, 8, 9, 12)
- Test Scores out of 20: (0, 11, 15, 8, 18, 19, 7, 8, 9, 15)

Data Characterization

- It has been shown that the mean (average) can be used to describe the data characteristics.
- However, it is possible to find two sets of data to have equal averages, but different degrees of scatter
- It has been a common mistake in many cases of applications to put all emphasis on the average but overlook the scatter of the data
- Such a mistake usually leads to unnecessary erroneous conclusions which could have been easily avoided if the scatter of the data had been considered.

Data Presentation

Table 1. Modified Table in Lee and Kim's Research (Adapted from Korean J Anesthesiol 2017; 70: 39-45)

Variable	Group	Baseline	After drug	1 min	3 min	5 min
SBP	C	135.1 ± 13.4	139.2 ± 17.1	186.0 ± 26.6*	160.1 ± 23.2*	140.7 ± 18.3
	D	135.4 ± 23.8	131.9 ± 13.5	165.2 ± 16.2*,†	127.9 ± 17.5†	108.4 ± 12.6†,‡
DBP	C	79.7 ± 9.8	79.4 ± 15.8	104.8 ± 14.9*	87.9 ± 15.5*	78.9 ± 11.6
	D	76.7 ± 8.3	78.4 ± 6.3	97.0 ± 14.5*	74.1 ± 8.3†	66.5 ± 7.2†,‡
MBP	C	100.3 ± 11.9	103.5 ± 16.8	137.2 ± 18.3*	116.9 ± 16.2*	103.9 ± 13.3
	D	97.7 ± 14.9	98.1 ± 8.7	123.4 ± 13.8*,†	95.4 ± 11.7†	83.4 ± 8.4†,‡

Values are expressed as mean ± SD. Group C: normal saline, Group D: dexmedetomidine. SBP: systolic blood pressure, DBP: diastolic blood pressu
MBP: mean blood pressure, HR: heart rate.

Table 2. Difference between a Regular Table and a Heat Map

Example of a regular table				Example of a heat map			
SBP	DBP	MBP	HR	SBP	DBP	MBP	HR
128	66	87	87	128	66	87	87
125	43	70	85	125	43	70	85
114	52	68	103	114	52	68	103
111	44	66	79	111	44	66	79
139	61	81	90	139	61	81	90
103	44	61	96	103	44	61	96
94	47	61	83	94	47	61	83

All numbers were created by the author. SBP: systolic blood pressure, DBP: diastolic blood pressure, MBP: mean blood pressure, HR: heart rate.

