

# Introduction to Research

Arun K. Tangirala

**Data Analysis**



## Objectives

To learn the following:

- ▶ What is data analysis?
- ▶ Types of analyses
- ▶ Different types of data in analysis
- ▶ Systematic procedure

With two hands-on examples in R<sup>®</sup> (a popular software for data analysis) . . .

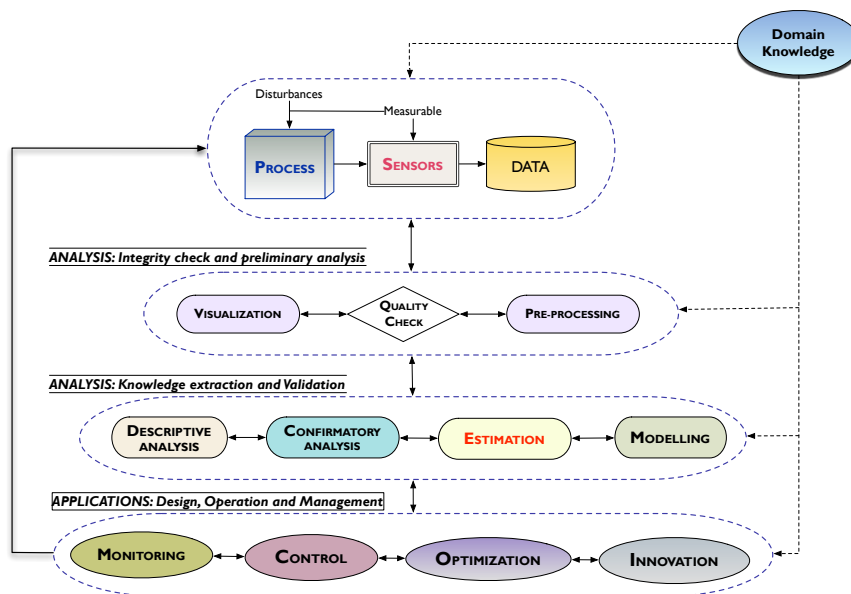
# What is data analysis?

Extracting **useful, relevant and meaningful** information from observations in a **systematic** manner

For the purposes of

- ▶ Parameter estimation (inferring the unknowns)
- ▶ Model development and Prediction (forecasting)
- ▶ Feature extraction (identifying patterns) and classification
- ▶ Hypothesis testing (verification of postulates)
- ▶ Fault detection (process monitoring)
- ▶ ...

# Where does analysis fit in?



## Types of analyses

1. Exploratory vs. Confirmatory
2. Quantitative vs. Qualitative
3. Descriptive vs. Inferential
4. Predictive vs. Prescriptive
5. ...

Exploratory and descriptive analyses are typically good starting points

The objective of analysis determines the type of analysis and the tools that are used.

## Types of data

Several classifications (e.g., numerical vs. categorical, steady-state vs. dynamic, etc.) exist. In data analysis, the primary distinction is w.r.t the **data generating process** (DGP) or mechanism;

**Deterministic** (non-random) and **Stochastic** (non-deterministic) data

1. **Deterministic:** The DGP is perfectly known OR a mathematical function can perfectly explain the data OR there is no uncertainty in the observed values OR process is accurately predictable.
2. **Stochastic:** Observations are only a subset of many (countably finite or infinite) possibilities OR no mathematical function can explain the observed data OR observations have uncertainties in them. Stochasticity does NOT necessarily mean zero predictability

In practice, no process is deterministic since perfect knowledge is never or rarely available.

## Why is this assumption important?

The **theory**, tools for analysis and **interpretations of the results** heavily depend on the assumptions of determinism or otherwise.

- ▶ Example: definitions and computation of periodicities, mean, variance, stationarity

## In practice: deterministic or stochastic?

**Q: When do we assume the DGP to be deterministic or stochastic? Can we know a priori?**

- ▶ Determinism or stochasticity (random) rests with the knowledge of the user (analyst)
- ▶ Whenever the causes are unknown or the underlying mechanism is too complex to be understood, it is reasonable to assume the data to be stochastic (e.g. sensor noise, econometric data).
- ▶ Where the physical process is well-understood and/or the causal variables are known, and the errors in observations are “negligible”, data may be assumed to be (predominantly) deterministic.

**Q: Is it possible to have a mix of both situations?**

- ▶ Yes. Several processes fall into this category (e.g. engineering data).

