# Overview Over The Cycle Multivariate Statistics

Stephan Huckemann

May 4, 2019

Statistics is essentially *scientific divination*, e.g. fortune telling, as it learns about a model and adds probabilistic bounds in order to come up with a prediction of the future.

**Example 1.**[The Central Limit Theorem (CLT)] *Let $X_1, ..., X_n \in \mathbb{R}$ be independent identically distributed random variables (with finite second moments), then the famous CLT states that*

$$\sqrt{n}\,\frac{\bar{X} - \mu}{\sqrt{\mathrm{var}[X_1]}} \xrightarrow{D} N(0,1)$$

*with the true mean $\mu$.*

Hence upon a sample, one can predict the true $\mu$ with a given confidence (almost like a probability). Typical questions, turning multivariate:

- clarify the notion "confidence" vs. probability;
- what if $\mathrm{var}[X_1]$ unknown?
- can we do better if we have some parametric prior knowledge (belief)? e.g. normal errors, exponential (waiting times), Poisson (photon noise in nano-microscopy)?
- non-parametric inference?
- Bayesian "magic" may considerably improve estimation and inference;
- what if data are vectors (classical multivariate setting), can we increase power of tests if we consider all components simultaneously? The more data simultaneously considered the better!
- what if data are more general objects?
  - functions (growth curves, climate time series);
  - (projective) directions (geology, wind, astronomy, say);
  - fingerprints where rotation and translation do not matter $\rightarrow$ quotient space;
  - shapes of objects (2D images of faces, where additionally size does not matter);
  - points on a manifold stratified space (boundaries can be lower dimensional manifolds,where many "pieces" meet);
- statistical dimension reduction (what does that mean on a manifold, say?);
- CLTs may change with topology and geometry.

**Typical topics covered:**

1. *Introduction*, usually accompanied by a *forensics* (mainly *fingerprint analysis*) seminar:
   - parametric (exponential) families;
   - sufficient and complete statistics;
   - optimal estimation of parameters including the EM algorithm;
   - Bayesian estimation and Markov Chain Monte Carlo (MCMC).
2. *Advances*, usually accompanied by a *biometrics* seminar:
   - hypothesis testing;
   - confidence regions (optimal and reasonable estimation of parameters);
   - credibility regions ( = the Bayes version of the above).
3. *Specialization*, usually accompanied by a *statistical learning* seminar:
   - multivariate statistics (MVA) and (generalized) linear models, ANOVA, MANOVA, factor analysis, Kalman filters, etc.;
   - dimension reduction by (kernel) principal component analysis (PCA) and multidimensional scaling (MDS);
   - more statistical (classifier) learning techniques, resampling and regularisation, e.g. support vector machines, boosting, bootstrap, LASSO.
4. *Aspects*, usually accompanied by a *shape spaces* seminar:
   - parametric and nonparametric (e.g. Procrustes) statistics on Non-Euclidean spaces (e.g. circular, spherical, shape and phylogenetic)
   - generalized Fréchet means and
   - their limit theorems on stratified spaces leading to *stickiness* and *smeariness*;
   - Non-Euclidean dimension reduction;
   - functional data analysis (FDA), on Lie groups, say;
   - non-parametric statistics (M-estimators, some empirical process theory);
   - stochastic processes on manifolds;
   - (Gaussian) kinematic formula.

**Requirements, necessary and helpful:**

- Intro, Advances and Specialization: necessary are
  - measure and integration theory ($\sigma$-fields, measures, null sets, Fubini). In particular, Bayes people fancy *conditional expectation* of *random variables* where the latter are measurable mappings between measure spaces, – and so are the former;
  - basic knowledge of statistics: mean, variance, median, moments, some distributions (normal, binomial, Bernoulli), strong law and the CLT (central limit theorem).
- Aspects: it helps
  - (tremendously) to know basic differential geometry (tangent space, bundles, tensors, curvature, geodesics, Lie-groups);
  - and to know basic stochastic processes process theory.

**Master and bachelor theses topics**, preferably *research oriented* (upon sufficient commitment papers emerge):

- on theory of
  - Non-Euclidean limit theorems;
  - stochastic processes on manifolds;
  - data spaces and dimension reduction thereon;
- and on applications to
  - forensics/fingerprint analysis;
  - differentiation of stem cells;
  - gait analysis of the knee joint;
  - protein 3D structure;
  - radiation therapy;
  - phylogenetic tree analysis.

# References


**Introduction:**

Bickel, P. and K. Doksum (2001). *Mathematical statistics: basic ideas and selected topics, Vol. 1* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Lehmann, E. L. and G. Casella (2003). *Theory of Point Estimation*. Springer.

Robert, C. and G. Casella (2004). *Monte Carlo statistical methods*. Springer Verlag.

Shao, J. (2003). *Mathematical Statistics*. Springer, New York.

**Advances:**

Robert, C. P. (2007). The Bayesian choice: From decision-theoretic foundations to computational implementation (springer texts in statistics) by.

Romano, J. P. and E. L. Lehmann (2005). *Testing statistical hypotheses*. Springer, Berlin.

**Specialization:**

Friedman, J., T. Hastie, and R. Tibshirani (2017). *The elements of statistical learning*, Volume 1. Springer series in statistics New York, NY, USA:. 2nd edition, corrected 12th printing.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1980). *Multivariate Analysis*. Academic press.

Sengupta, D. and S. Jammalamadaka (2003). *Linear models: an integrated approach*. World Scientific Pub Co Inc.

**Aspects:**

Dryden, I. L. and K. V. Mardia (1998). *Statistical Shape Analysis*. Chichester: Wiley.

Hsu, E. P. (2002). *Stochastic analysis on manifolds*, Volume 38. American Mathematical Soc.

Mardia, K. V. and P. E. Jupp (2000). *Directional Statistics*. New York: Wiley.

van der Vaart, A. (2000). *Asymptotic statistics*. Cambridge Univ. Press.