# Grip on Graphs

**Rajesh Kayala, Senior Statistical Programmer**
**Hyderabad**
**June 17, 2017**

NOVARTIS

# Disclaimer:

All opinions expressed in this presentation are the presenter's personal views, and do not reflect the views or opinions of Novartis.

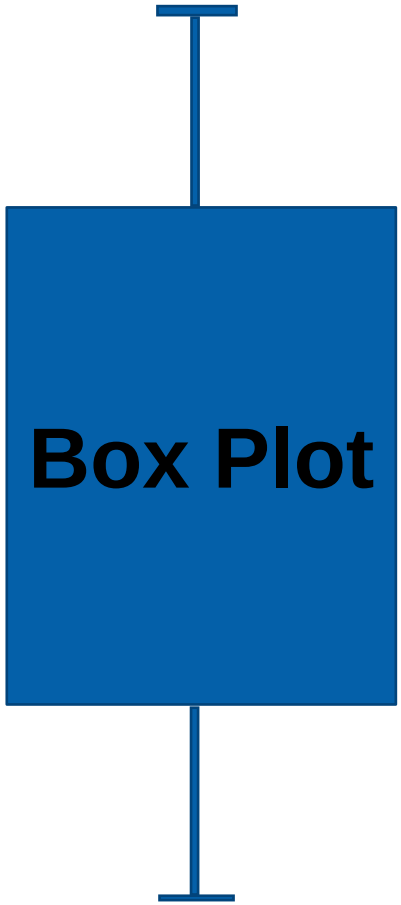
# Introduction

The expectation of this presentation is,

- Basic understanding of graphs.
- Usage of graphs.
- Interpretation of graphs.

# Types of Graphs

- **Box Plot**
- **Scatter Plot**

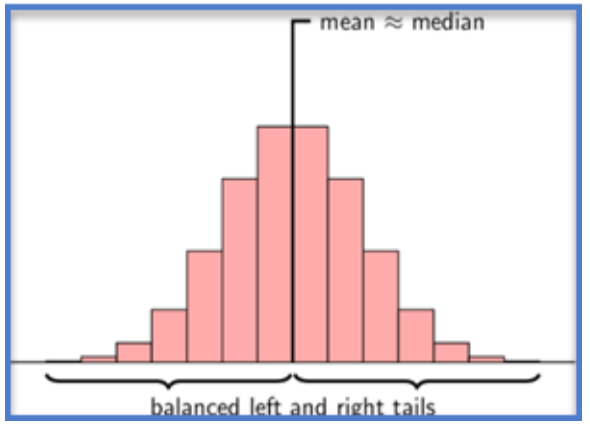





NOVARTIS

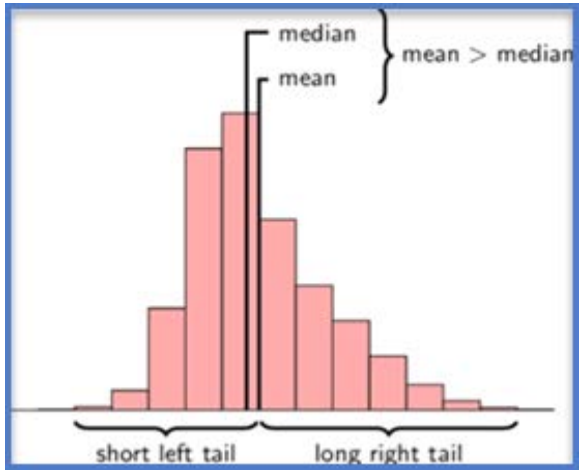
# Shape of the data

- Symmetric/ Normal Distribution : A symmetric distribution is one where the left and right hand sides of the distribution are roughly equally balanced around the mean. The Mean is approximately equal to the Median.




- Right skewed distribution : The mean is typically greater than the median and also know as **Positive Skewed**.

- Left skewed distribution : The mean is typically less than the median and also know as **Negative Skewed**. It is exactly opposite to the right skewed distribution.






NOVARTIS

# Box Plot

## What is Box Plot?

- A box plot is a graphical representation of continuous data based on the minimum, first quartile (Q1), median, third quartile (Q3), and maximum (also called as 5 number statistics or summary statistics).

# Box Plot

## Advantages of Box-plot

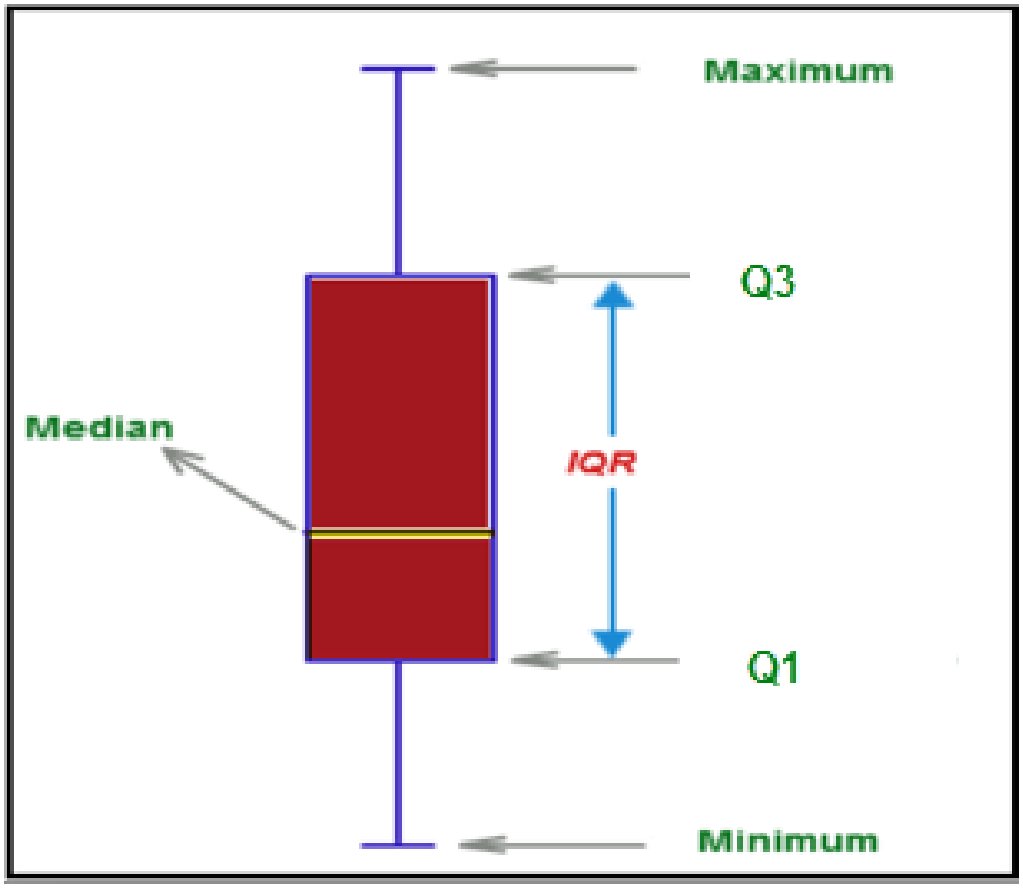
- Box-Plot is able to handle and present summary of a large amount of data.
- Box plot will give you the idea about the shape of the distribution (symmetric or skewed).
- A box plot is a highly visually effective way of viewing a clear summary of one or more sets of data. It is particularly useful for quickly summarizing and comparing different sets of results.
- Box plots can help to identify the presence of outliers.

