AN INTRODUCTION TO DATA ANALYTICS

# Chapter 6
# Statistical Data Analysis

Vi Bảo Ngọc

0983408885 - ngocvb@lqdtu.edu.vn

Computer Science Department, Le Quy Don Technical University

# Outline

1. Distribution fitting
2. Kernel Density Estimation
3. **Determining confidence intervals for mean, variance, and standard deviation**
4. **Exploring extreme values**
5. **Correlating variables with correlation**
6. **Evaluating relationships between variables with ANOVA**

# Determining confidence intervals for mean, variance, and standard deviation

- A confidence intervals are an estimated range usually associated with a certain confidence level quoted in percentages.
- You can calculate confidence intervals for many kinds of statistical estimates, including:
  - Proportions
  - Population means
  - Differences between population means or proportions
  - Estimates of variation among groups

# Determining confidence intervals for mean, variance, and standard deviation

- If you want to calculate a confidence interval on your own, you need to know:
  - The **point estimate** you are constructing the confidence interval for
  - The **critical values** for the test statistic
  - The **standard deviation** of the sample
  - The **sample size**

# Confidence interval for the mean of normally-distributed data

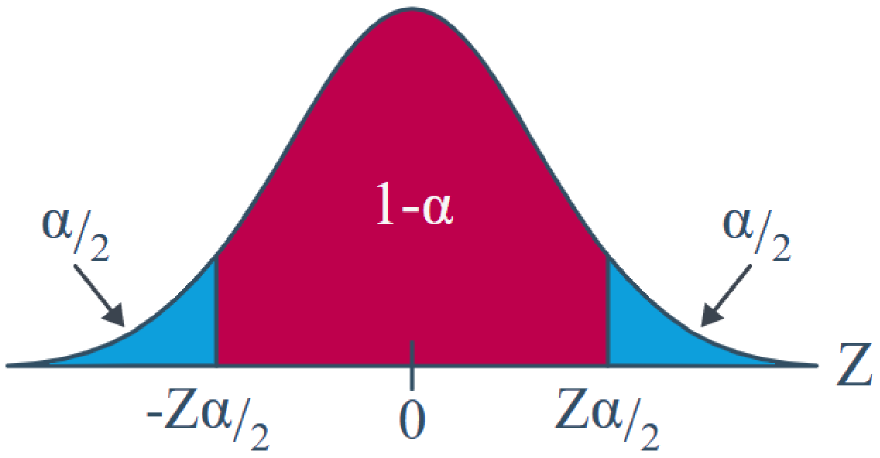
- **Z-interval** for a mean by making the unrealistic assumption that we know the population variance.
- **t-interval** for a mean for the more realistic situation that we don't know the population variance

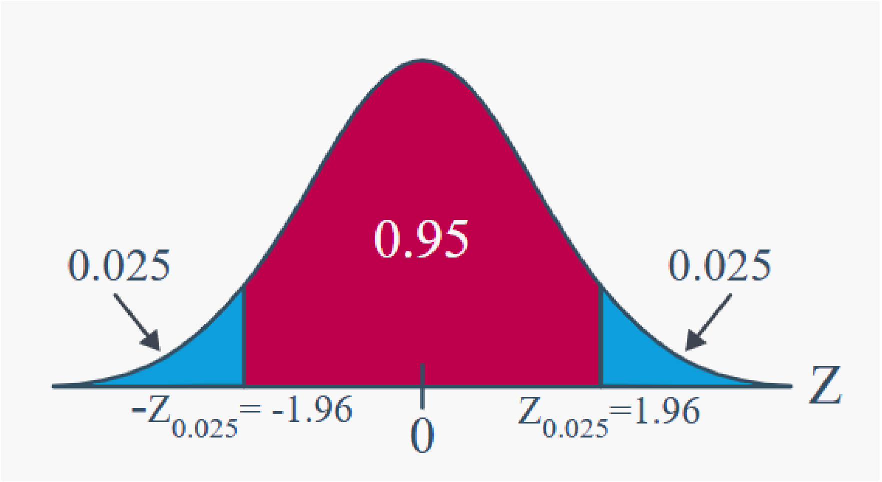
# Confidence interval for the mean of normally-distributed data

- **Z-interval**

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ and } Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$



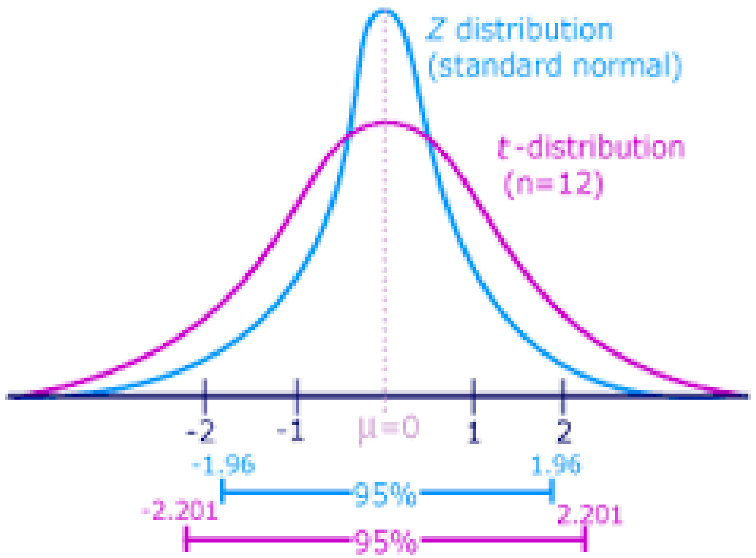
# Confidence interval for the mean of normally-distributed data

