Basic Concepts in Statistics

Presenter: Bui Minh Tri

Objective + Outline

- Objective

Basic concepts in statistics

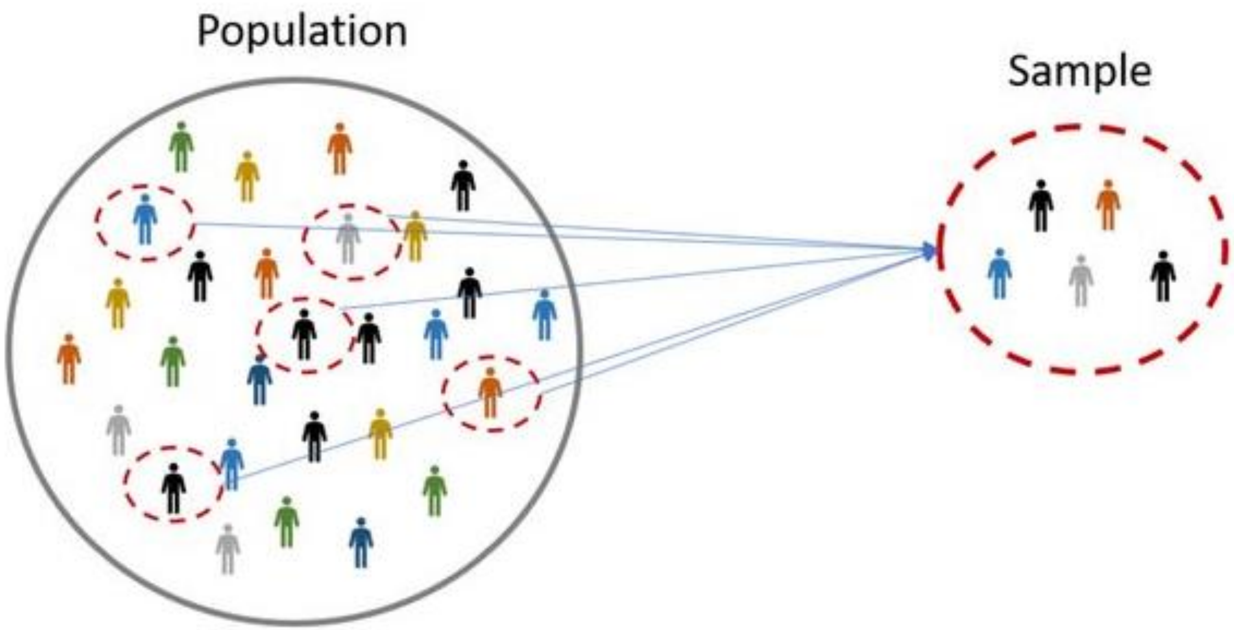
- Outline

1. **Population vs Sample**
2. **Mean vs median**
3. **Interquartile range and Boxplot**
4. **Outliers**
5. **Assessing the fit of mean**

Population vs Sample

Population vs Sample

Population	Sample
Entire group - you want to draw conclusions	A specific group - you will collect data from.
Difficult to collect data	Easy to collect data
Parameter - descriptive measure of population	Statistic - descriptive measure of the sample
Reports – true representation	Reports – confidence interval



Mean vs median

Mean

- Mean = **Average of a data set**.

- Adding up all the values ➢ Dividing by the total number of value.

$$Mean = \frac{Sum\ of\ all\ values}{Number\ of\ values}$$

Mean

```
# Create random data
```{r}
set.seed(123)  # Set seed for reproducibility

n <- 20  # Number of values
num_outliers <- 3  # Number of outliers

# Generate 17 random integer values greater than 0
x <- sample(1:100, n - num_outliers, replace = TRUE)

# Generate 3 outliers
outliers <- sample(201:300, num_outliers, replace = TRUE)

# Combine the random values and outliers
x <- c(x, outliers)

x
```
```

```
 [1]  31  79  51  14  67  42  50  43  14  25  90  91  69  91  57  92   9 293 299 272
```

```
## Mean
```{r}
mean(x)  # Calculate the mean
```
```

```
[1] 88.95
```

$$\text{Mean} = \frac{31+79+51+14+67+42+50+43+14+25+90+91+69+91+57+92+9+293+299+272}{20} = 88.95$$

Median

- **The middle value** when observations are **ordered from least to most**.
- **Unaffected by extreme values**.
- Step by step:

1. Order scores from least to most:

   2 3 3 6 10 **2 3 3 5 6 10**

2. Find the middle position: (n + 1) / 2.

$$\frac{5+1}{2} = 3 \qquad \frac{6+1}{2} = 3.5$$

Median

3. If the value is a **whole number**, median is value at the **middle position**.

2 3 3 6 10

$$M = 3$$

2 3 3 5 6 10

If not, median = two middlemost scores / 2.

2 3 3 5 6 10

$$M = \frac{3+5}{2} = 4$$

Median

1. **Order scores from least to most (n=20)**

9 14 14 25 31 42 43 50 51 57 67 69 79 90 91 91 92 272 293 299

2. **Find the middle position**: $\frac{20+1}{2} = 10.5$

3. **Median = two middlemost scores / 2.**

9 14 14 25 31 42 43 50 51 **57** **67** 69 79 90 91 91 92 272 293 299

$$Median = \frac{57+67}{2} = 62$$

Median

```
# Create random data
```{r}
set.seed(123)  # Set seed for reproducibility

n <- 20  # Number of values
num_outliers <- 3  # Number of outliers

# Generate 17 random integer values greater than 0
x <- sample(1:100, n - num_outliers, replace = TRUE)

# Generate 3 outliers
outliers <- sample(201:300, num_outliers, replace = TRUE)

# Combine the random values and outliers
x <- c(x, outliers)

x
```
```

```
 [1]  31  79  51  14  67  42  50  43  14  25  90  91  69  91  57  92   9 293 299 272
```

```
```{r}
# Calculate the median
median(x)
```
```

```
[1] 62
```