# Box Plot

## Disadvantages of Box-plot

- Original data is not clearly shown in the box plot (It shows only summary of the data).
- Mode cannot be identified in a box plot.
- They will be used only with numerical data.

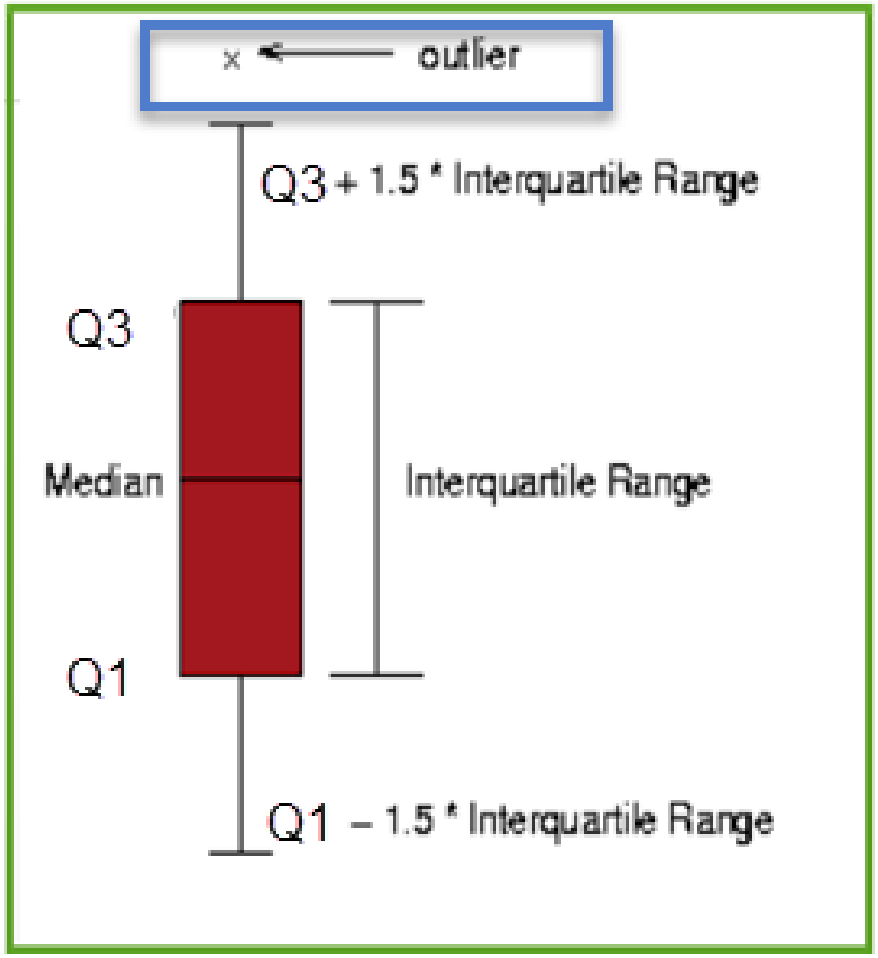
# Composition of Box and whisker plot






NOVARTIS

# Way to Representing outliers




**Q3 + 1.5 ( IQR )**

**Q1 - 1.5 ( IQR )**

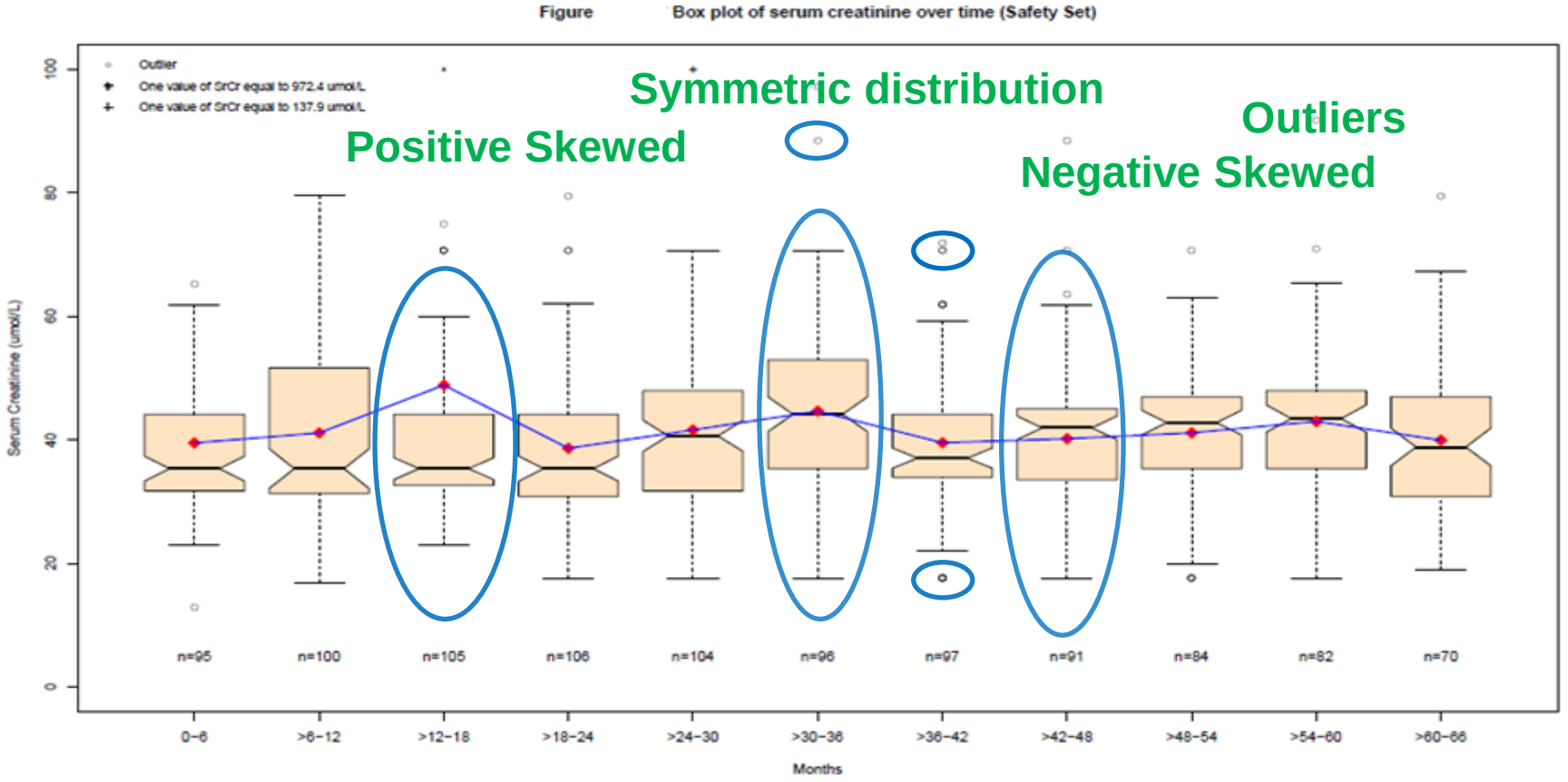
**IQR (Inter Quartile Range) = Q3 – Q1;**



NOVARTIS