- **t-interval**

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$



# Confidence interval for the mean of non-normal data

- When the sample size increases, the ratio:

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

approaches an approximate normal distribution

$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right) \qquad \bar{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

# Examples

1. A random sample of 64 guinea pigs yielded the following survival times (in days):

36; 18; 91; 89; 87; 86; 52; 50; 149; 120; 119; 118; 115; 114; 114; 108; 102; 189; 178; 173; 167; 167; 166; 165; 160; 216; 212; 209; 292; 279; 278; 273; 341; 382; 380; 367; 355; 446; 432; 421; 421; 474; 463; 455; 546; 545; 505; 590; 576; 569; 641; 638; 637; 634; 621; 608; 607; 603; 688; 685; 663; 650; 735; 725

What is the mean survival time (in days) of the population of guinea pigs?

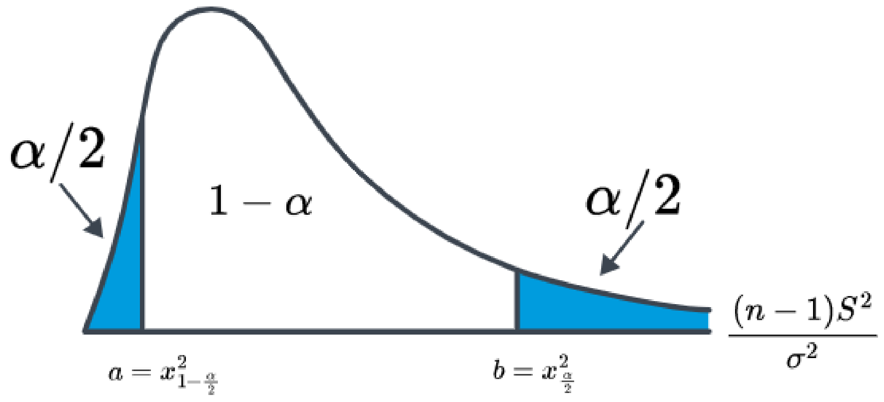
2. Calculate the confidence interval of mean of sepal length in Iris dataset.

# Confidence interval for the variance/std of normally-distributed data

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}$$

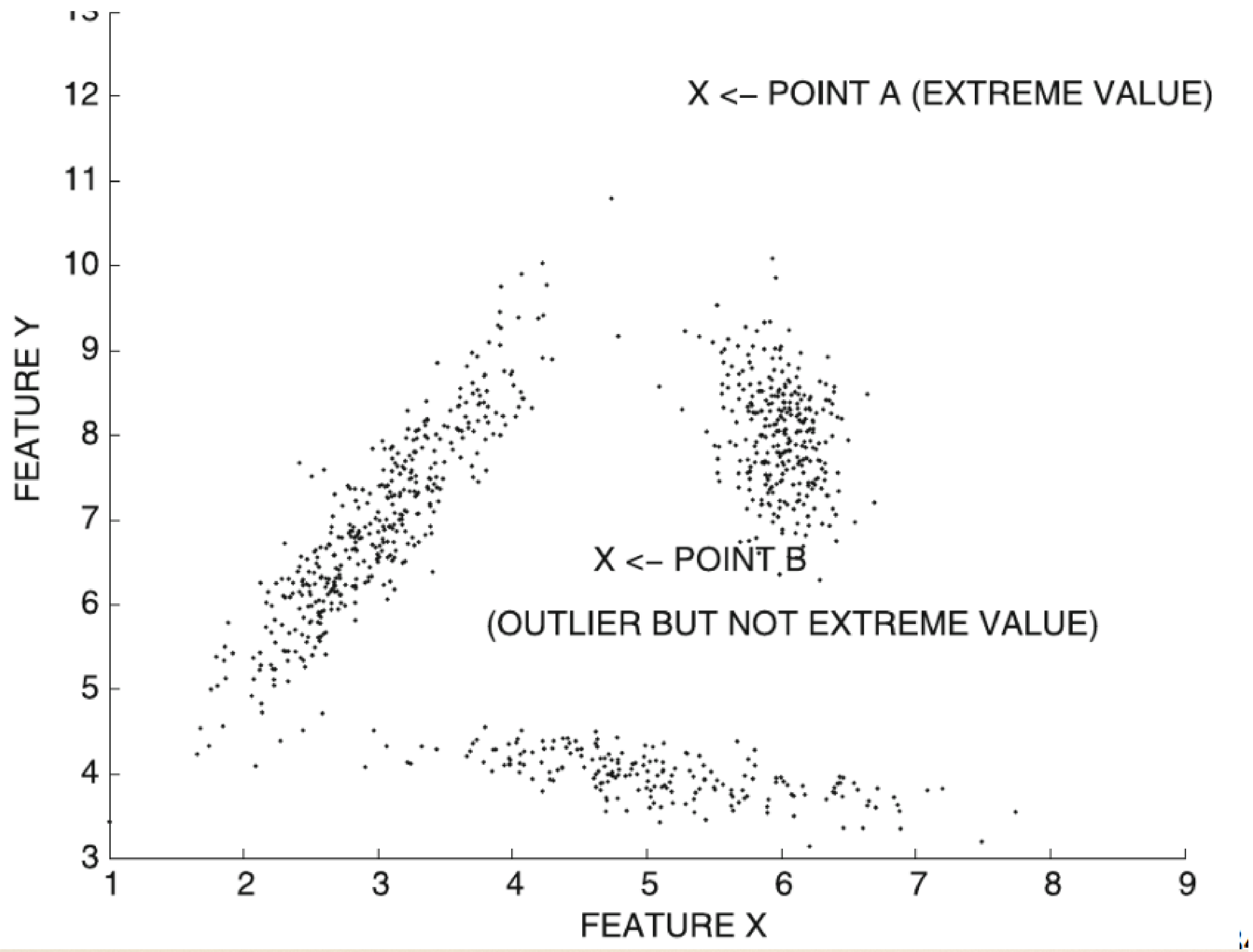
$$\frac{\sqrt{(n-1)S^2}}{\sqrt{b}} \leq \sigma \leq \frac{\sqrt{(n-1)S^2}}{\sqrt{a}}$$



