Basics of Inference

Sofia Olhede[1]

[1]Department of Statistical Science, University College London, UK.

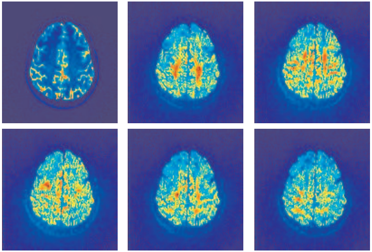
UCL, Nov 24th.

Material to be Covered:

- Statistical problems and models.
- Uncertainty.
- Parametric Methods of producing estimators.
- Bayesian methods.

Why do we collect data?

- We wish to make decisions *given* we have observed data generated from some scenario.
- If our decision is not to be influenced by the data, it makes **no sense** to collect the data.
- Not all data yields information about all models. Statistical problems are *inductive* rather than *deductive*.



What is a Model?

- A model is how we describe the generation of an observable quantity.
- A model can be mechanistic, describing the underlying mechanism that explains the observed data (think Newton, laws of motion).
- A model can be empirical, explaining observed variability (think Kepler).

Uncertainty

- In most data collection scenarios there is *uncertainty*.
- Uncertainty arises because there is stochastic uncertainty.
- Uncertainty also arises due to inductive uncertainty.
- Whatever modelling decisions you make – you probably could have made another... “All models are wrong, but some are useful...”

Uncertainty

- The condition of being uncertain; doubt.
- Something uncertain: the uncertainties of modern life.
- The estimated amount or percentage by which an observed or calculated value may differ from the true value.

Parametrics etc

- Most models are parametric, e.g. depend on a set number of parameters in which terms the model is specified.
- Sometimes models are semi-parametric, e.g. certain aspects of the model are parametric.
- Or models can be non-parametric, e.g. not depend on parameters.
- Modern problems often contain more variables p than data points n, as we shall return to.
- If you try to estimate more parameters than you have data points, a number of fallacies will most often arise. Typical example is using SVD when you didn't have many replicates.

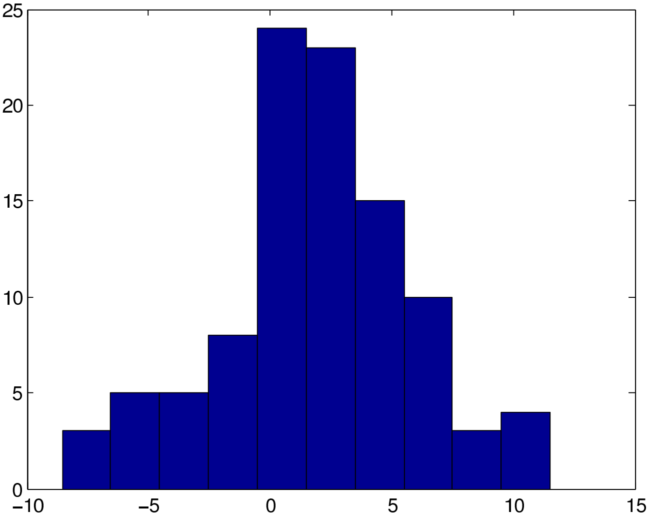
Basics

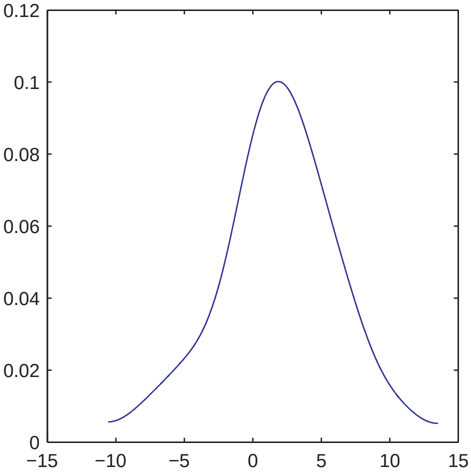
- A parametric model usually corresponds to specifying a cdf on a scalar:
$$F_X(x) = P\left(X \leq x|\boldsymbol{\theta}\right), \tag{1}$$
or a pdf corresponding to a derivative of $\frac{d}{dx}F_X(x) = f_X(x)$, or a vector $\mathbf{X} = (X_1, \ldots, X_n)^T$
$$F_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) \equiv P\left(X_1 \leq x_1, \ldots, X_n \leq x_n|\boldsymbol{\theta}\right).$$
- The set of random variables $X_1, \ldots, X_n$ is said to be a **random sample** of size n from a population with pdf $f_X(x|\boldsymbol{\theta})$ if the joint pdf has the form
$$f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) = f_X(x_1|\boldsymbol{\theta}) \cdots f_X(x_n|\boldsymbol{\theta}).$$
This means that the joint density function is a product of marginal densities.