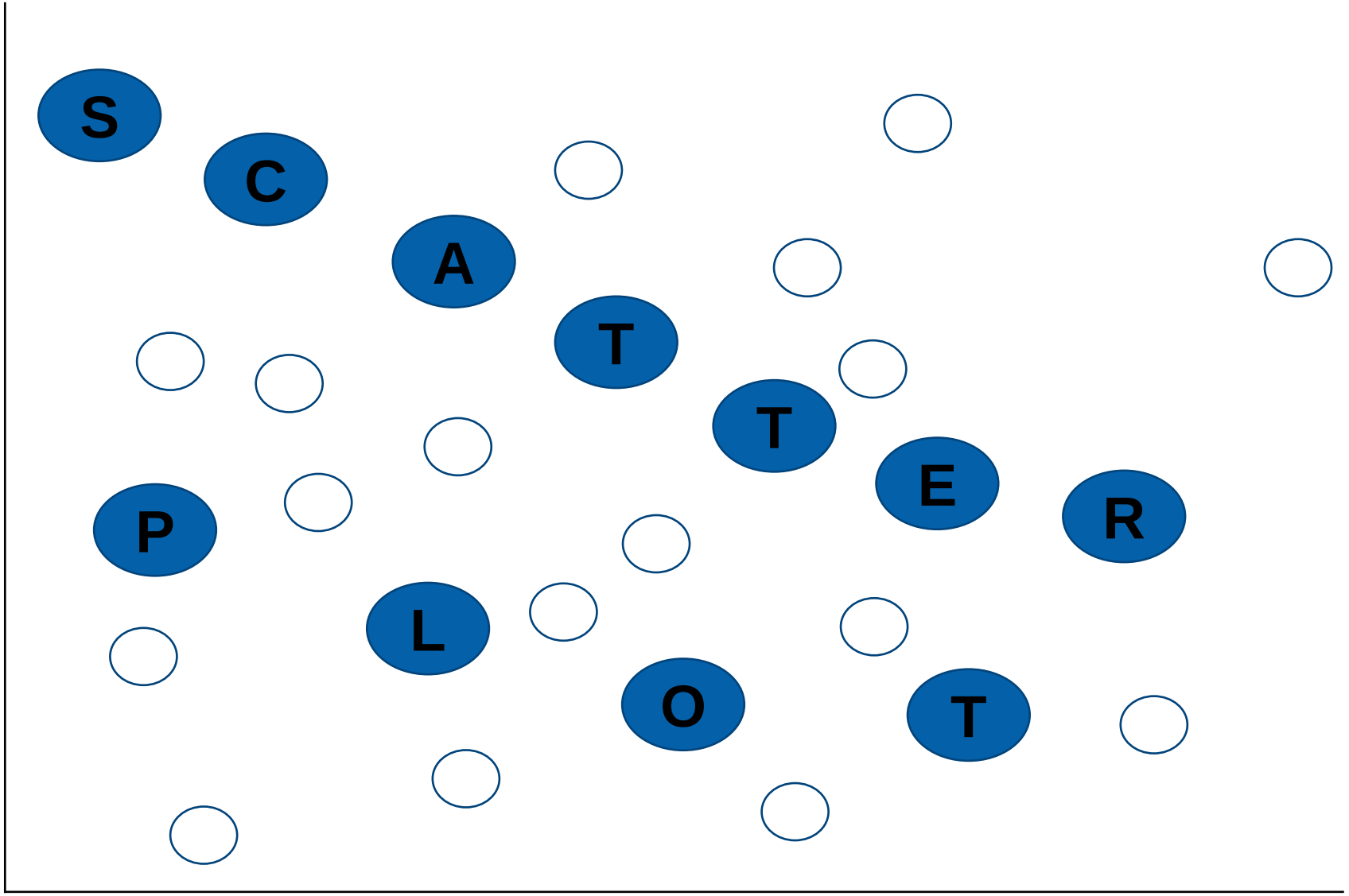
# Interpreting Box plots

**Example :** The below graph is generated to see the Serum Creatinine across the different time points.






NOVARTIS






NOVARTIS

# Scatter plot

- A scatter plot is a plot of data points on a horizontal and vertical axis to show how much one variable is affected by another and gives good visual picture of the relationship.
- The relationship between two variables is called **Correlation**.
- Continuous data should be used to plot these.