## Further

- ▶ Be clear on why the data is assumed to be non-deterministic.
- ▶ Suppress the sources of randomness, if possible, during experimentation.
- ▶ **IMPORTANT: Any estimate inferred from data with uncertainties also has uncertainty in it!** Therefore, always attempt to **compute the size of the error in the estimate**.
- ▶ Choose an estimation method that delivers estimates with the lowest possible error and improves with the sample size  $N$ , especially error goes to zero as  $N \rightarrow \infty$ .

## Terminology in data analysis / statistical inferencing

### 1. Population and Sample:

Population refers to the collection of all events or possibilities, typically also known as the ensemble. Sample refers to the subset of population (observations) that has been obtained through sampling.

### 2. Truth and Estimate:

True value refers to the population characteristics in descriptive statistics or to that of true parameters / signals in parameter / signal estimation. Estimate is the inferred value of the unknown parameter / signal from observations.

- ### 3. Statistic and Estimator:
- A statistic is a mathematical function of the given observations. An estimator is also a mathematical function of the observations, but devised for the purpose of inferring or estimating an unknown quantity / variable.

## Terminology

### 4. Ensemble and sample averages:

Ensemble average is the average computed across all set of possibilities (population), at a single point in time (or another domain). Technically this is known as **Expectation**. Sample average is the average that is computed over all observations for one data record.

5. **Bias:** It is the systematic error in an estimate  $\hat{\theta}$  of a parameter  $\theta$  introduced by virtue of the estimation method. Bias is a measure of **accuracy**. Mathematically,  $\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta_0$ , where  $E(\hat{\theta})$  is the ensemble average of estimates across all data records.

6. **Variability:** It is a measure of the spread of estimates obtained from all experimental records. Variability is a measure of the **precision**. Mathematically,  $\text{var}(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2$ , i.e., the ensemble average of the squared errors in the estimate across all records.

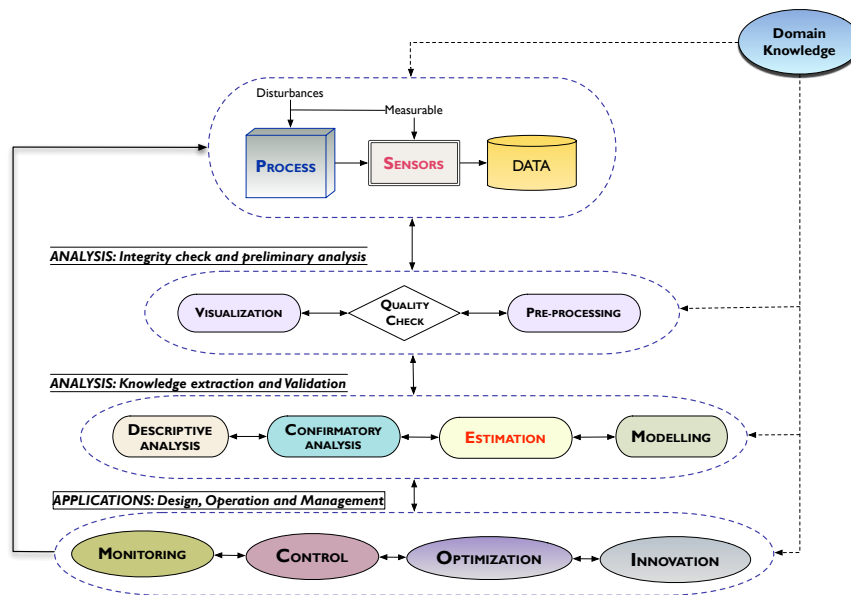
Ideally one would like to obtain estimates with zero bias and as low variability as possible!

## Data analysis procedure

Almost all data analysis exercises are **iterative**, not necessarily sequential.

The success of any data analysis exercise depends on two factors,  
**quality of data** and **methodology of data analysis**.