# Extreme value analysis

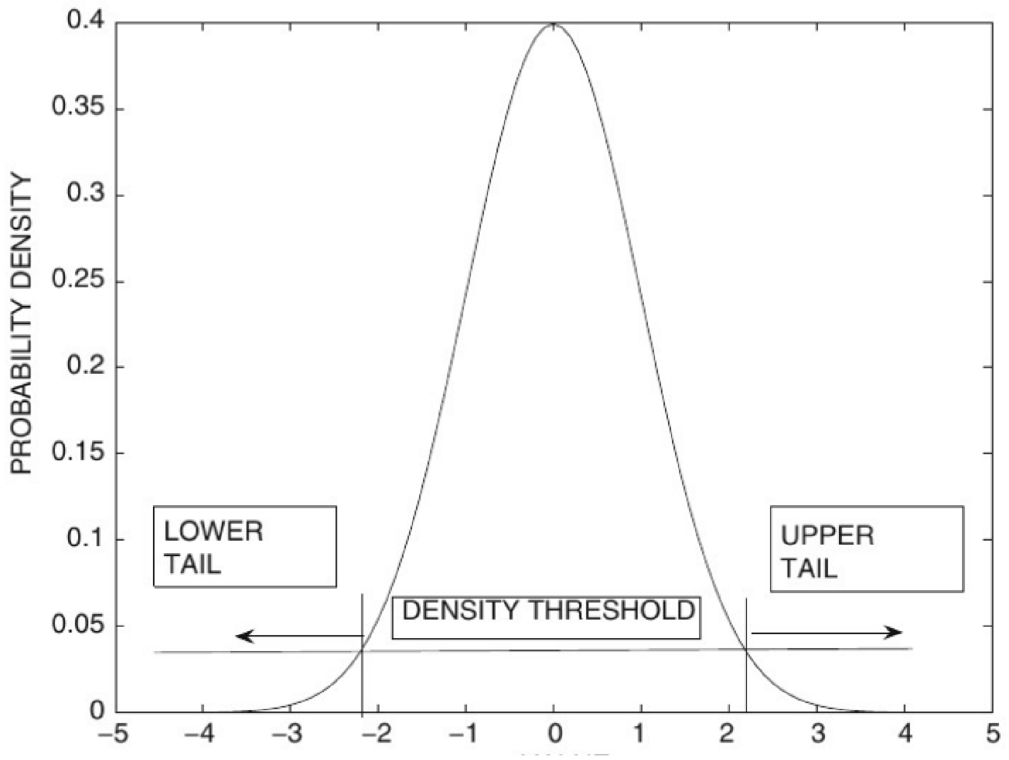
- Extreme value: data point lying at one end of a probability distribution
- Extreme values are specialized types of outliers: All extreme values are outliers, but the reverse may not be true
- Example of univariate extreme values {1,3,3,3,50,97,97,97,100}
  - 1 and 100: extreme values outliers
  - 50 is the mean of the data set not an extreme value
  - 50 is the most isolated point outlier from a generative perspective