- X axis is the quantitative ***Explanatory*** variable.
  - An **explanatory variable** is one that explain changes in that variable. Ex: Winter season in Months.

- Y axis is the quantitative ***Response*** variable.
  - The **response variable** is the focus of a question in a study or experiment. Ex: Ice cream sales.



NOVARTIS

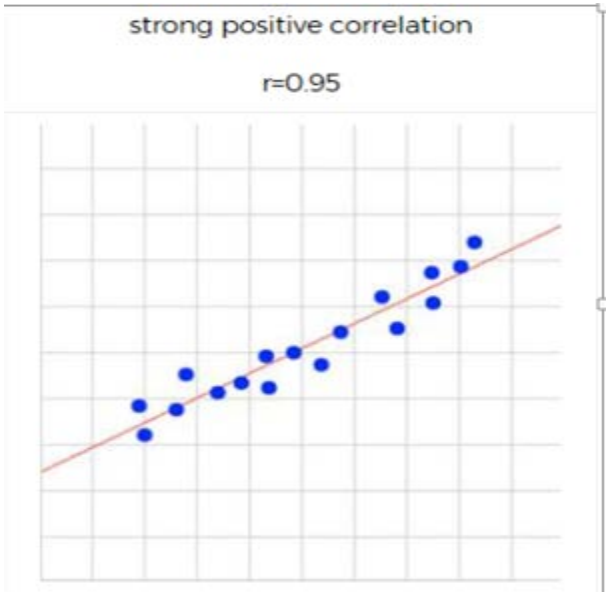
# Correlation

- Correlations have two properties: strength and direction. The **strength** of a correlation is determined by its numerical value. The **direction** of the correlation is determined by whether the correlation is positive or negative.
- Strength of a correlation is measured by **Correlation coefficient (*r*).**
  - Always between −1 and +1 (inclusive)
    - r = +1 ⇒ all points on upward sloping line
    - r = - 1 ⇒ all points on downward line
    - r = 0 ⇒ no line or horizontal line
  - The closer ***r*** is to +1 or –1, the **stronger** the correlation

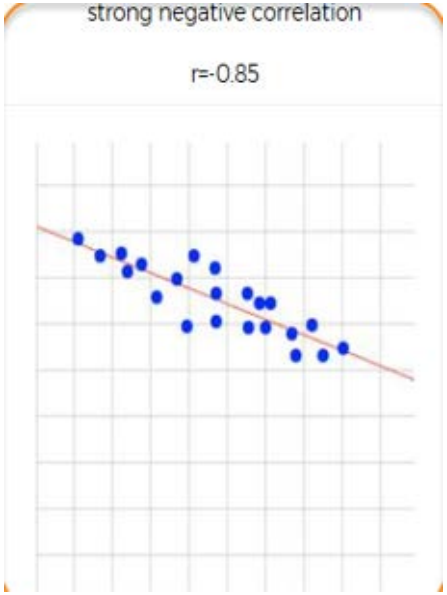
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