## A Systematic Procedure



## Preliminary questions in analysis

- ▶ **What type of analysis is required?** Objectives or purpose of data analysis (e.g., exploratory or confirmatory, descriptive or predictive)
- ▶ **Where does the data come from?** Sources of data (domain)
- ▶ **How was the data acquired?** Method of data acquisition / collection (sampling mechanism, hardware used, etc.)
- ▶ **How informative is the data?** Quality of data (e.g., outliers, missing data, anomalies, noise levels)
- ▶ Lastly, but perhaps the most important, **what assumptions are being made on the data generating process?** (e.g., deterministic / stochastic, linear / non-linear, stationary / non-stationary)

## Preparing your data

- ▶ **Data pre-processing**
  - ▶ Cleaning up the outliers or faulty data, re-scaling, filling in missing data
  - ▶ Robust methods can handle outliers or faulty data without an explicit treatment
  - ▶ Data reconciliation - adjusting data to satisfy fundamental conservation laws
  - ▶ Synchronization of time-stamps in multivariate data
- ▶ **Partitioning the data**
  - ▶ Partition into “training” and “test” data sets.
  - ▶ In certain applications, several data sets may have to be merged.
- ▶ **Assessing quantization or compression errors** for data from historians.

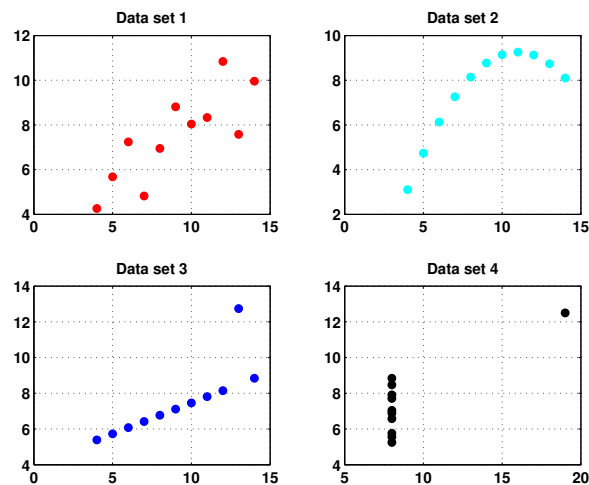
## Visualization

Always convert numbers to pictorial representations before, during and after mathematical operations - visual analysis reveals a wealth of information

- ▶ Simple trend to high-dimensional plots
- ▶ Bar charts, Box plots, histograms
- ▶ Color maps, contour plots
- ▶ Specialized plots
- ▶ ...



## Anscombe data sets



- ▶ Four data sets having identical mean, variance, line fit, correlation coefficient, but completely different features!

## Core steps





- ▶ **Select the domain of analysis** (e.g, time, frequency).
  - ▶ Transformation of data does wonders (e.g., Fourier transform for detecting periodicities)
- ▶ **Dimensionality reduction** for large-dimensional data
  - ▶ Multivariate data analysis tools (e.g., PCA, NMF)
- ▶ **Choose the right mathematical / statistical tool** appropriate for the purpose of analysis and commensurate with the assumptions made on the data.
  - ▶ In hypothesis tests, the appropriate statistic and significance level should be chosen
  - ▶ In analysis of variance (ANOVA) problems, sources of variability should be clearly identified.
  - ▶ In modelling, appropriate regressor set and model structure must be chosen.
- ▶ **Always factor in prior and/or domain knowledge** at each step.

## Assessing and reporting your results

- ▶ **Compute confidence regions and errors** in estimates.
- ▶ **Conduct hypothesis tests** on estimates.
- ▶ **Cross-validate models** on fresh data sets (did the model over learn?).
- ▶ If **multiple models** are identified, use a mix of information-theoretic measures (e.g., Akaike Information Criterion) and end-use criteria to **select an appropriate model structure**. Remember we seek the **best working model**, not necessarily the right model!
- ▶ Where forecasts are computed, confidence bounds should be provided.

Any report from a data analysis exercise should include estimates, their error bounds and results from statistical tests of validation.

## Bibliography I

-  Bendat, J. S. and A. G. Piersol (2010). *Random Data: Analysis and Measurement Procedures*. 4<sup>th</sup> edition. New York, USA: John Wiley & Sons, Inc.
-  Johnson, R. A. (2011). *Miller and Freund's: Probability and Statistics for Engineers*. Upper Saddle River, NJ, USA: Prentice Hall.
-  Montgomery, D. C. and G. C. Runger (2011). *Applied Statistics and Probability for Engineers*. 5<sup>th</sup> edition. New York, USA: John Wiley & Sons, Inc.
-  Ogunnaike, B. A. (2010). *Random Phenomena: Fundamentals of Probability and Statistics for Engineers*. Boca Raton, FL, USA: CRC Press, Taylor & Francis Group.