Mean vs Median

9 14 14 25 31 42 43 50 51 **57** **67** 69 79 90 91 91 92 272 293 299

Mean = 88.95 **Median = 62**

9 14 14 25 31 42 43 50 51 **57** **67** 69 79 90 91 91 92 **500 550 600**

Mean = 128.25 **Median = 62**

-20 **-10** **0** 25 31 42 43 50 51 **57** **67** 69 79 90 91 91 92 272 293 299

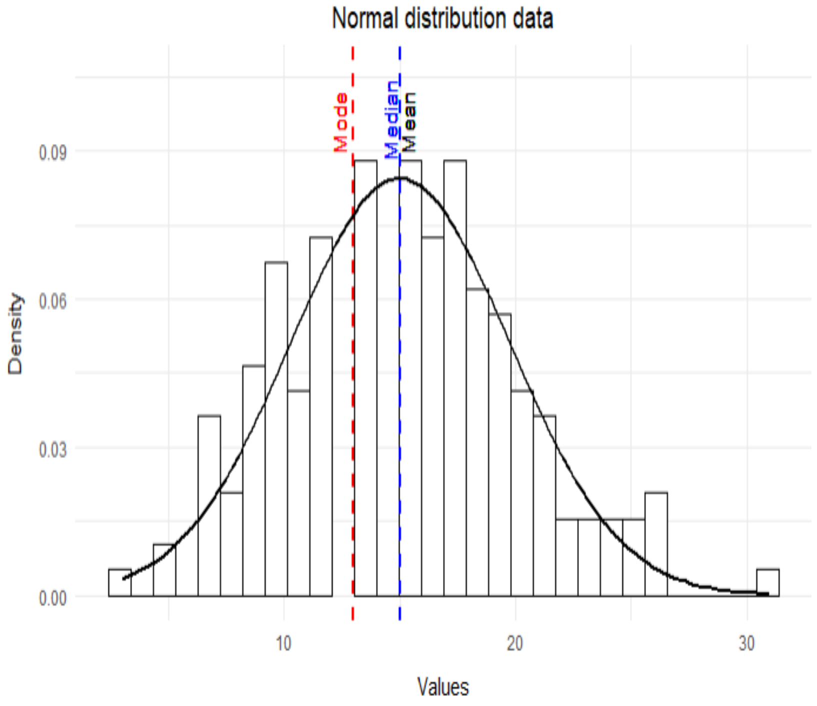
Mean = 85.6 **Median = 62**

Mean vs Median

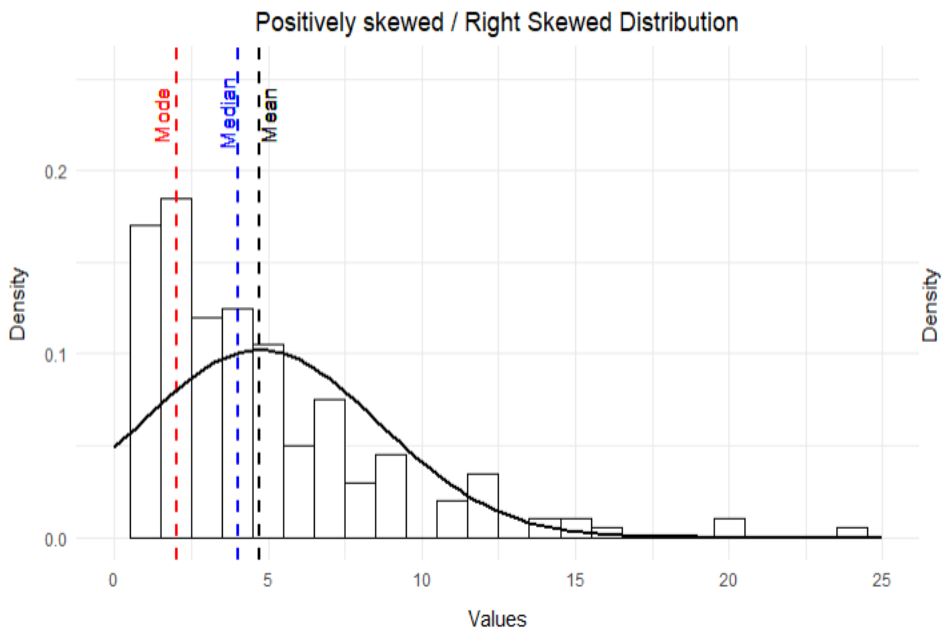
	Mean	Median
Concept	Average value of data	Middle value of ranked data
Extreme values	Sensitivity	Not affected
Skewed distributions	Pulled in skew direction	Less affected by skewness

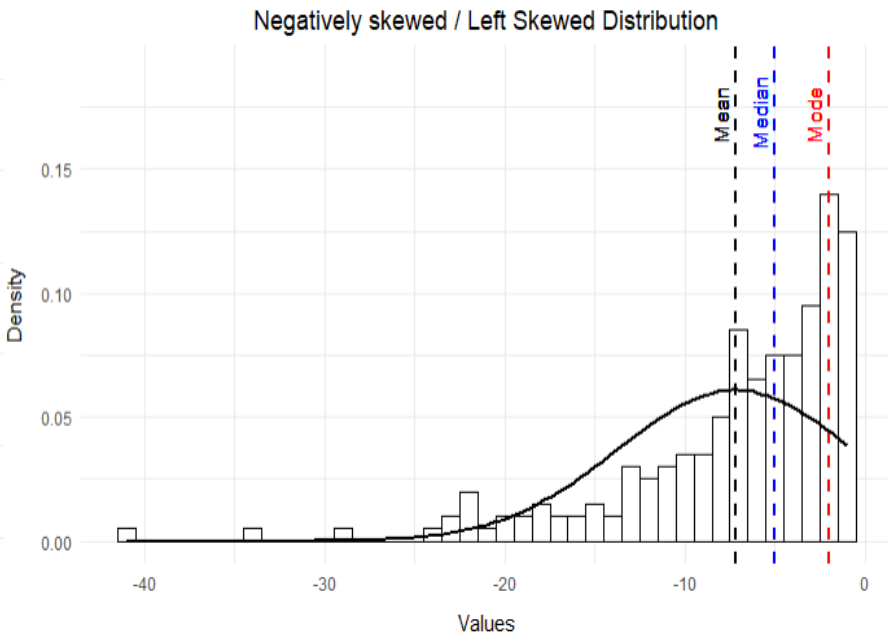
Normal vs Skewed Distributions

- ✓ Vertical line through centre of the distribution ➢ Both sides are similar.
- ✓ Mean ≈ Median
- ✓ Majority of scores at the centre.



Normal vs Skewed Distributions





	Right Skewed Distribution	Left Skewed Distribution
Long tail	Right side	Left side
Extreme values	High end	Low end
Mean vs Median	Mean > Median	Mean < Median

Interquartile range

Boxplot

Interquartile range (IQR)

- **IQR = Q3 – Q1** : Measure of how spread out the data.
- **Resistant** to distorting effects of **extreme scores**
- Q1: 25th percentile of the data.
- Q2: 50th percentile of the data = median of the dataset.
- Q3: 75th percentile of the data.
- Find IQR:
  - ✓ Find Q1 position: median of the **first 25% of values =** 0.25 (n + 1)
  - ✓ Find Q3 position: median of the **last 25% of values =** 0.75 (n + 1)
  - ✓ Identify Q3 and Q1 values.
  - ✓ IQR = Q3 – Q1.