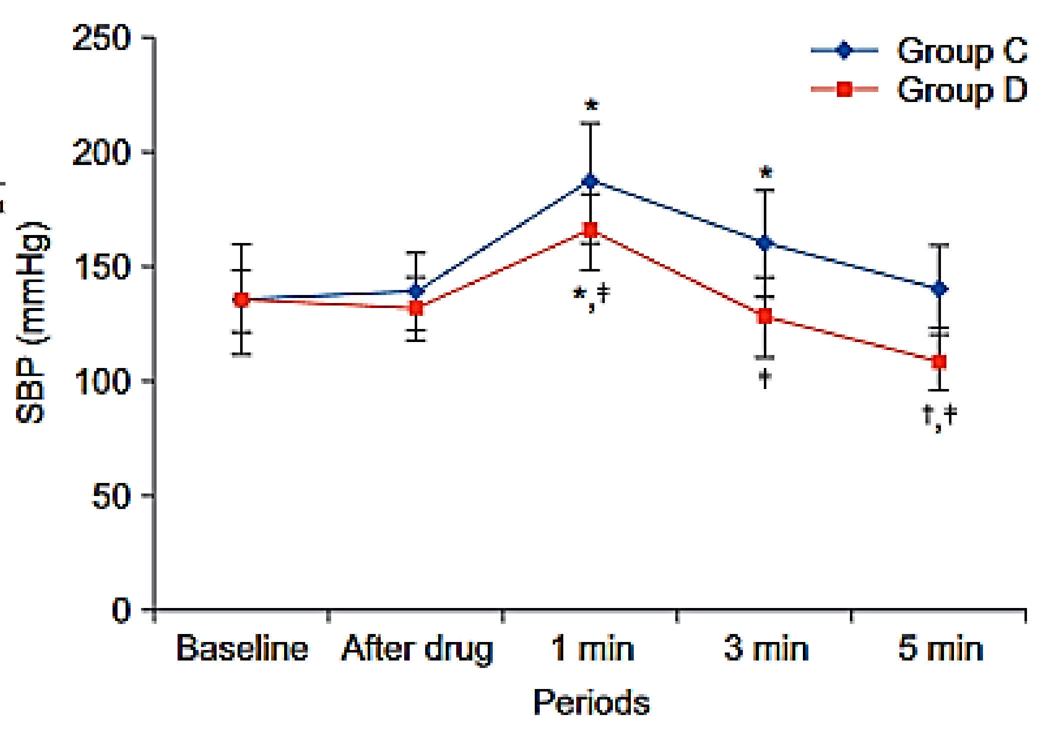
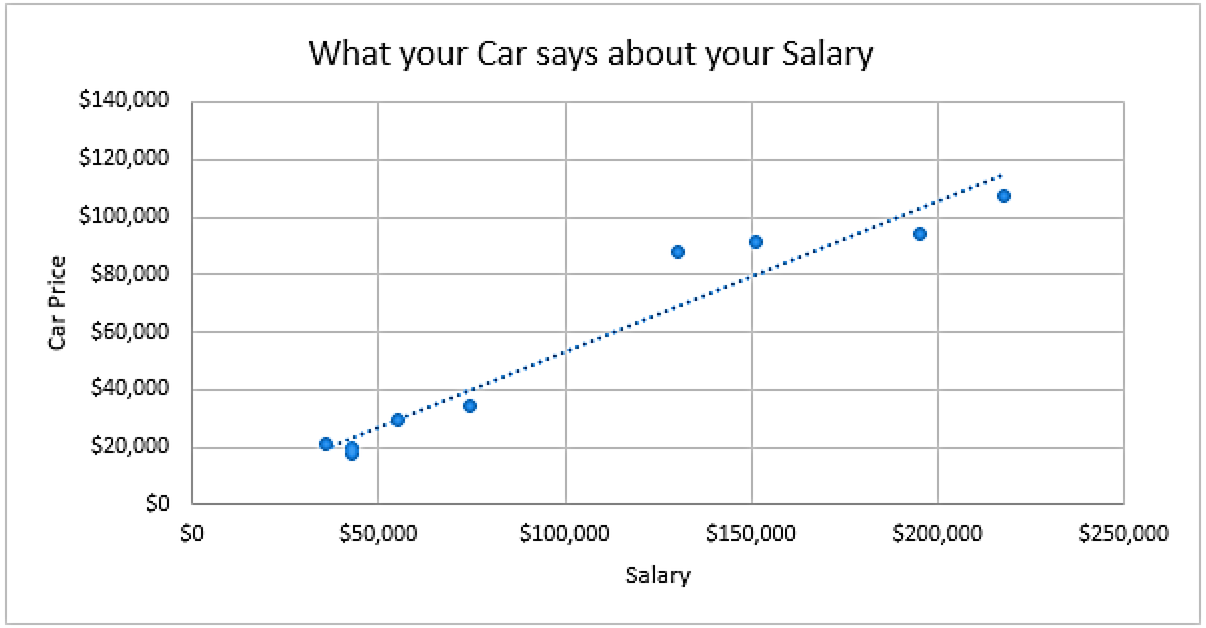


Fig. 1. Line graph with whiskers. Changes in systolic blood pressure (SBP) in the two groups. Group C: normal saline, Group D: dexmedetomidine.

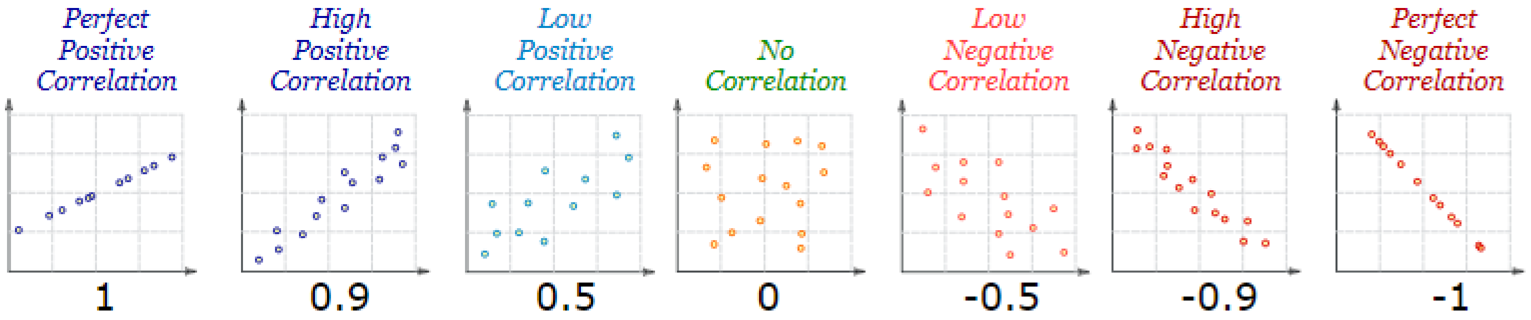
Reference: In, Junyong, and Sangseok Lee. "Statistical data presentation." *Korean journal of anesthesiology* 70.3 (2017): 267.