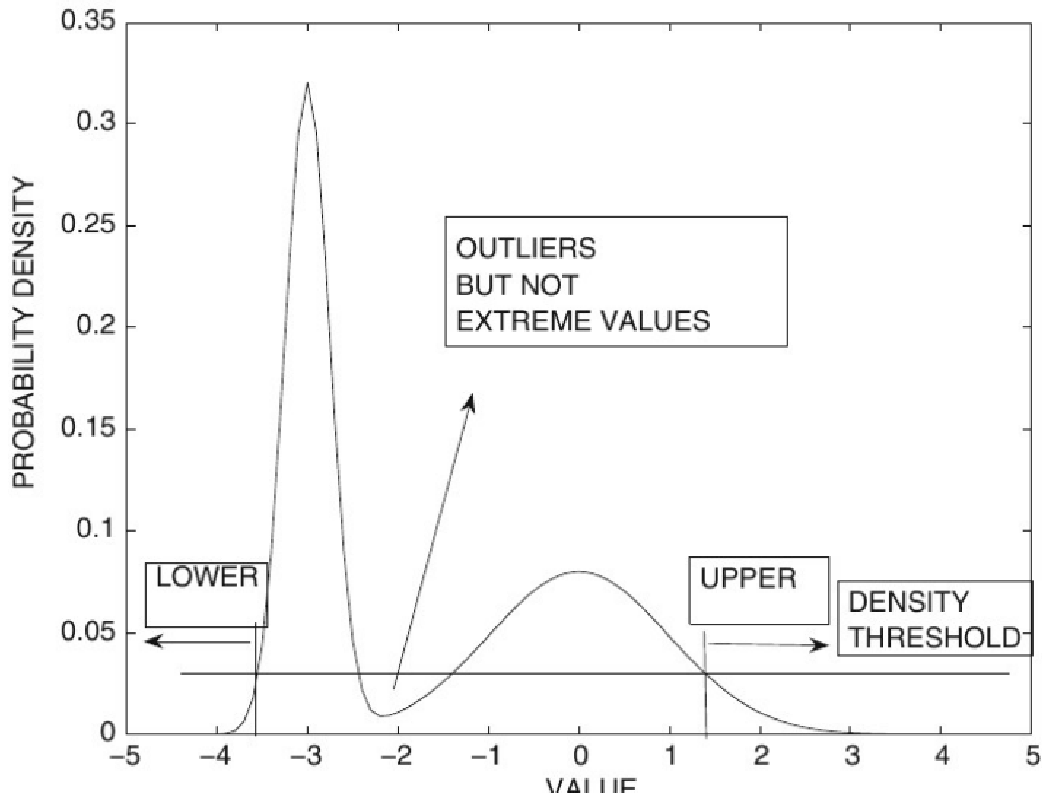
# Extreme value analysis



# Extreme value analysis

- Univariate Extreme Value Analysis: $f_X(x) \leq \theta$



# Extreme value analysis

- Univariate Extreme Value Analysis: The most commonly used model for quantifying the tail probability is the normal distribution

$$f_X(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{\frac{-(x-\mu)^2}{2 \cdot \sigma^2}}$$

- Compute the Z-value for a random variable:

$$z_i = \frac{(x_i - \mu)}{\sigma}$$

  - Large positive values of $zi$ correspond to the upper tail
  - Large negative values correspond to the lower tail

# Extreme value analysis

- Therefore:

$$f_X(z_i) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{\frac{-z_i^2}{2}}$$

- If $zi$ > 3, $xi$ is considered extreme value
- The cumulative area inside the tail can be shown to be less than 0.01% for the normal distribution

# Extreme value analysis

- Multivariate Extreme Values: A multivariate Gaussian model is used

$$f(\overline{X}) = \frac{1}{\sqrt{|\Sigma|} \cdot (2 \cdot \pi)^{(d/2)}} \cdot e^{-\frac{1}{2} \cdot (\overline{X} - \overline{\mu}) \Sigma^{-1} (\overline{X} - \overline{\mu})^T}$$