# Types of Correlation

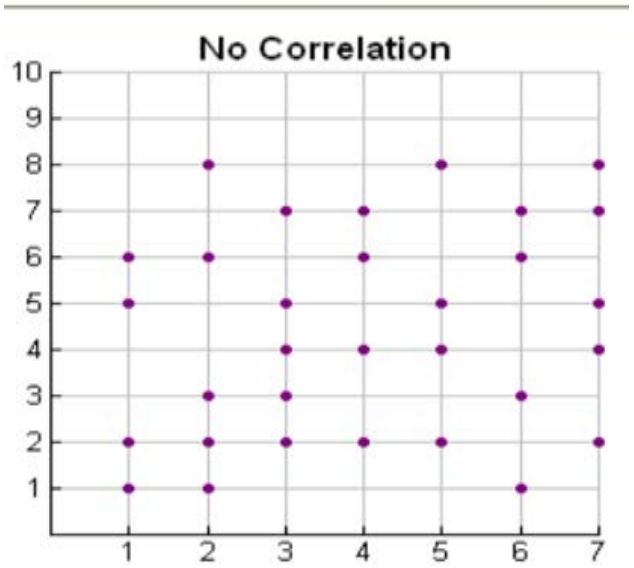
## Positive correlation




## Negative Correlation




## No Correlation

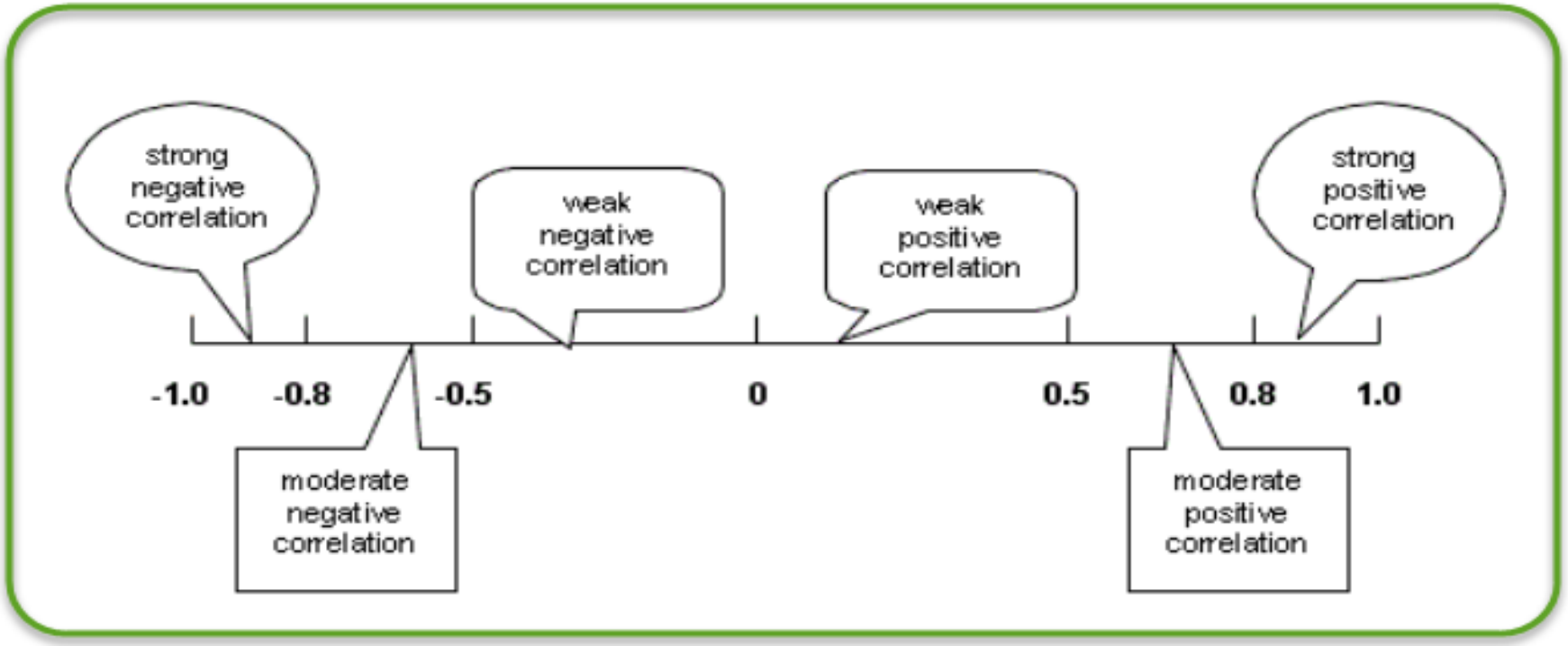





NOVARTIS

# Correlation coefficient (cont.)

- The following diagram helps us to understand the relationship between the data based on **r** value and the word descriptors.






NOVARTIS

# Calculating *r*

**Example:**

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

| Subject | Age (X) | Glucose Level (Y) | XY | $X^2$ | $Y^2$ |
|---|---|---|---|---|---|
| 1 | 43 | 99 | 4257 | 1849 | 9801 |
| 2 | 21 | 65 | 1365 | 441 | 4225 |
| 3 | 25 | 79 | 1975 | 625 | 6241 |
| 4 | 42 | 75 | 3150 | 1764 | 5625 |
| 5 | 57 | 87 | 4959 | 3249 | 7569 |
| 6 | 59 | 81 | 4779 | 3481 | 6561 |
| **Σ** | **247** | **486** | **20485** | **11409** | **40022** |

# Calculating *r* (cont.)

From the table:

$\Sigma x = 247$

$\Sigma y = 486$

$\Sigma xy = 20{,}485$

$\Sigma x^2 = 11{,}409$

$\Sigma y^2 = 40{,}022$

sample size (n) = 6

**Correlation coefficient (r) =**

$6(20{,}485) - (247 \times 486) / [\sqrt{[[6(11{,}409) - (247^2)] \times [6(40{,}022) - 486^2]]}]$

$= 0.5298$ $\Rightarrow$ moderate positive correlation

# Line of Best Fit

- A **line of best fit** (or 'trend' line) is a straight line that best represents the data.
- A Line of Best Fit does not touch all of the data points.
- A Line of Best Fit is useful because it allows us to:
  - Understand the type and strength of the relationship between two sets of data.
  - Predict missing Y values for given X values.
- The equation of a line of best fit using a method called **linear regression.**
- A more accurate way of finding the line of best fit is the **least square method**.

# Line of Best Fit (cont.)

- Use the following steps to find the equation of line of best fit.

Step 1: Plot the data values and calculate the mean of x and y -values.

Step 2: The following formula gives the slope of the line of best fit:

$$m = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}$$

Step 3: Compute the **y -intercept** of the line by using the formula:

$$b = Y - m X$$

Step 4: Use the slope (m) and the y -intercept (b) to form the equation of the line.

$$Y = b + m (x)$$

# Line of Best Fit (cont.)

| Age (X) | Height (Y) |
|---|---|
| 18 | 76.1 |
| 19 | 77 |
| 20 | 78.1 |
| 22 | 78.8 |
| 23 | 79.7 |
| 24 | 79.9 |
| 25 | 81.1 |
| 26 | 81.2 |
| 27 | 82.8 |
| 29 | 83.5 |

$\overline{X}$ = 23.3 ; $\overline{Y}$= 79.82
Slope (m) = 0.6618
y –intercept (b) = 64.399

- The slope (m) of the line is **0.6618** and
- The y –intercept (b) is **64.399**
- Therefore, the equation is