Interquartile range (IQR) - Practice

n=11 1, 4, **6**, 9, 15, **21**, 22, 27, **35**, 40, 41

- Position Q1 and Q3:
  - ✓ Q1 = 0.25 (n +1) = 0.25 * 12 = 3
  - ✓ Q3 = 0.75 (n +1) = 0.75 * 12 = 9
- Value Q1 and Q3:
  - ✓ Q1 = 6
  - ✓ Q3 = 35
- IQR = Q3 – Q1 = 35 – 6 = 29

Interquartile range (IQR) - Practice

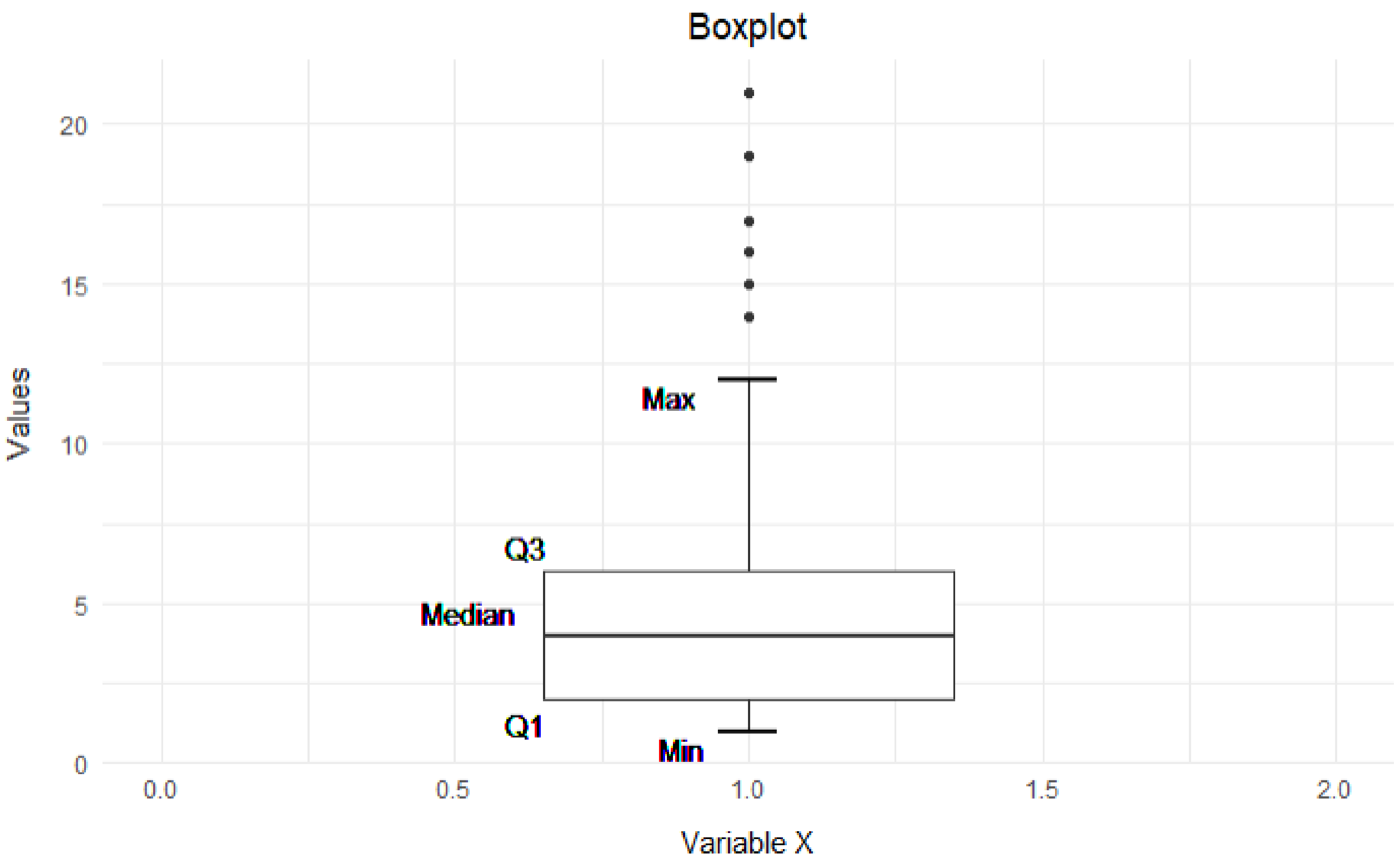
n=12: 1, 4, **6, 9**, 15, 21, 22, 27, **35, 40**, 41, 56

- Position Q1 and Q3:
  - ✓ Q1 = 0.25 (n +1) = 0.25 * 13 = 3.25
  - ✓ Q3 = 0.75 (n +1) = 0.75 * 13 = 9.75
- Value Q1 and Q3:
  - ✓ Q1 = (6+9) / 2 = 7.5
  - ✓ Q3 = (35+40) / 2 = 37.5
- IQR = Q3 – Q1 = 37.5 – 7.5 = 30

Boxplot - Concepts

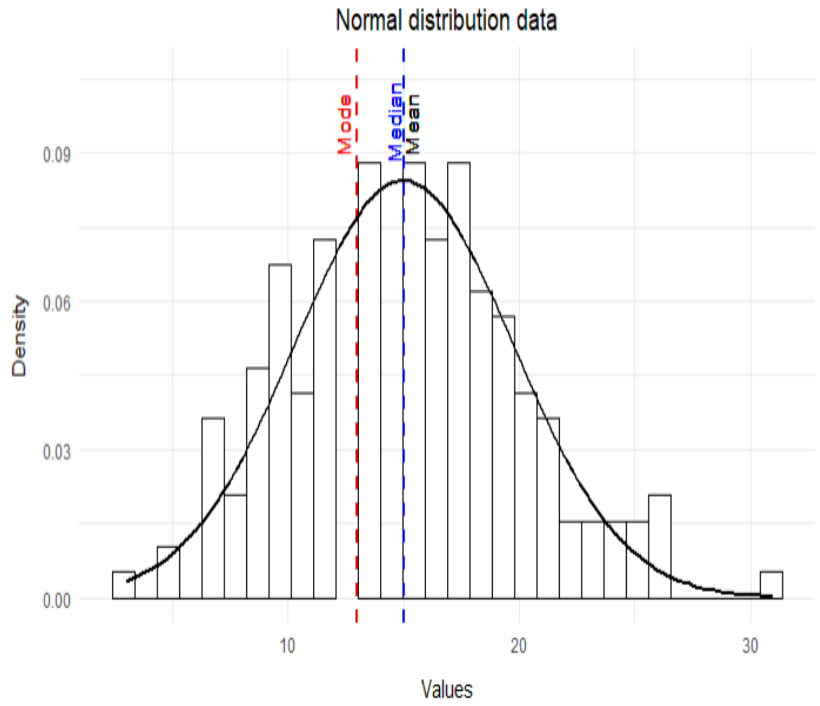
- Visualize a quantitative variable by 5 common location summary
  - ✓ Median
  - ✓ Q1 First quartiles
  - ✓ Q3 Third quartiles
  - ✓ Minimum
  - ✓ Maximum
  - ❖ Suspected outlier using IQR criterion