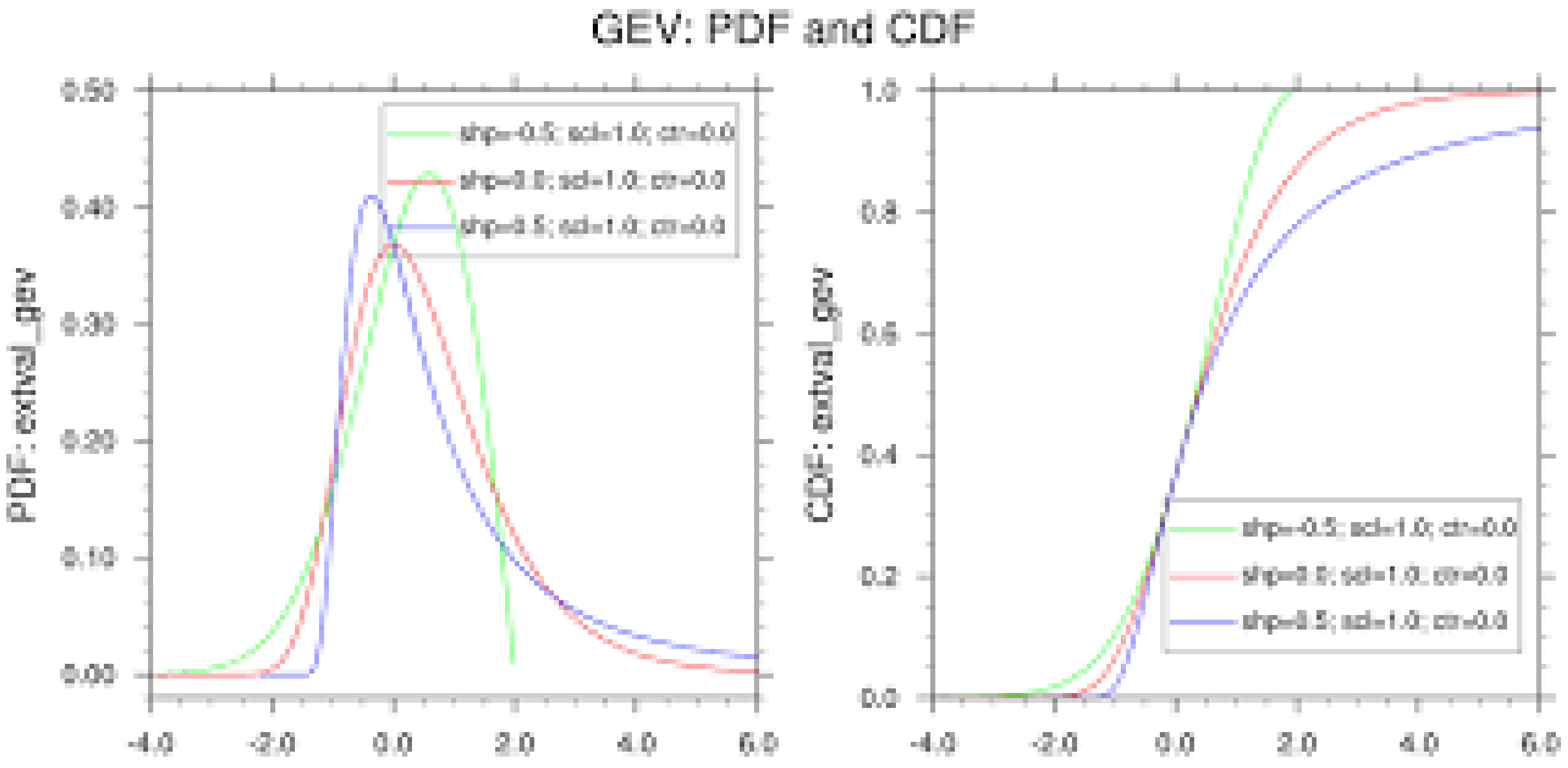
$$= \frac{1}{\sqrt{|\Sigma|} \cdot (2 \cdot \pi)^{(d/2)}} \cdot e^{-\frac{1}{2} \cdot Maha(\overline{X}, \overline{\mu}, \Sigma)^2}$$

- For $f(X)$ less than a particular threshold
  - $Maha(.)$ needs to be larger than a threshold
  - $Maha(.)$ can be used as an extreme-value score