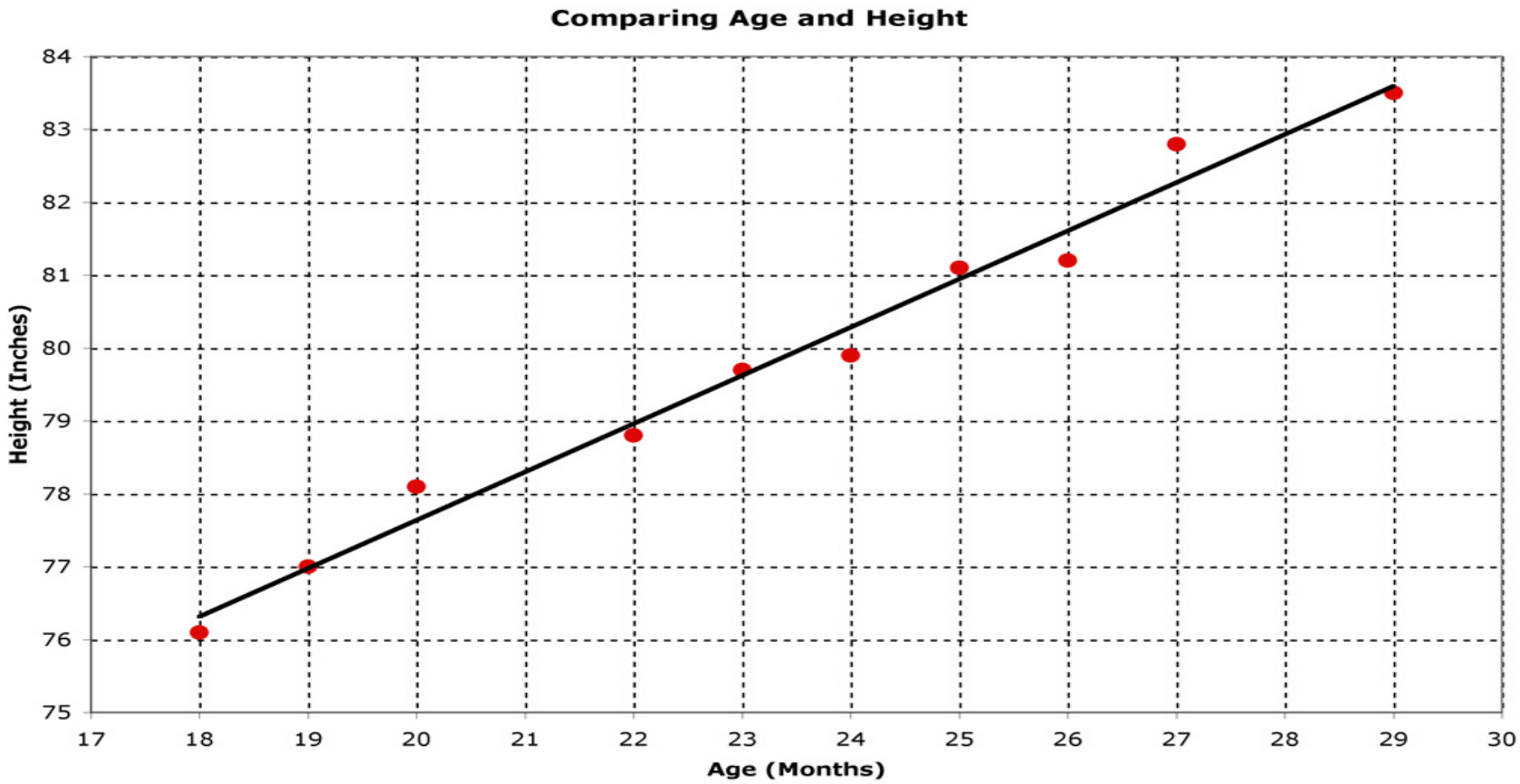
$$Y = b + m(x)$$

**Y= 64.399 + 0.6618(X)**



NOVARTIS

# Line of Best Fit (cont.)






NOVARTIS

# Predicting Data Points

Estimating values from scatter plots can be done as below.

Interpolation - Making a prediction for an unknown Y value based on a given X value **within** a range of known data.
Ex: Estimating Height (Y) value at Age (X) 28 months. 28months is within the range and close to the pattern.

Extrapolation - Making a prediction for an unknown Y value based on a given X value **outside** of a range of known data.
Ex: Estimating Height (Y) value at Age (X) 240 months. 240 months is too far away to the pattern. Could be consider as **outlier**. Typically it is discouraged.

# Predicting Data with Line of best fit

How to make a prediction?

- We can make a prediction based on Slope and y – intercept values for **Y** when **X** is given.
- Lets predict the Height at the time of Age (months) 21, 28 and 240.

| Age (X) | Height (Y) |
|---|---|
| 18 | 76.1 |
| 19 | 77 |
| 20 | 78.1 |
| 21 | |
| 22 | 78.8 |
| 23 | 79.7 |
| 24 | 79.9 |
| 25 | 81.1 |
| 26 | 81.2 |
| 27 | 82.8 |
| 28 | |
| 29 | 83.5 |