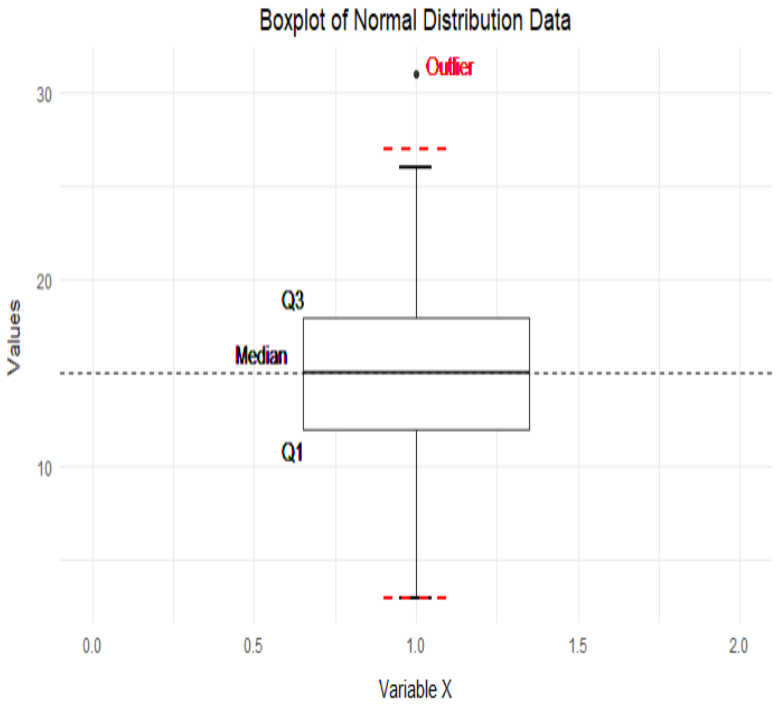
Boxplot - Concepts

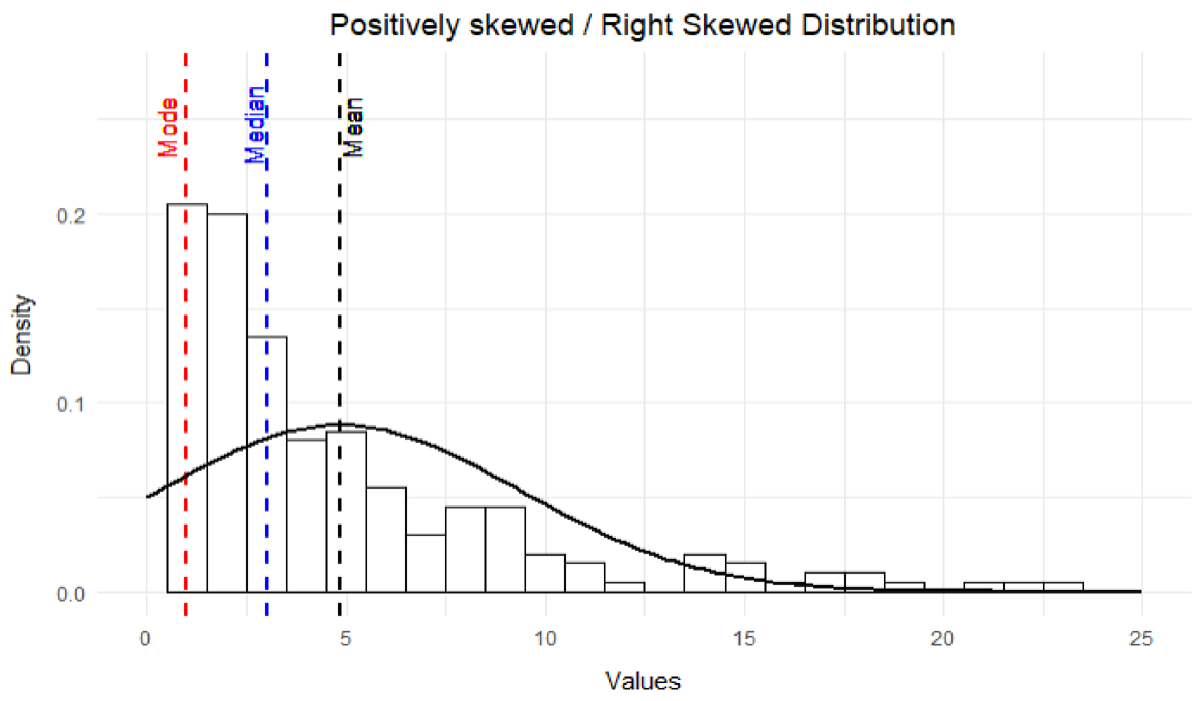


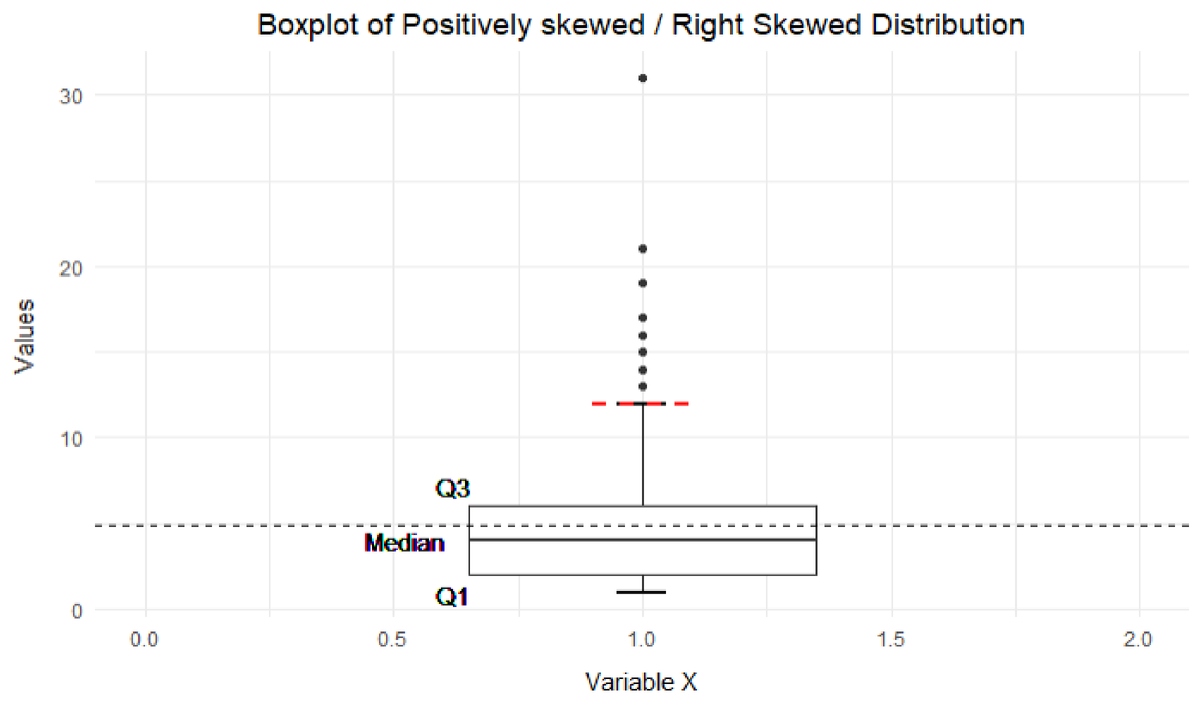
Boxplot – Normal Distributions

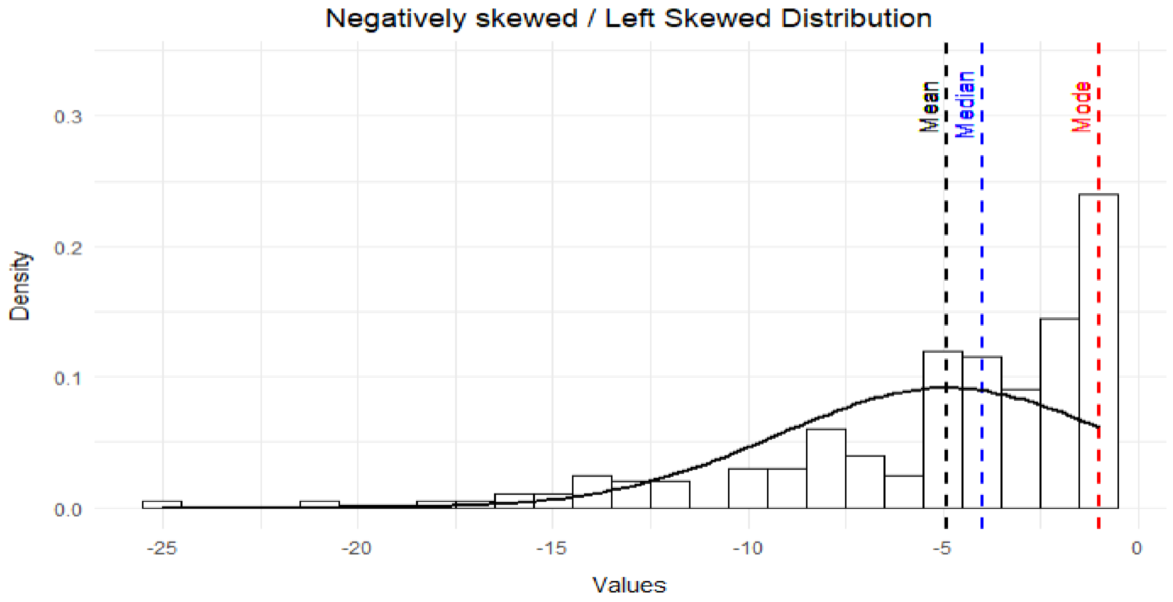
- ✓ Vertical line through centre of the distribution ➢ Both sides are similar.
- ✓ Mean ≈ Median
- ✓ Majority of scores at the centre.

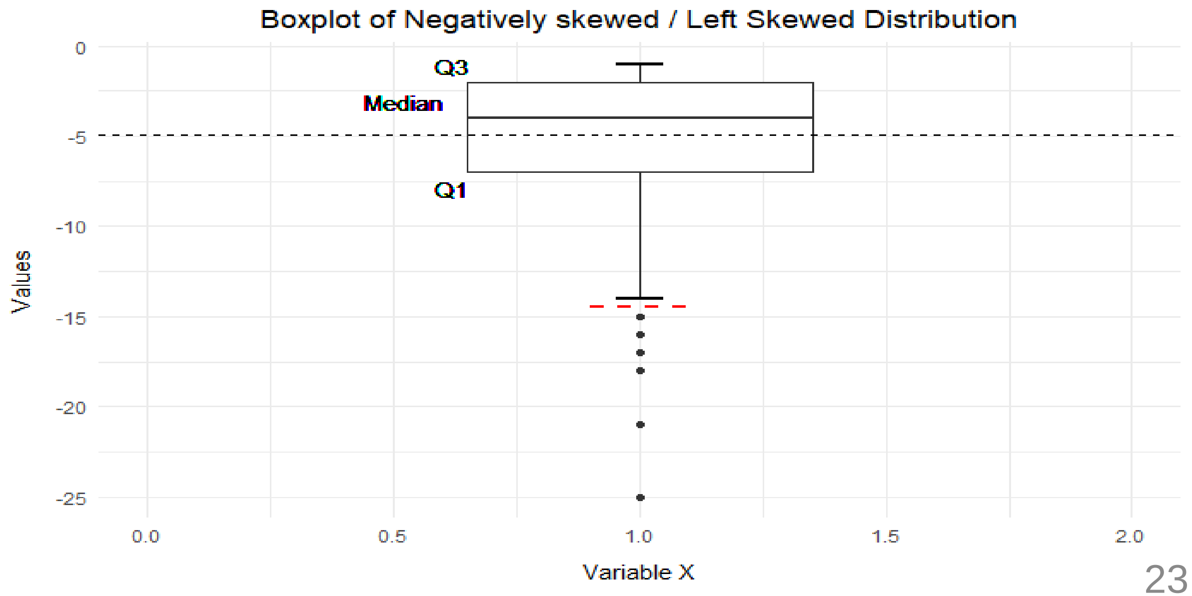





Positively skewed / Right Skewed Distribution
Mode
Median
Mean
Density
0.2
0.1
0.0
0
5
10
15
20
25
Values


Boxplot of Positively skewed / Right Skewed Distribution
30
20
10
0
Values
Q3
Median
Q1
0.0
0.5
1.0
1.5
2.0
Variable X


Negatively skewed / Left Skewed Distribution
Mean
Median
Mode
Density
0.3
0.2
0.1
0.0
-25
-20
-15
-10
-5
0
Values


Boxplot of Negatively skewed / Left Skewed Distribution
0
-5
-10
-15
-20
-25
Values
Q3
Median
Q1
0.0
0.5
1.0
1.5
2.0
Variable X

Outliers

Overview

2013, Herman Aguinis et al:

- 14 definitions
- 39 identification techniques
- 20 handling techniques

Definitions - Error outliers

- Data points that lie at a distance from other data points.
- The result of inaccuracies:
  - ✓ Not targeted population of interest.
  - ✓ Outside the possible range of values.
  - ✓ Observation.
  - ✓ Recording.
  - ✓ Preparing data.
  - ✓ Coding or in data manipulation.

Definitions - Interesting Outliers

- Outlying data points that are accurate.
- Not confirmed as actual error outliers.
- These cases may contain potentially valuable or unexpected knowledge.

Definitions

Table 1. Outlier Definitions Based on a Review of Methodological and Substantive Organizational Science Sources.