Data Presentation

Scatter Plot

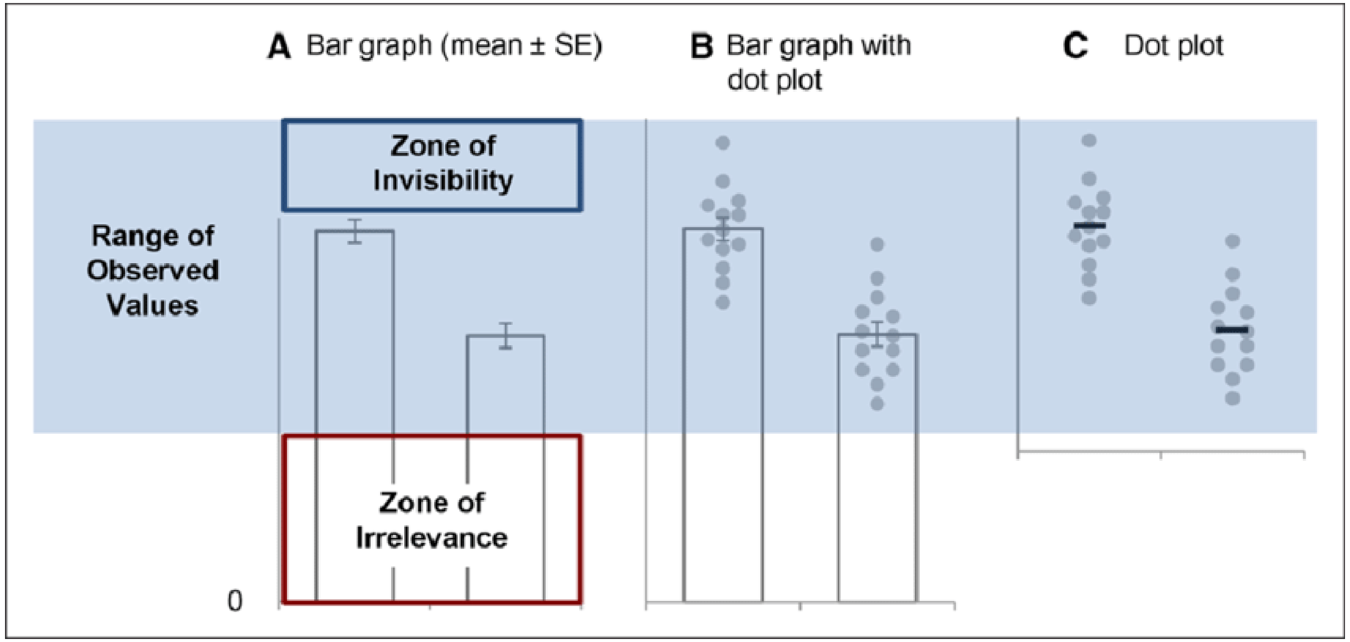


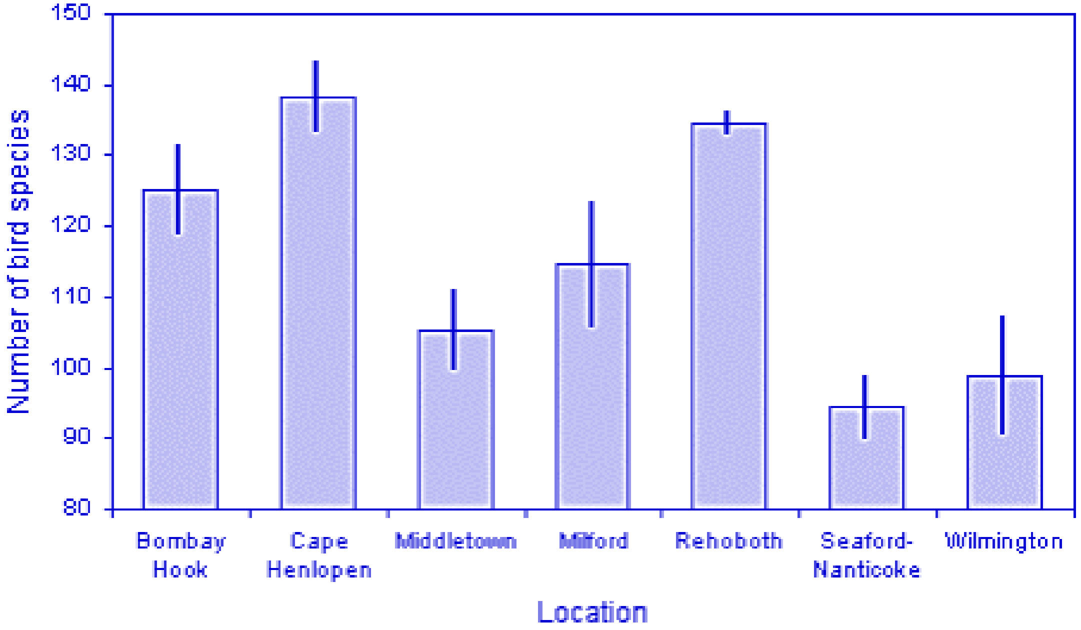
Scatter plots are used to investigate association between two variables