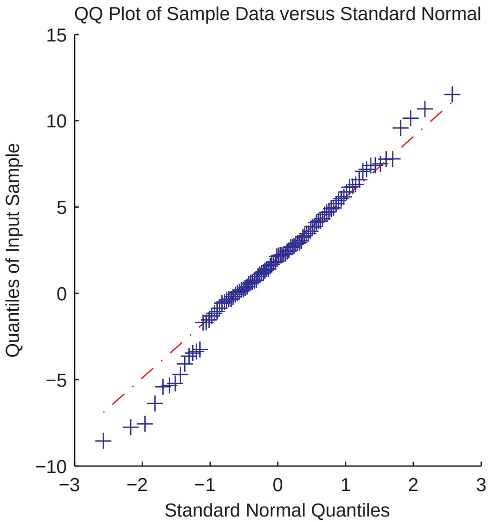
Nonparametrics

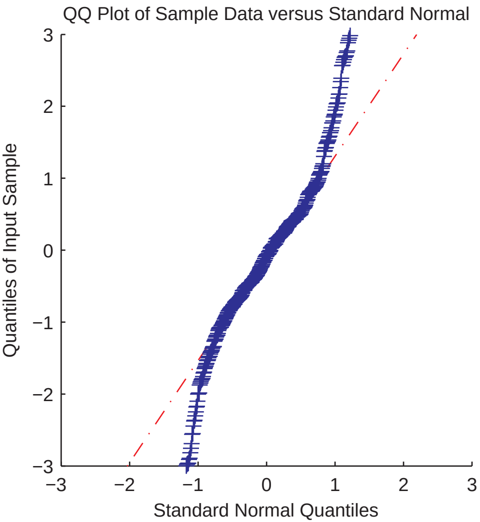
- Before formulating a parametric model it may be sensible to start by describing the data non-parametrically.
- For example we can estimate the pdf using a histogram or a kernel density estimator.
- Histograms are very crude, and their characteristics depend on their bin size.
- To investigate whether a given distribution is appropriate, our first choice is a q-q plot, e.g. using the ordered data $X_{(1)}, \dots, X_{(n)}$ and plotting

$$\left\{ \left(F_X^{-1}\left(\frac{j}{n+1}\right), X_{(j)} \right), \right\} \tag{2}$$


25
20
15
10
5
0
−10
−5
0
5
10
15


0.12
0.1
0.08
0.06
0.04
0.02
0
−15
−10
−5
0
5
10
15


QQ Plot of Sample Data versus Standard Normal
Quantiles of Input Sample
Standard Normal Quantiles
15
10
5
0
−5
−10
−3
−2
−1
0
1
2
3


QQ Plot of Sample Data versus Standard Normal
Quantiles of Input Sample
Standard Normal Quantiles
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

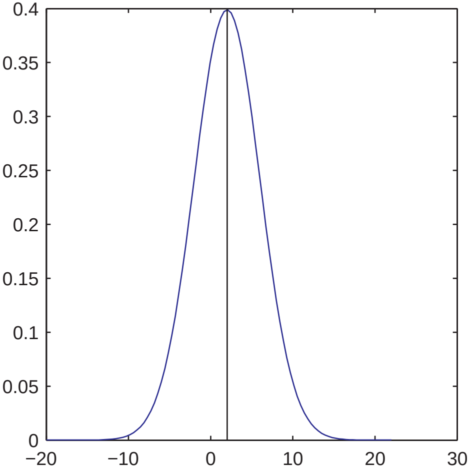
Expectation

- The expectation of a random variable is

$$\mathrm{E}(X) = \int x f_X(x)\, dx = \mu.$$

- The variance of a random variable is

$$\mathrm{var}(X) = \int (x-\mu)^2 f_X(x)\, dx.$$


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−20
−10
0
10
20
30

Estimation

- Usually one wishes to learn about $\boldsymbol{\theta}$ from the data, by calculating statistics.
- A statistic is a function of the data which does not depend on any unknown parameters.
- A statistic that is used to estimate the value of the parameter θ is called an **estimator** of θ, and an observed value of the statistic is called an **estimate** of θ.