1. Single construct outliers	Data values that are unusually large or small compared to the other values of the same construct. These points typically fall in the tails of a data distribution.
2. Error outliers	Data points that lie at a distance from other data points because they are the result of inaccuracies. More specifically, error outliers include outlying observations that are caused by not being part of the population of interest (i.e., an error in the sampling procedure), lying outside the possible range of values, errors in observation, errors in recording, errors in preparing data, errors in computation, errors in coding, or errors in data manipulation.
3. Interesting outliers	Accurate (i.e., nonerror) data points that lie at a distance from other data points and may contain valuable or unexpected knowledge.
4. Discrepancy outliers	Data points with large residual values, with possibly (but not necessarily) large influence on model fit and/or parameter estimates.
5. Model fit outlier	An influential outlier whose presence influences the fit of the model.
6. Prediction outlier	An influential outlier whose presence affects the parameter estimates of the model.
7. Influential meta-analysis effect size outlier	A data point that is unusually large or small compared to others in a meta-analytic database, specifically regarding the size of the effect or relationship.
8. Influential meta-analysis sample size outliers	In the context of a meta-analysis, these are single construct outliers in terms of their sample size compared to the other studies' sample sizes.

Identification techniques

Table 2. Outlier Identification Techniques Based on a Review of Methodological and Substantive Organizational Science Sources.