Data Presentation

Bar Graph

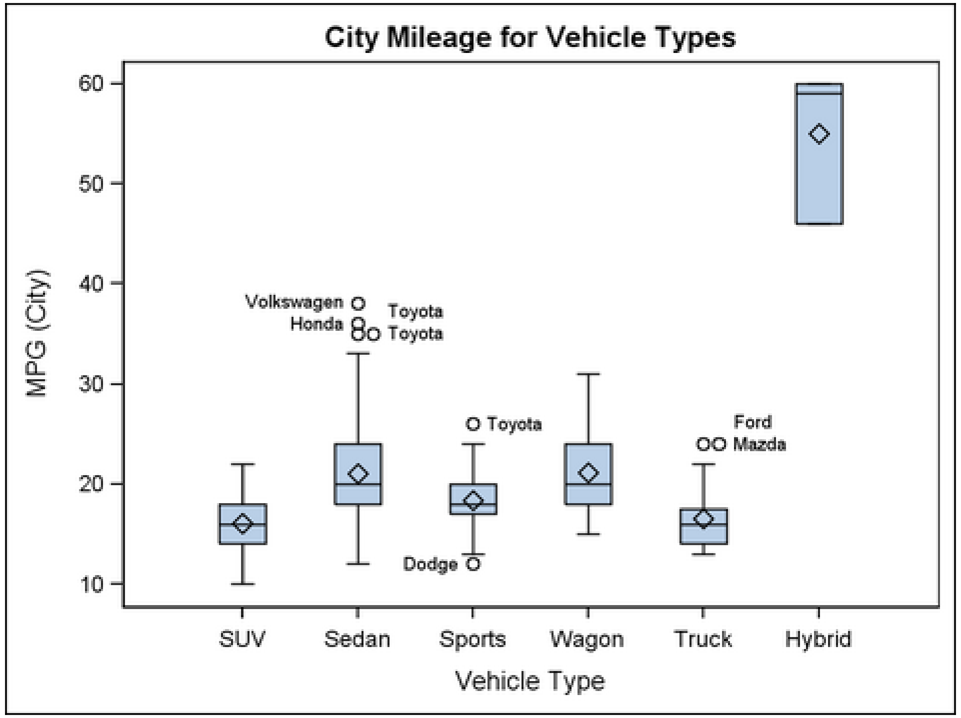


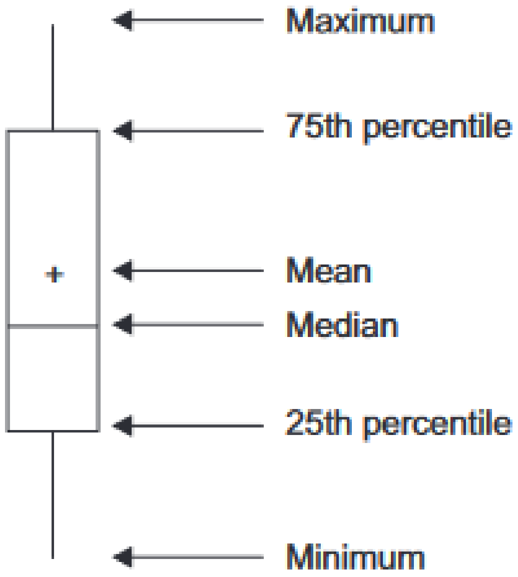


Bar graphs are used to indicate and compare values in discrete category or groups

Reference: In, Junyong, and Sangseok Lee. "Statistical data presentation." *Korean journal of anesthesiology* 70.3 (2017): 267.