# Predicting Data with Line of best fit cont.

## Equation for Line of Best Fit

$$Y = 0.6618 (X) + 64.399$$

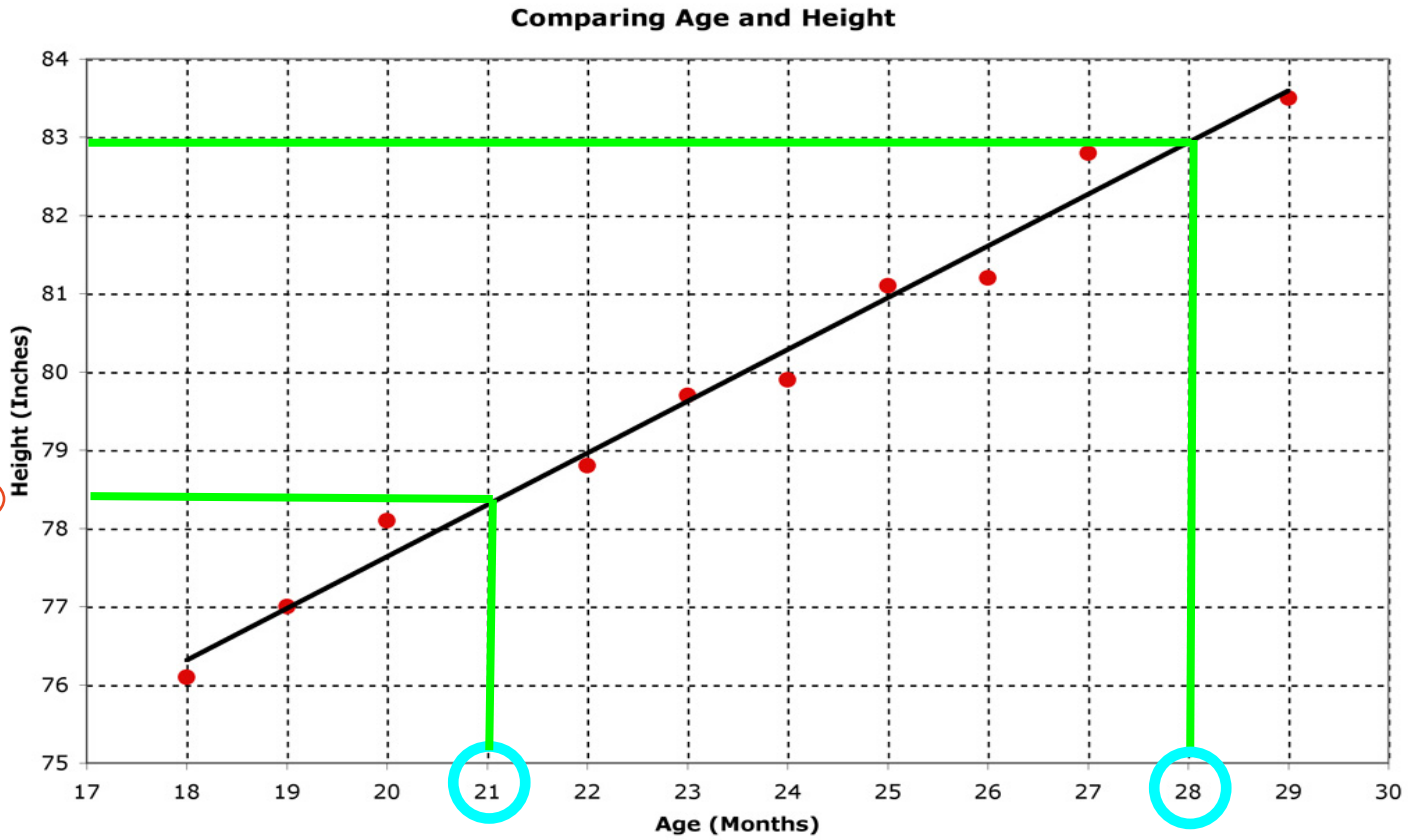
| X (months) | Formula | Y (inches) |
|---|---|---|
| 21 | 0.6618(21) + 64.399 | 78.3 |
| 28 | 0.6618(28) + 64.399 | 82.9 |
| 240 | 0.6618(240) + 64.399 | 223.3 |

# Predicting data with Line of best fit cont.

## Interpolation

82.9

78.3

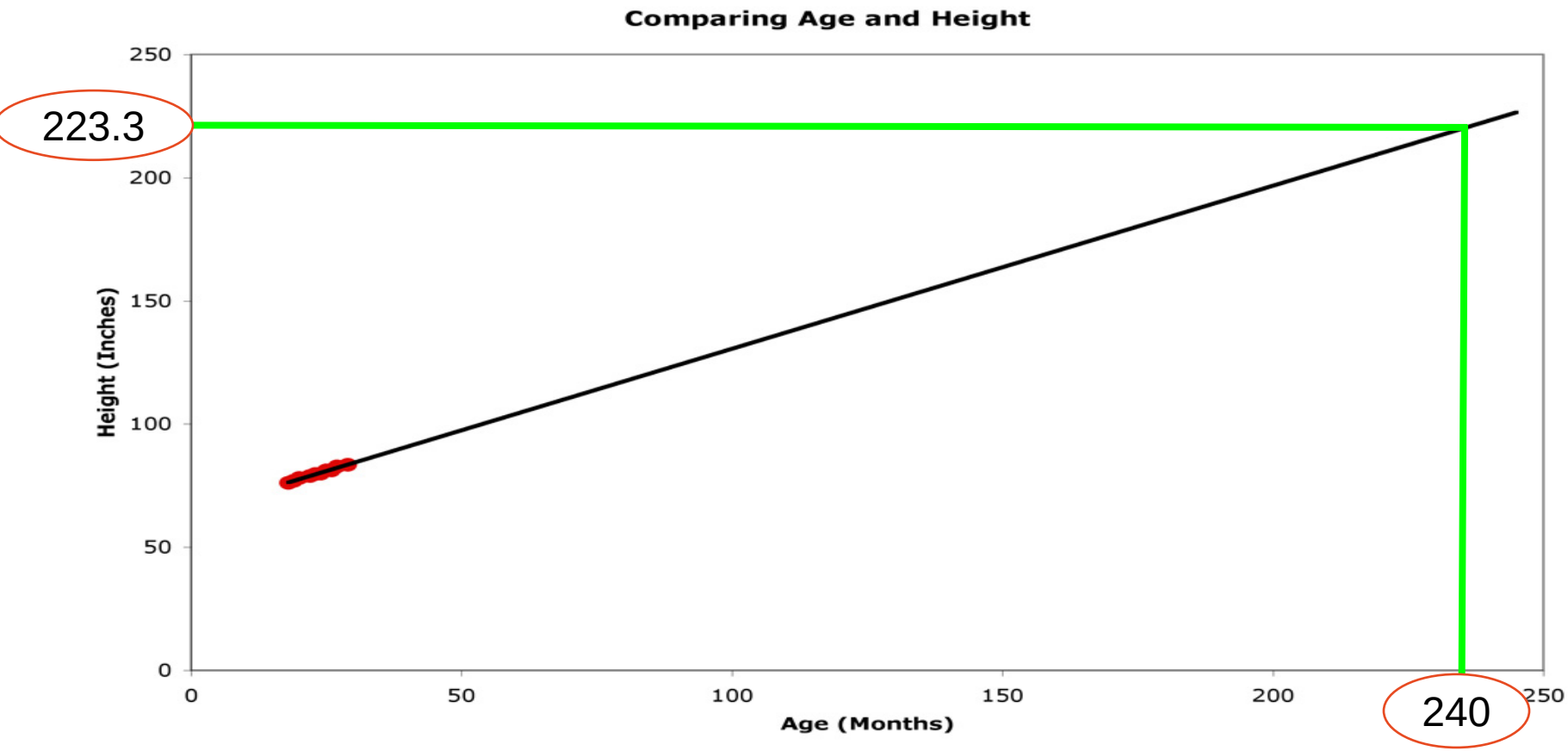





NOVARTIS

# Predicting Data with Line of best fit cont.

## Extrapolation






NOVARTIS

# Interpretation of Scatterplots

We need to keep in mind the below points at the time of interpreting scatter plots.