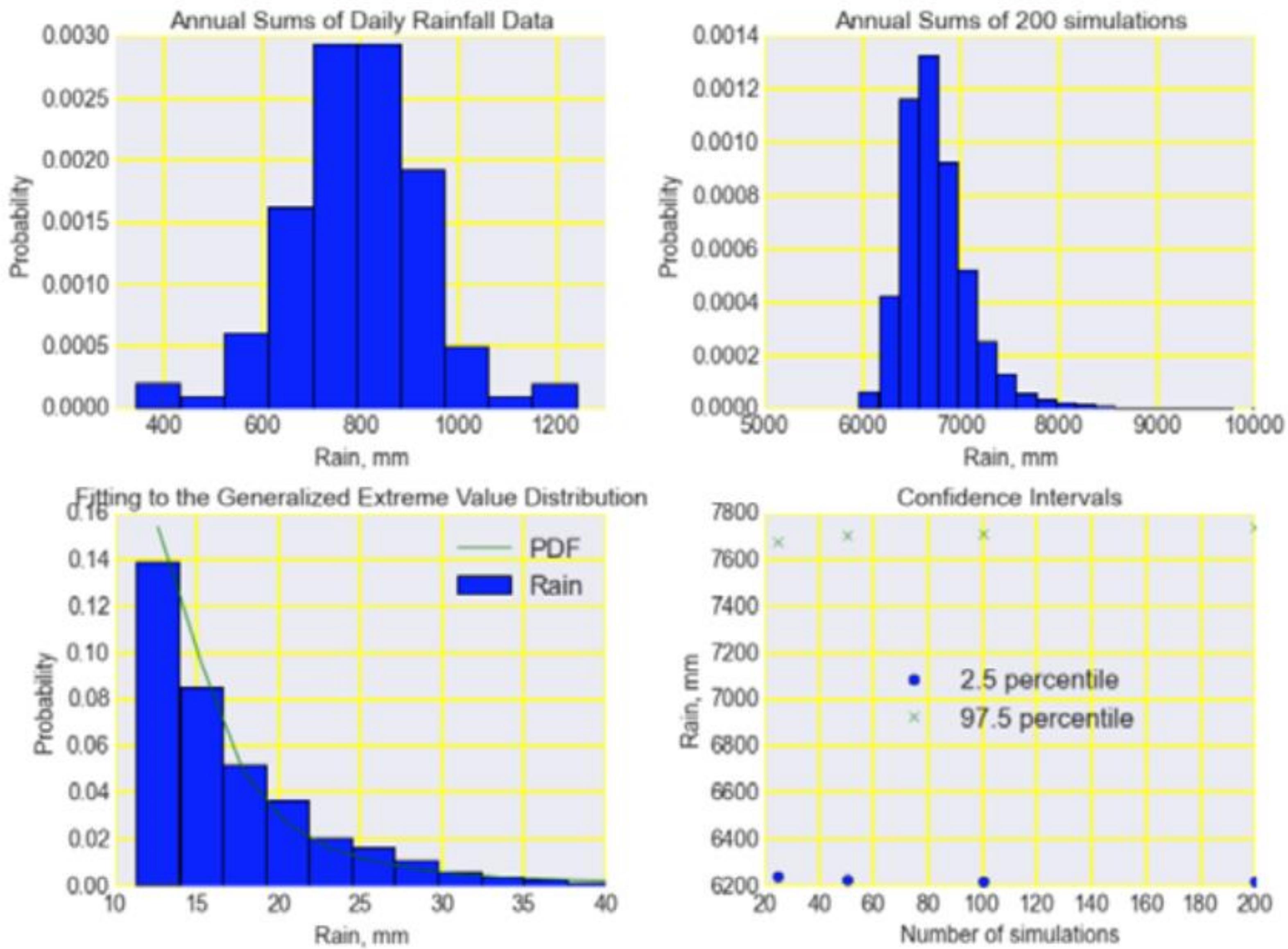
# Extreme value analysis

- The Extreme Value Theorem (aka the Fisher-Tippett-Gnedenko Theorem) states that for a certain class of distributions, the maximum value for a sufficiently large sample will have a **GEV distribution**.
- If a sample comes from a beta distribution (including the uniform distribution) then the maximum value (for a sufficiently large sample) has a reverse Weibull distribution.
- If the sample comes from a Pareto, Fréchet or t-distribution, then the maximum value has a Fréchet distribution.
- Finally, if the sample comes from a Weibull, exponential, gamma, logistic, normal or log-normal distribution then the maximum value has a Gumbel distribution.
- The GEV combines three distributions into a single framework.

# Extreme value analysis



# Extreme value analysis



# Extreme value analysis

- Identify the extreme values of sepal witdth in Iris dataset by different methods (z-score, GEV, and IQR)

# Correlating variables with correlation

| Correlation coefficient | Type of relationship | Levels of measurement | Data distribution |
|---|---|---|---|
| **Pearson's r** | Linear | Two quantitative (interval or ratio) variables | Normal distribution |
| **Spearman's rho** | Non-linear | Two ordinal, interval or ratio variables | Any distribution |
| **Point-biserial** | Linear | One dichotomous (binary) variable and one quantitative (interval or ratio) variable | Normal distribution |
| **Cramér's V (Cramér's φ)** | Non-linear | Two nominal variables | Any distribution |
| **Kendall's tau** | Non-linear | Two ordinal, interval or ratio variables | Any distribution |

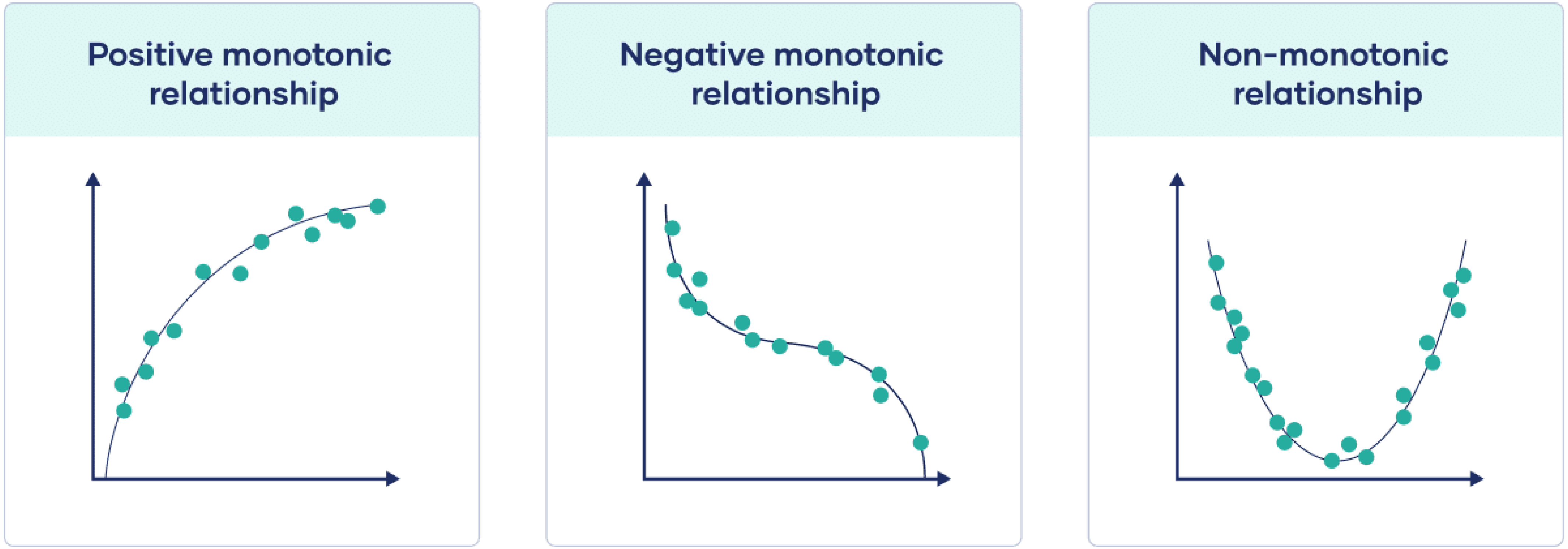
# Correlating variables with correlation