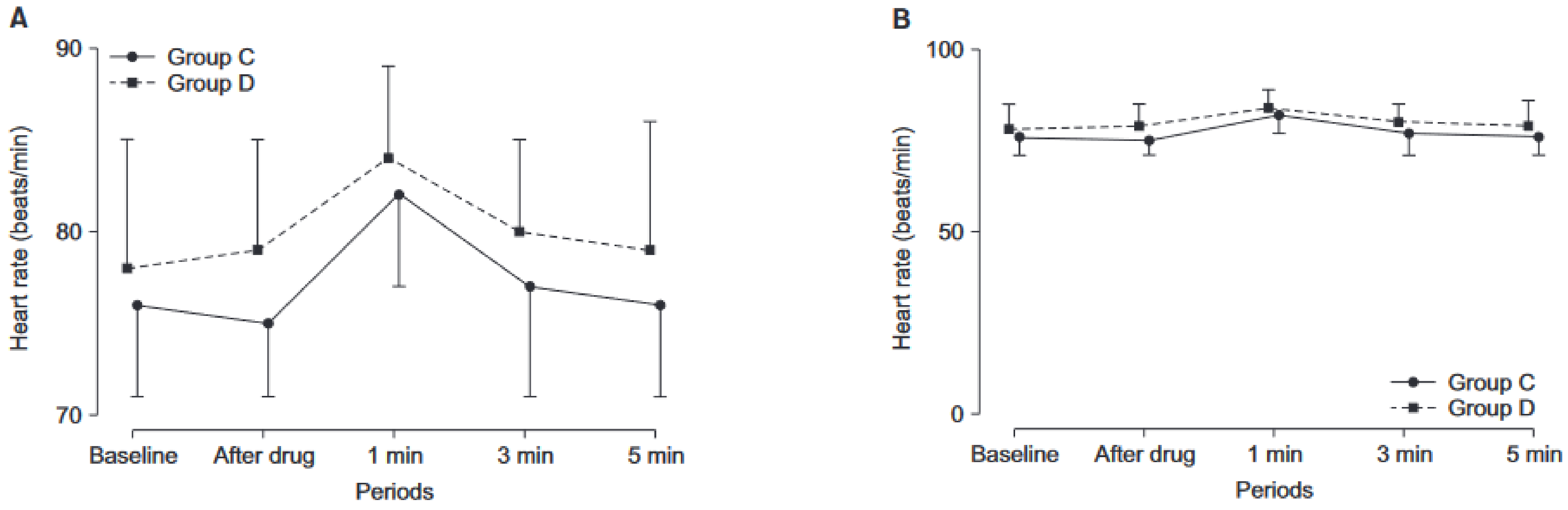
Box and Whisker Graph





Reference: In, Junyong, and Sangseok Lee. "Statistical data presentation." *Korean journal of anesthesiology* 70.3 (2017): 267.

Data Presentation



Misleading Plot

Reference: In, Junyong, and Sangseok Lee. "Statistical data presentation." *Korean journal of anesthesiology* 70.3 (2017): 267.

Data Presentation

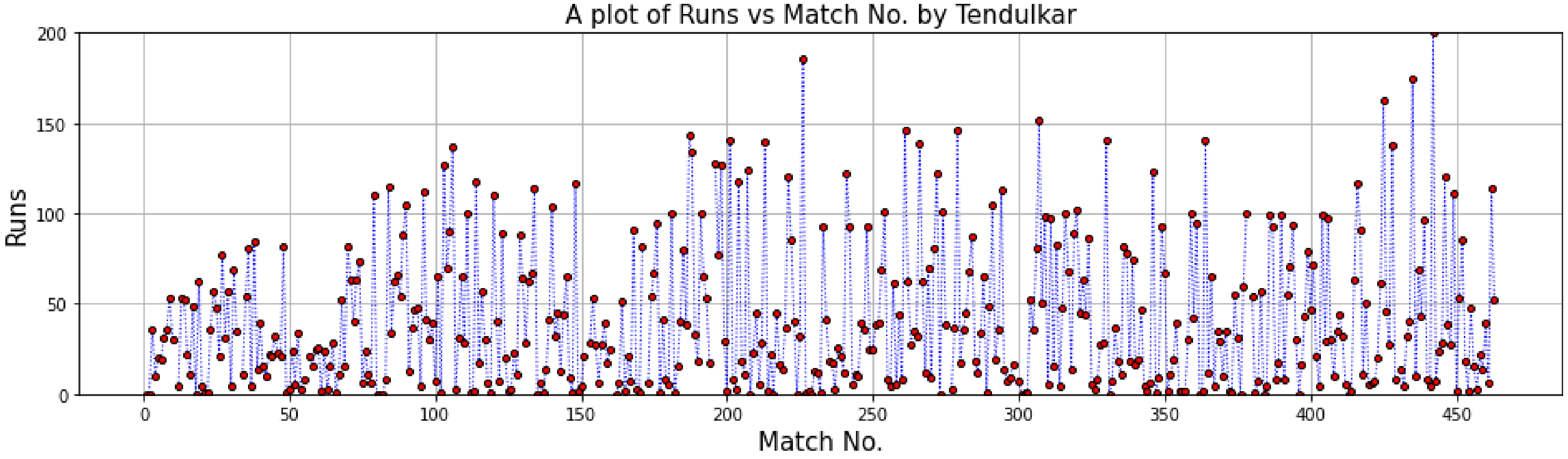
Table 3. Types of Charts depending on the Method of Analysis of the Data