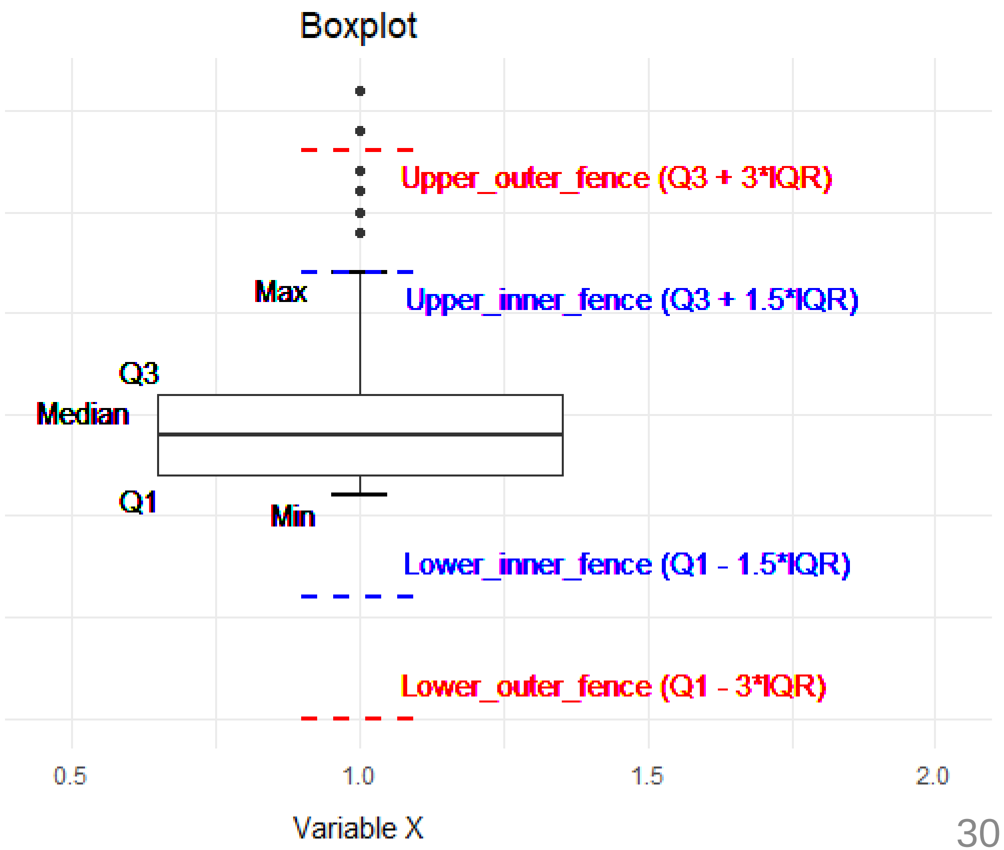
Single-construct techniques	
1. Box plot	A plot that depicts a summary of the smallest value of a construct (excluding outliers), lower quartile (Q1), median (Q2), upper quartile (Q3), and largest value (excluding outliers). Outliers can be identified as those points that lie beyond the plot's whiskers (i.e., the smallest and largest values, excluding outliers).
2. Stem and leaf plot	A plot that simultaneously rank-orders quantitative data and provides insight about the shape of a distribution. Stem-and-leaf pairs that are substantially far away from the rest of the pairs signal the presence of outliers.
3. Schematic plot analysis	Similar to a box plot, but used specifically for effect sizes in the context of a meta-analysis.
4. Standard deviation analysis	Distance of a data point from the mean in standard deviation units.
5. Percentage analysis	Relative standing of a data point in a distribution of scores as indexed by its percentile.
Multiple-construct (i.e., "distance") techniques	
6. Scatter plot	A plot of the values of two variables, with one variable on the x-axis (usually the independent variable) and the other variable on the y-axis (usually the dependent variable). A potential outlier can be identified by a data point lying far away from the centroid of data.
7. q-q plot	A plot (q stands for quantile) that compares two probability distributions by charting their quantiles against each other. A nonlinear trend indicates the possible presence of outlier(s).
8. p-p plot	A plot (p stands for probability) that assesses the degree of similarity of two data sets (usually the observed and expected) by plotting their two cumulative distribution functions against each other. A nonlinear trend indicates the possible presence of outlier(s).

Outliers - Boxplot

- An observation that lies an abnormal distance from other values in a random sample from a population.

- **Mild Outliers**:
  - ✓ Lower Inner fence: x < Q1 – 1.5*IQR
  - ✓ Upper Inner fence: x > Q3 + 1.5*IQR

- **Extreme Outliers**
  - ✓ Lower outer fence: x < Q1 – 3*IQR
  - ✓ Upper outer fence: x > Q3 + 3*IQR



Outliers - Practice

n=12: 1, 4, **6, 9**, 15, 21, 22, 27, **35, 40**, 41, 56

- Q1 = 7.5
- Q3 = 37.5
- IQR = Q3 – Q1 = 30
- [Q1 - 1.5(IQR) , Q3 + 1.5(IQR)] = [7.5 - 1.5*30 , 37.5 + 1.5*30] = **[-37.5 , 82.5]**

Handling Techniques

Table 3. Outlier Handling Techniques Based on a Review of Methodological and Substantive Organizational Science Sources.

1. Correct value	Correcting a data point to its proper value.
2. Remove outlier	Elimination of the data point from the analysis.
3. Study the outlier in detail	Conducting follow-up work to study the case as a unique phenomenon of interest.
4. Keep	Acknowledging the presence of an outlier, but doing nothing to the outlier value prior to the analysis.
5. Report findings with and without outliers	Reporting substantive results with and without the outliers—which also includes providing an explanation for any difference in the results.
6. Winsorization	Transforming extreme values to a specified percentile of the data. For example, a 90th percentile Winsorization would transform all the data below the 5th percentile to the 5th percentile, and all the data above the 95th percentile would be set at the 95th percentile.
7. Truncation	Setting observed values within a believable range and eliminating other values from the data set.
8. Transformation	Applying a deterministic mathematical function (e.g., log function, ln function) to each value to not only keep the outlying data point in the analysis and the relative ranking among data points, but also reduce the error variance and skew of the data points in the construct.