- **Form:** Can relationship be described by straight line (linear)? Or by a curved line (Non–linear)?
- **Outliers?**: Any deviations from overall pattern?
- **Direction** of the relationship either:
  - Positive association (upward slope)
  - Negative association (downward slope)
  - No association (flat)
- **Strength:** The strength of the pattern is related to how tightly clustered the points are around the underlying form. Strength of the relationship, whether negative or positive?

# Interpretation of Scatterplots

**Example 1:**


**r = 0.9**

**Strength: Positive**

Form : **linear**

Outliers : **No**

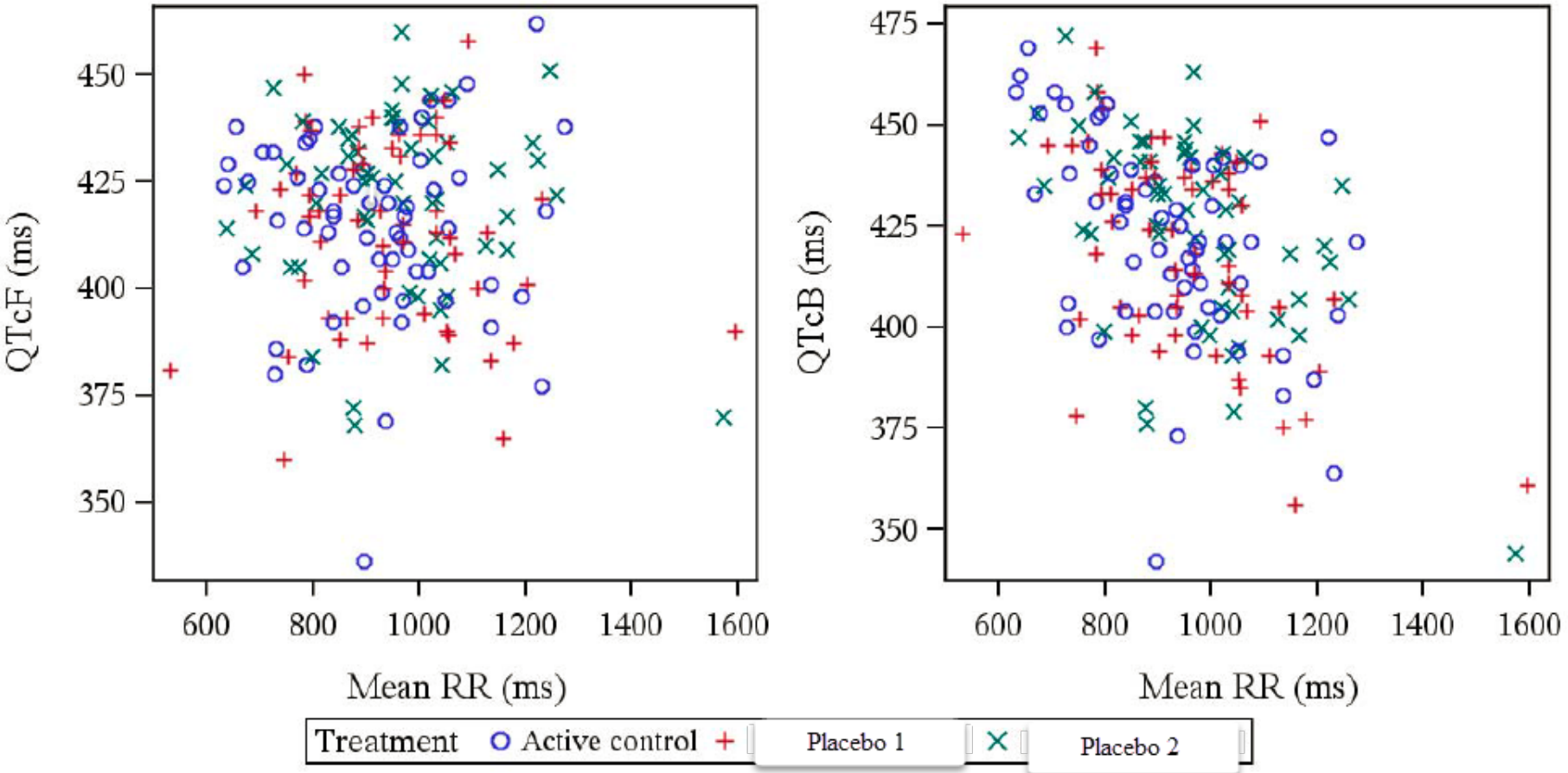
Direction: **Upward Slope**



NOVARTIS

# Interpretation of Scatterplots

Example 2 :






NOVARTIS

# Basic syntax of Box and Scatter plots using SAS

```
PROC SGPLOT DATA=DATASET;
VBOX VARIABLE / category = VARIABLE; /* Vertical Box*/
HBOX VARIABLE / category = VARIABLE; /* Horizontal Box*/
RUN;
```

```
PROC SGPLOT DATA=DATASET;
SCATTER X=variable Y=variable</option(s)>;
RUN;
```

# References


1) http://www.wellbeingatschool.org.nz/information-sheet/understanding-and-interpreting-box-plots

2) http://web.pdx.edu/~stipakb/download/PA551/boxplot.html

3) www.radicalmath.org/Powerpoints/LineOfBestFit.ppt

4) www.sjsu.edu/faculty/gerstman/hs67/Chapt04_BPS.ppt

5) http://www.everythingmaths.co.za/maths/grade-11/11-statistics/11-statistics-05.cnxmlplus




NOVARTIS