

AbstractID: 9815 Title: Models in Medicine. I. Introduction to Statistical Inference and the General Linear Model.

The Institute of Medicine notes that, "... a [mathematical] *model is the tool for converting data into insight.*" This remark suggests that it could be useful to some medical physicists to be able to better understand and use methods of formulating *deterministic* and *stochastic* models, and of fitting and testing them against real data. This series of four lectures offers an introduction to some of these models and methods. The first lecture provides a general introduction to the construction, evaluation and deployment of deterministic and stochastic models, empirical and mechanistic, in observational and experimental studies. We provide first a brief survey of *dynamical systems* – sets of ordinary/stochastic differential equations. Solutions are typically obtained by Runge-Kutta methods. We then turn to statistical inference, beginning with a taxonomy of data based on the *measurement scale* (i.e., nominal, ordinal, interval, ratio) and *distribution type* (e.g., discrete, continuous) that determine the model (e.g., Student's t or Wilcoxon) most appropriate to the research. Statistical inference is concerned with drawing conclusions, based on *observed* quantities, i.e., data, about quantities that are *not observed*, e.g., parameters. The validity, range, and usefulness of statistical inferences depend decisively upon how the data are obtained. Thus, for example, *random selection* and *random allocation* of subjects are required for *population* and *causal* inferences, respectively. There are two principal schools of statistical inference: Neyman-Pearson and Bayes. Neyman-Pearson is covered in the first lecture and Bayes in the second. In Neyman-Pearson inference statistical modeling proceeds in three steps: 1) specification of the model; 2) estimation of the parameters; 3) making inferences, that is, assessing the goodness-of-fit, testing hypotheses and estimating confidence intervals. We examine Neyman-Pearson inference in the context of the General Linear Model (GLM), $y = X\beta + \varepsilon$, where y is an $(n \times 1)$ *random* vector response variable (*interval* scale; *continuous* distribution), X is an $(n \times p)$ matrix of *fixed* explanatory variables, β is a $(p \times 1)$ vector of *fixed* but unknown parameters, ε is an $(n \times 1)$ vector of random "errors". The GLM includes Student's t-test, analysis of variance, analysis of covariance and multiple linear regression models; we examine each. The research questions addressed by the first three involve comparison of means; emphasis is on *hypothesis testing*. In regression, emphasis is on *estimation*. Typically, X-variables are nominal in the t-test and analysis of variance and *interval/ratio* in regression; analysis of covariance includes both of these types of X-variables. In the GLM the stochastic component, ε , is assumed to be a statistically independent, identically and Normally distributed set with mean 0 and unknown constant variance σ^2 . Least squares methods are used to obtain estimates of goodness-of-fit and β and σ^2 in all three models. Model *diagnostics* and *sensitivity tests* are examined as well as robust and imputation methods for dealing with outlying, influential and missing observations. Ridge regression methods for constructing regression models of multicollinear data are described. Model validation methods are discussed. Medical applications of the GLM are described. **Educational Objectives:** Understanding of 1) The Neyman-Pearson model of statistical inference. 2) The General Linear Model.