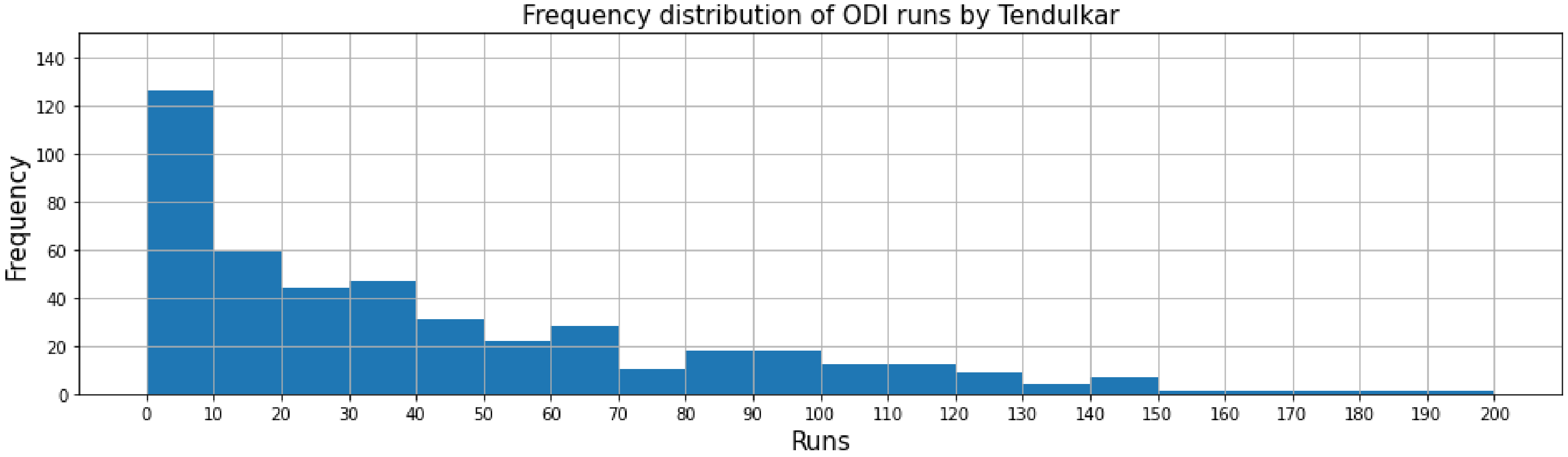
Analysis	Subgroup	Number of variables	Type
Comparison	Among items	Two per items	Variable width column chart
		One per item	Bar/column chart
	Over time	Many periods	Circular area/line chart
		Few periods	Column/line chart
Relationship		Two	Scatter chart
		Three	Bubble chart
Distribution		Single	Column/line histogram
		Two	Scatter chart
		Three	Three-dimensional area chart
Comparison	Changing over time	Only relative differences matter	Stacked 100% column chart
		Relative and absolute differences matter	Stacked column chart
	Static	Simple share of total	Pie chart
		Accumulation	Waterfall chart
		Components of components	Stacked 100% column chart with subcomponents

Reference: In, Junyong, and Sangseok Lee. "Statistical data presentation." *Korean journal of anesthesiology* 70.3 (2017): 267.