Describing variability

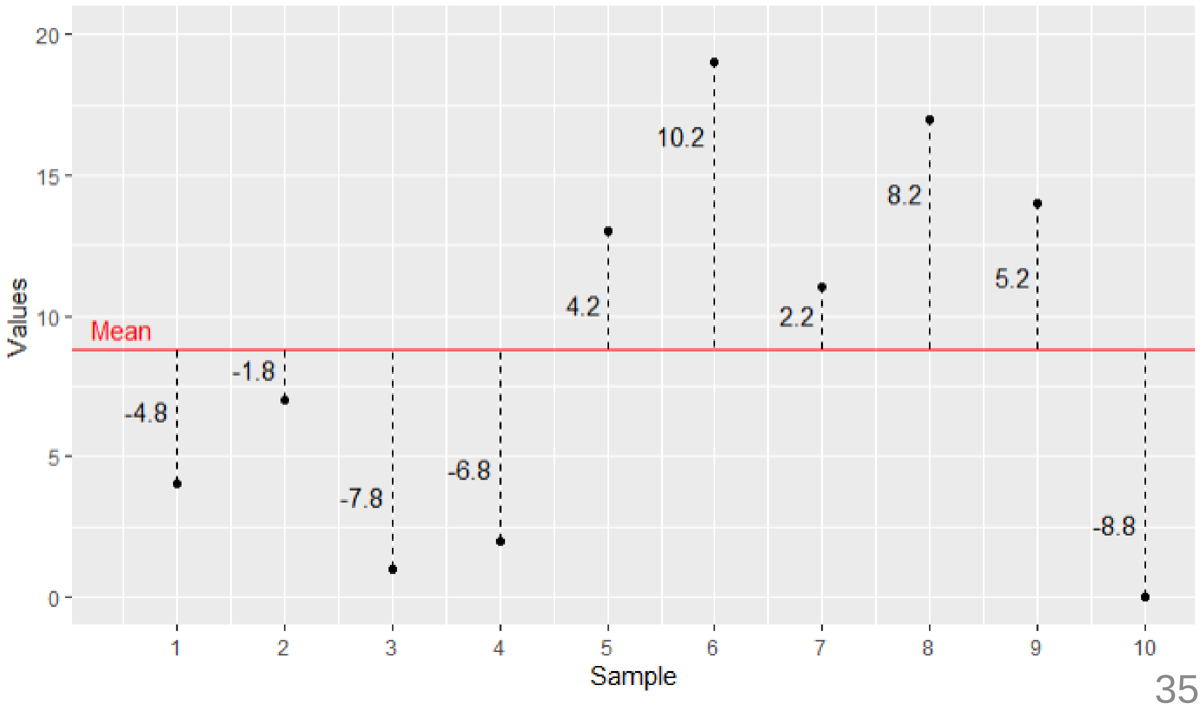
Assessing the fit of mean

Describing variability
Assessing the fit of mean

1. “Deviation from the mean”
2. Sum of squares (SS)
3. Variance
4. Standard Deviation

Deviation from the mean

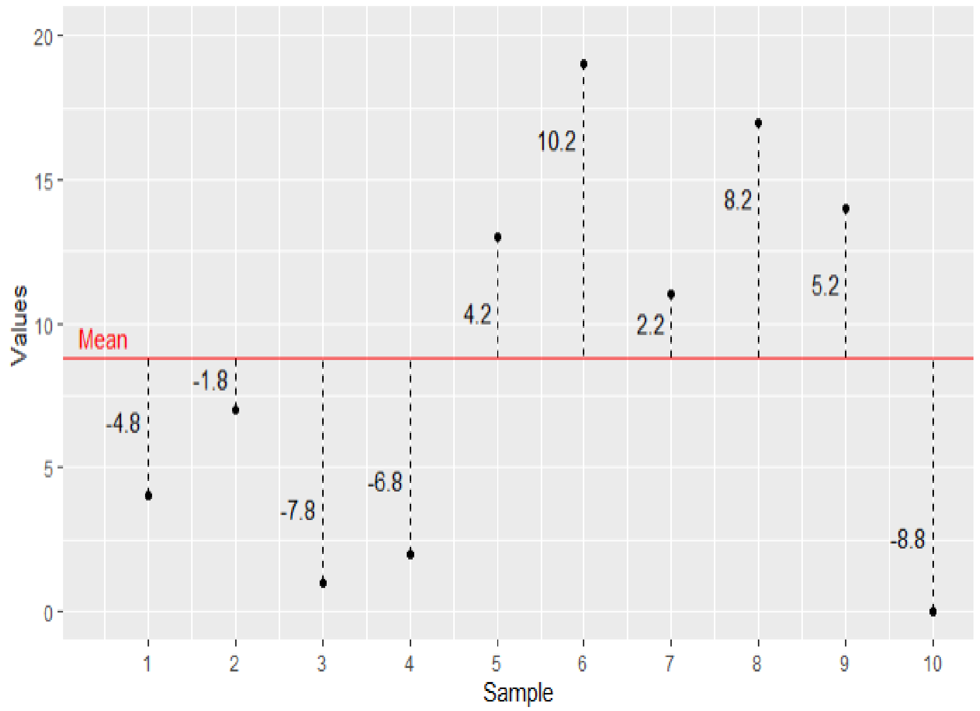
- **Mean:**
  - ✓ An estimate of the central tendency of the data.
  - ✓ A reference point for measuring deviation of individual data points from the average.
- Dash line:
  - ✓ "**Deviation from the mean**"



Sum of squares (SS)

- Why not Sum **"Deviation from the mean"**?
  - ✓ Positive value + negative value = 0

- **Sum of squares (SS)**: $\Sigma(x_i - \overline{x})^2$
  - ✓ Drawback: Data points ↑ , SS ↑
    - ➢ Average of Sum of squares



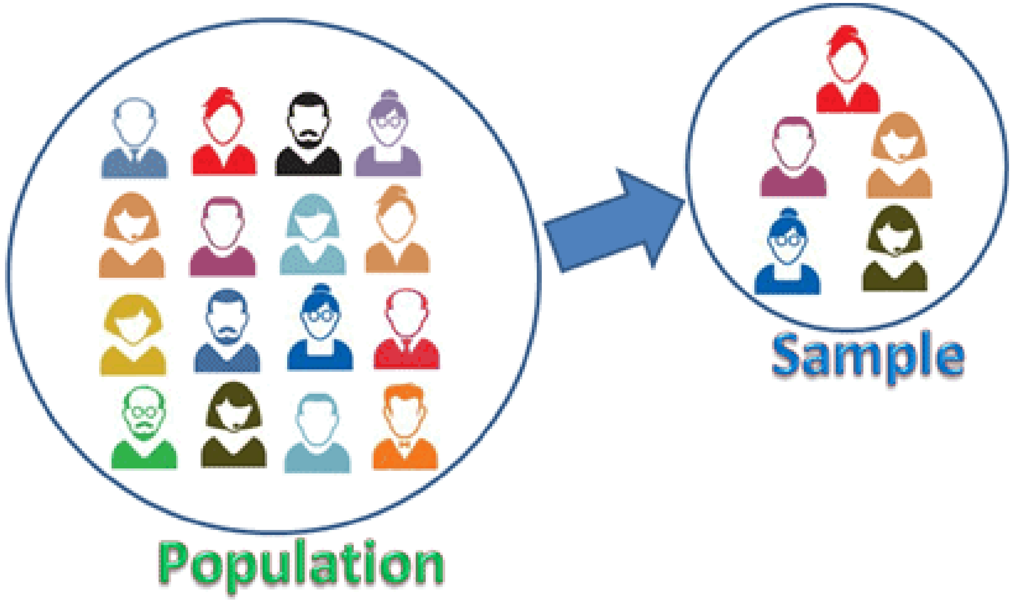
Variance - Concept

- Average error between the mean and the observations made.