- Pearson's correlation coefficient:
  - These are the assumptions your data must meet if you want to use Pearson’s r:
    - Both variables are on an interval or ratio level of measurement
    - Data from both variables follow normal distributions
    - Your data have no outliers
    - Your data is from a random or representative sample
    - You expect a linear relationship between the two variables

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

# Correlating variables with correlation

- Spearman’s rank correlation coefficient:
  - While the Pearson correlation coefficient measures the linearity of relationships, the Spearman correlation coefficient measures the monotonicity of relationships.

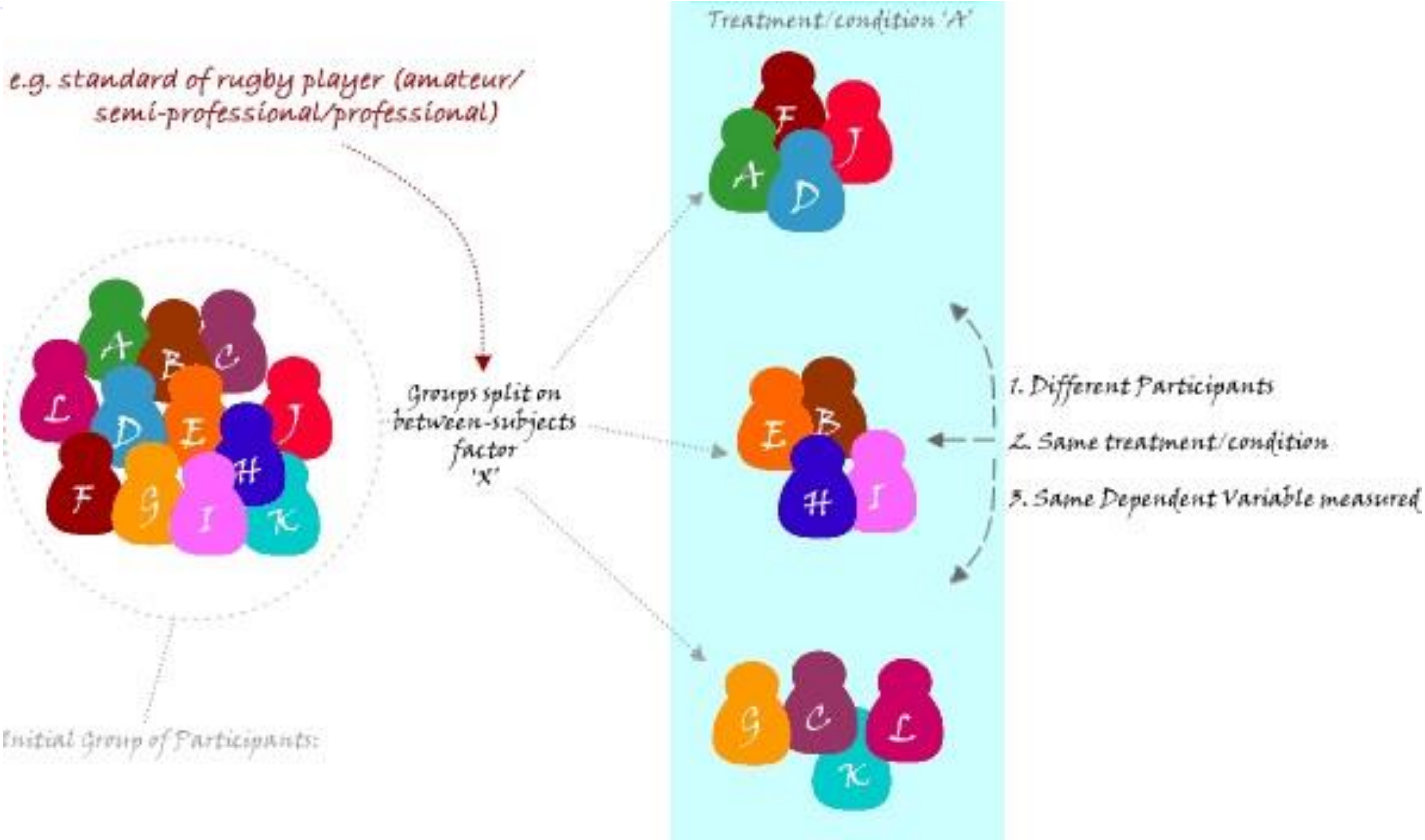


# Correlating variables with correlation

- Spearman's rank correlation coefficient:
  - To use this formula, you'll first rank the data from each variable separately from low to high: every datapoint gets a rank from first, second, or third, etc.
  - Then, you'll find the differences ($d_i$) between the ranks of your variables for each data pair and take that as the main input for the formula.