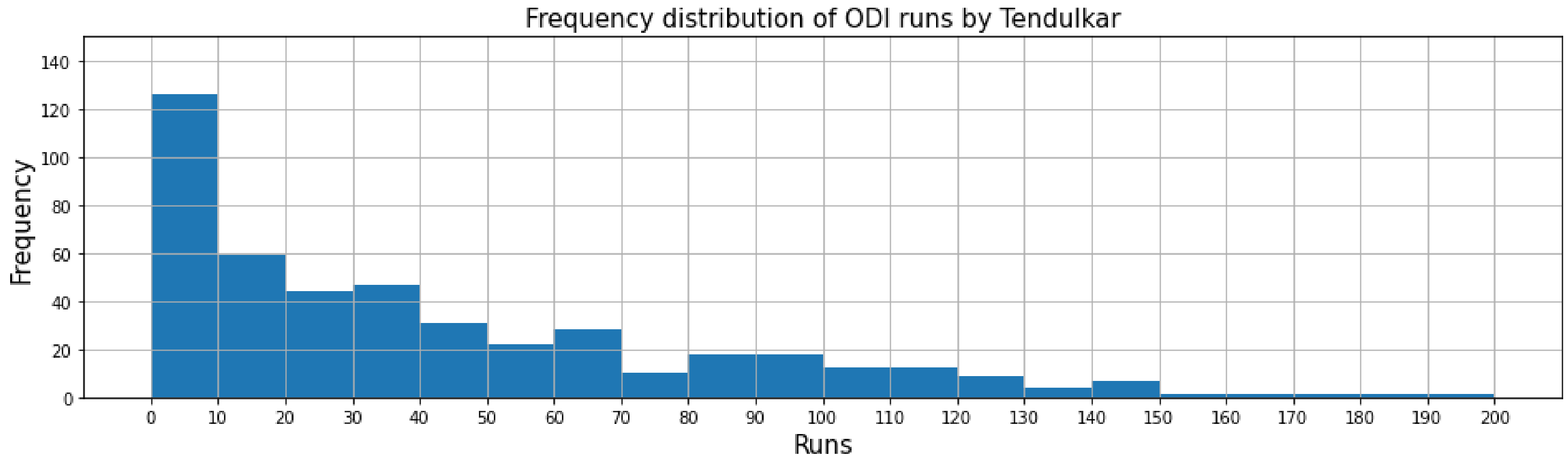
Shape of Frequency Distribution

CEP2022_Notebook (1.4)





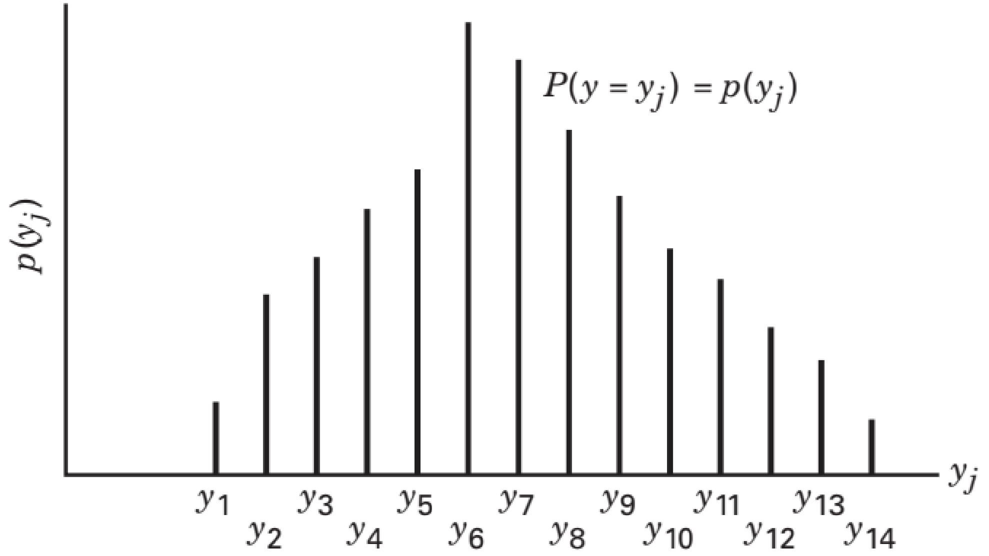
Shape of Frequency Distribution



Questions:

- What is the area under the curve?
- Given such data, how would you calculate the probability of Tendulkar scoring a given number of runs?
- How would you then convert the Y-axis to probability?
- What happens when the bin size $\rightarrow 0$