- **Formula ?** $$\frac{\Sigma(x_i - \bar{x})^2}{N}$$

Variance - Example

- Assume the average weight of **entire population** in A: 70 Kg.
  - ✓ In fact, we don't know this.
- Measure weight of 10 people: **sample** with 10 people



Variance - Example

Not know true **population** mean:

- **Sample** mean is 77 Kg.
- Variance = $\frac{\Sigma(x_i - \bar{x})^2}{N}$ = **201.8**

$\mu = 70$	x_i	x_i - x	$(x_i - x)^2$
Observation 1	72	-5	25
Observation 2	64	-13	169
Observation 3	72	-5	25
Observation 4	102	25	625
Observation 5	65	-12	144
Observation 6	89	12	144
Observation 7	55	-22	484
Observation 8	97	20	400
Observation 9	78	1	1
Observation 10	76	-1	1
Mean (x)	77		
Variance			201.8

Variance - Example

If known true **population** mean:

- **Population** mean = 70 Kg.

➢ Variance = $\frac{\Sigma(x_i - \bar{x})^2}{N}$ = **250.8**

➢ **201.8** ≠ **250.8**

➢ Due to **bias**.

$\mu = 70$	x_i	$x_i - x$	$(x_i - x)^2$		$x_i - \mu$	$(x_i - \mu)^2$
Observation 1	72	-5	25		2	4
Observation 2	64	-13	169		-6	36
Observation 3	72	-5	25		2	4
Observation 4	102	25	625		32	1024
Observation 5	65	-12	144		-5	25
Observation 6	89	12	144		19	361
Observation 7	55	-22	484		-15	225
Observation 8	97	20	400		27	729
Observation 9	78	1	1		8	64
Observation 10	76	-1	1		6	36
Mean (x)	77					
Variance			201.8			250.8
Variance (sample)			224.2222			

➢ Variance less than what it should be if population mean is considered.

➢ **Bessel correction**

Variance - Bessel correction

- Only when population mean unknown.
- Variance and standard deviation.
- **Variance of sample =** $\frac{\Sigma(x_i - \bar{x})^2}{N - 1}$
  - ✓ Decrease the denominator.
  - ✓ Increase Variance value.

μ = 70	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$		$x_i - \mu$	$(x_i - \mu)^2$
Observation 1	72	-5	25		2	4
Observation 2	64	-13	169		-6	36
Observation 3	72	-5	25		2	4
Observation 4	102	25	625		32	1024
Observation 5	65	-12	144		-5	25
Observation 6	89	12	144		19	361
Observation 7	55	-22	484		-15	225
Observation 8	97	20	400		27	729
Observation 9	78	1	1		8	64
Observation 10	76	-1	1		6	36
Mean ($\bar{x}$)	77					
Variance			201.8			250.8
Variance (sample)			224.2222			

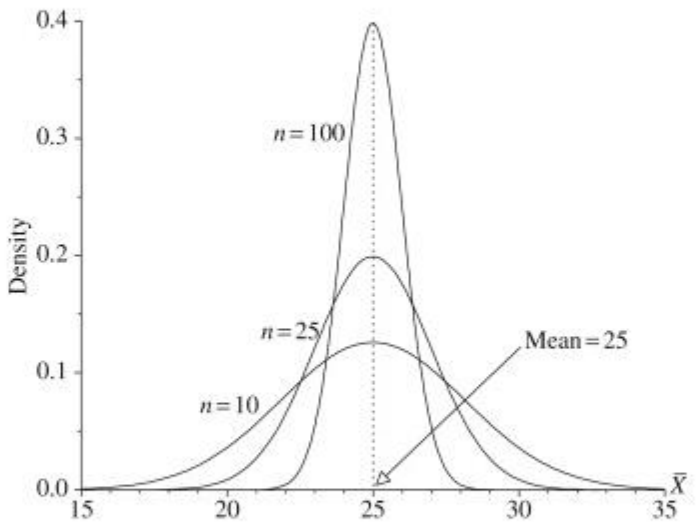
Variance of Sample

- Average error between the mean and the observations made.
- **Variance of Sample =** $\frac{\Sigma(x_i - \bar{x})^2}{N - 1}$
- Drawback:
  - ✓ Squaring
    - ➢ Cannot compare variance to individual data points.
    - ➢ Square root of the variance.

Standard Deviation (SD)

- Square root of the variance: $SD = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$

- How spread out the data points are on either side of the mean.
  - ✓ Small SD: less variability = more agreement in the data.



Mean – SD – Median – IQR

Affected by Extreme data	Less affected by Extreme data
Mean	Median
SD	IQR

Descriptive Table

Normal distribution data	Non-normal distribution data
Mean ± SD	**Median (IQR)**

Table 3. Characteristics of groups B and D of reclassified *versus* non-reclassified patients according to GOLD 2017 criteria.

Subjects	DB n=61	DD n=218	p-value	BB n=95	p-value
Age (years), mean ± SD	62.4 ± 9.68	65.5 ± 9.82	0.026	68.1 ± 10.5	0.001
Male, n (%)	57 (93.4%)	206 (94.5%)	0.83	88 (92.6%)	0.883
BMI°	21.1 (4.45)	20.6 (5.30)	0.773	21.4 (4.34)	0.226
FVC% pred°	60.0 (13.0)	66.5 (23.0)	0.001	80.0 (21.0)	<0.001
FEV_1%pred°	39.0 (12.0)	44.0 (19.8)	<0.001	61.0 (17.0)	<0.001
FEV_1/FVC°	48.0 (10.0)	52.0 (11.0)	0.002	58.0 (10.0)	<0.001
GOLD_therapy:			<0.001		0.073
LABA	16 (26.2%)	8 (3.67%)		43 (45.3%)	
LABA + ICS	10 (16.4%)	37 (17.0%)		11 (11.6%)	
LAMA	3 (4.92%)	7 (3.21%)		8 (8.42%)	
LAMA + LABA	26 (42.6%)	76 (34.9%)		30 (31.6%)	
LAMA + LABA + ICS	5 (8.20%)	88 (40.4%)		2 (2.11%)	
SABA/ SABA + SAMA	1 (1.64%)	2 (0.92%)		1 (1.05%)	

°: Non-normal distribution data were expressed as medians (interquartile range). DB: patients reclassified from group D to B; DD: patients remained in group D; BB: patients remained in group B; SD, standard deviation; BMI, body mass index; FVC% pred, percentage of forced vital capacity; FEV_1%pred, percentage of forced expiratory volume in the first second; LABA, long-acting beta agonists; ICS, inhaled corticosteroid; LAMA, long-acting muscarinic antagonist; SABA, short-acting beta agonists; SAMA, short-acting muscarinic antagonist.

References

1. Aguinis H, Gottfredson RK, Joo H. Best-practice recommendations for defining, identifying, and handling outliers. Organizational research methods. 2013 Apr;16(2):270-301.
2. Diez DM, Barr CD, Cetinkaya-Rundel M. OpenIntro statistics. Boston, MA, USA:: OpenIntro; 2012.
3. Schwertman NC, Owens MA, Adnan R. A simple more general boxplot method for identifying outliers. Computational statistics & data analysis. 2004 Aug 1;47(1):165-74.
4. Witte RS, Witte JS. Statistics. John Wiley & Sons; 2017.

THANK YOU FOR LISTENING