$$r_s = 1 - \frac{6\sum d_i^2}{(n^3 - n)}$$

# Evaluating relationships between variables with ANOVA



# Evaluating relationships between variables with ANOVA

**ANOVA- definition, one-way, two-way, table, examples, applications**

**ANOVA**

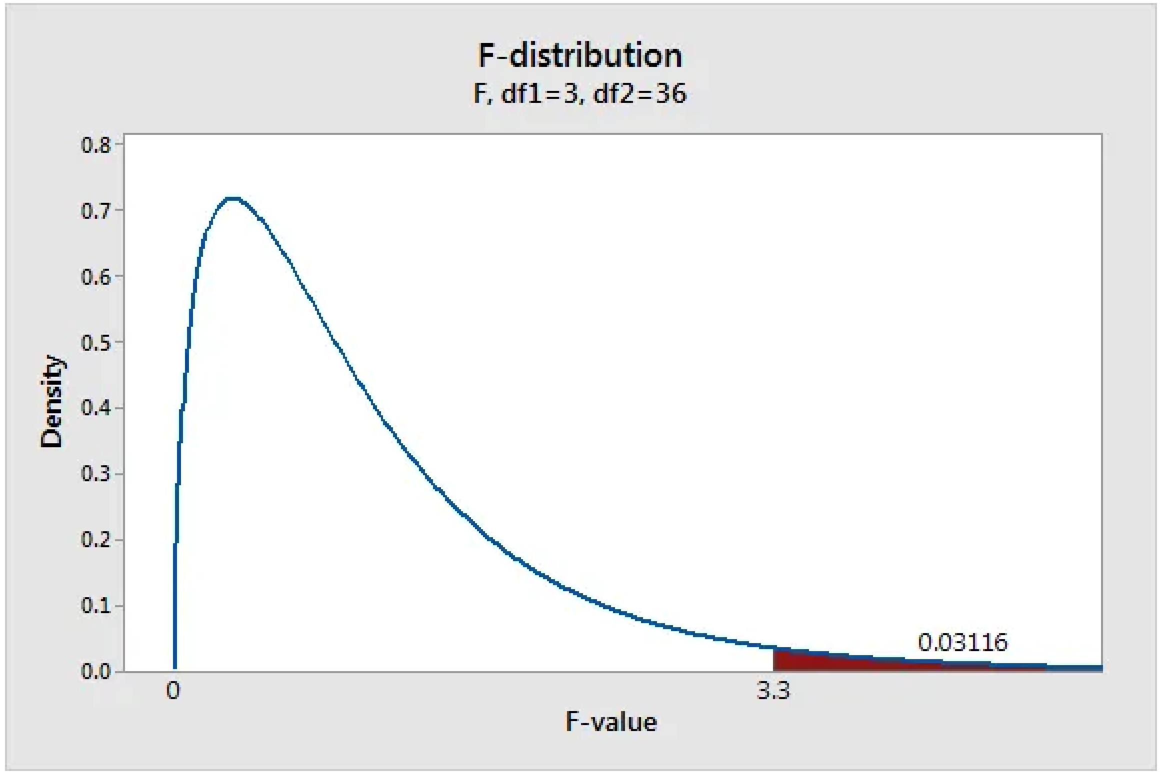
**One-way ANOVA**

**Two-way ANOVA**

| Sources of variation | Sum of squares (SS) | Degrees of freedom (d.f) | Mean sum of square (MS) | F-ratio |
|---|---|---|---|---|
| Between columns | $\sum \frac{(T_j^2)}{N_j} - \frac{(T^2)}{n}$ | (c-1) | $\frac{SS\ between\ columns}{(c-1)}$ | $\frac{MS\ between\ columns}{MS\ residual}$ |
| Between rows | $\sum \frac{(T_i^2)}{N_i} - \frac{(T^2)}{n}$ | (r-1) | $\frac{SS\ between\ rows}{(r-1)}$ | $\frac{MS\ between\ rows}{MS\ residual}$ |
| Residual error | Total SS-(SS between columns and SS between rows) | (c-1)(r-1) | $\frac{SS\ residual}{(c-1)(r-1)}$ | |
| Total | $\sum X_{ij}^2 - \frac{(T^2)}{n}$ | (c.r -1) | | |

# Evaluating relationships between variables with ANOVA

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Factor | 3 | 43.62 | 14.540 | 3.30 | 0.031 |
| Error | 36 | 158.47 | 4.402 | | |
| Total | 39 | 202.09 | | | |



# Evaluating relationships between variables with ANOVA

- Examples
  - https://www.javatpoint.com/anova-test-in-python

# Reference


- https://online.stat.psu.edu/stat415/
- Idris, Ivan. *Python data analysis cookbook*. Packt Publishing Ltd, 2016., Chapter 3.