Probability Distribution



(*a*) A discrete distribution

y discrete: $\quad 0 \le p(y_j) \le 1 \quad$ all values of y_j

$$P(y = y_j) = p(y_j) \quad \text{all values of } y_j$$

$$\sum_{\substack{\text{all values} \\ \text{of } y_j}} p(y_j) = 1$$

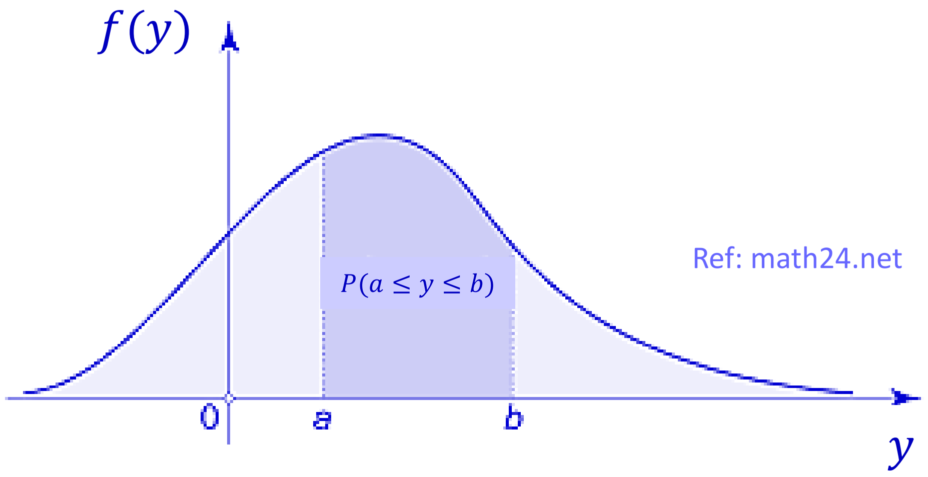
Probability Density/Distribution Function

- For a continuous random variable 'y', the probability behavior is described by a function called 'probability density function' (PDF) = $\boldsymbol{f}$(y)
- What are the properties of such PDF?

$$\boldsymbol{f(y) \geq 0}$$

$$\boldsymbol{\int_{-\infty}^{\infty} f(y)dy = 1}$$

$$\boldsymbol{Probability\ (a \leq y \leq b) = \int_a^b f(y)dy}$$



Ref: math24.net

- Cumulative distribution function (CDF) for a continuous random variable x with pdf $f(X)$

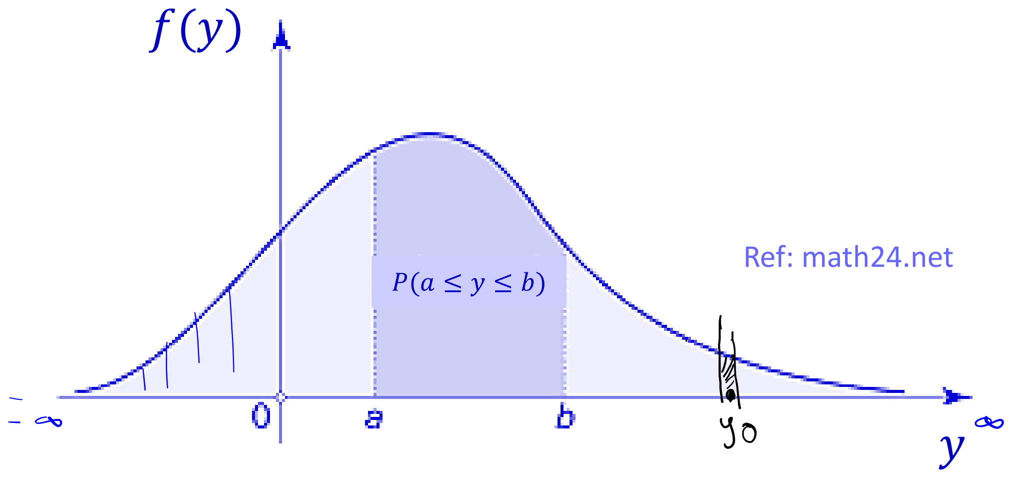
$$F(y) = Probability(Y \leq y) = \int_{-\infty}^{y} f(Y)dY$$

$$\text{Note:} \quad f(y) = \frac{dF(y)}{dy}$$

Probability Density Function

- Given $f(y)$, how would you find the *true arithmetic mean (μ)* value of 'y'?

$$\mu = \int_{-\infty}^{\infty} y\, f(y)\, dy$$



- What about *true variance (σ^2)*?

$$\sigma^2 = \int_{-\infty}^{\infty} (y-\mu)^2 f(y)\, dy$$

- The expectation of a function g(y) of a random variable 'y' with pdf 'f(y)' is defined as,

$$\boldsymbol{E}(g(y)) = \int_{-\infty}^{\infty} g(y) f(y) dy$$

$E(y) = \mu$

$E((y-\mu)^2) = \sigma^2$