Method of moments

- jth moment of X is given by

$$E_X(X^j|\theta) = \int x^j f_X(x|\boldsymbol{\theta}).$$

- jth **sample moment** of random sample $X_1, \ldots, X_n$ is given by

$$M_j = E_{X_j}(X^j) = \frac{1}{n}\sum_{i=1}^{n} X_i^j.$$

- The method of moments equates theoretical and empirical moments to determine the unknown parameters.
- We solve the k equations

$$E_{\boldsymbol{X}|f_X}(X^j|f_X) = M_j, \quad j = 1, \ldots, k.$$

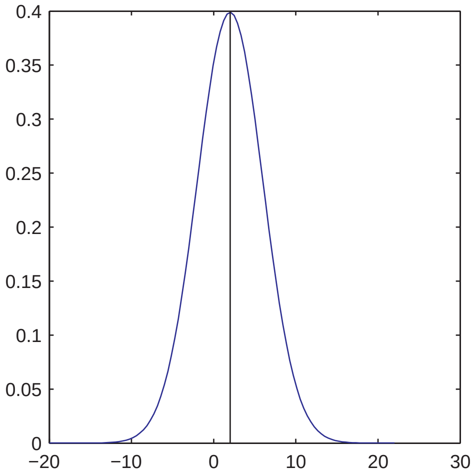
- How do we know if this is a good estimator?

How Good Is An Estimator

- In theory we could make up any amount of estimators.
- These are generally evaluated in terms of their mean square error, that is the aggregation of the bias square plus the variance, where the bias of estimator T for θ is

$$\mathrm{E}(T) - \theta.$$

- A good estimator has a small mean square error.


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−20
−10
0
10
20
30

Dimensionality Reduction

- We have focused on understanding **one** vector of observations in terms of their distribution, or a set of explanatory variables.
- Sometimes we wish to understand many variables simultaneously.
- Assume we have $Y_i^{(j)}$, N observations of each of p variables, or p vectors $\mathbf{Y}^{(j)}$ of length N.
- We can then define

$$\boldsymbol{\Sigma} = \begin{pmatrix} \operatorname{cov}\{Y_i^{(1)}, Y_i^{(1)}\} & \operatorname{cov}\{Y_i^{(1)}, Y_i^{(2)}\} & \dots & \operatorname{cov}\{Y_i^{(1)}, Y_i^{(p)}\} \\ \dots & \dots & \dots & \\ \operatorname{cov}\{Y_i^{(p)}, Y_i^{(1)}\} & \operatorname{cov}\{Y_i^{(p)}, Y_i^{(2)}\} & \dots & \operatorname{cov}\{Y_i^{(p)}, Y_i^{(p)}\} \end{pmatrix}$$

Estimation

- We can estimate the covariance using

$$\widehat{\sigma}_{kj} = \frac{1}{N} \sum (Y_i^{(k)} - \overline{Y}^{(k)})(Y_i^{(j)} - \overline{Y}^{(j)}) \tag{4}$$

- To convey most of the structure of the data wish to replace $\mathbf{\Sigma}$ by an approximation.
- Because $\mathbf{\Sigma}$ is symmetric it has an eigen-decomposition, e.g. it can be written as

$$\mathbf{\Sigma} = \sum_{j=1}^{p} \lambda_j \mathbf{v}_j \mathbf{v}_j^T. \tag{5}$$

- If only a few λ_j are large, then $\mathbf{\Sigma} \approx \sum_{j=1}^{p_0} \lambda_j \mathbf{v}_j \mathbf{v}_j^T$.

Likelihood Inference

- The joint density function of n random variables $X_1, \ldots, X_n$ evaluated at $x_1, \ldots, x_n$, say $f(x_1, \ldots, x_n; \theta)$ is referred to as a **likelihood function**.
- For a fixed data sample $x_1, \ldots, x_n$ the likelihood is only a function of θ and we shall denote it by $\ell(\theta)$.
- For a random sample $X_1, \ldots, X_n$ from $f(x; \theta)$,

$$\ell(\theta) = f(x_1; \theta) \cdots f(x_n; \theta) = \prod_1^n f(x_i; \theta).$$

- Usually convenient to deal with the log-likelihood

$$L(\theta) = \log(\ell(\theta)).$$

Linear and Generalized Linear Models

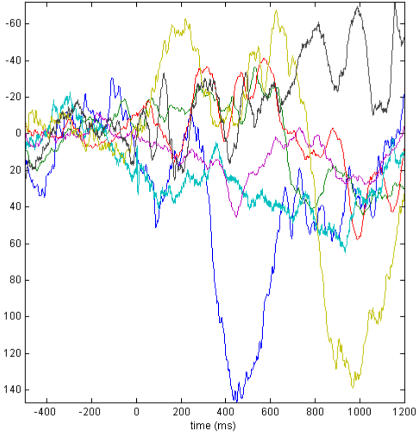
- A typical model is
$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$$
with a distribution on ϵ_i. This is a linear model.
- If Y_i is constrained to be positive or lie in a range, it may be more convenient to use a generalized linear model, or to say
$$\begin{aligned} \mathrm{E}Y_i &= \mu_i \qquad (6)\\ \mu_i &= g^{-1}\left(\mathbf{x}_i^T \boldsymbol{\beta}\right) \qquad (7)\end{aligned}$$
- You will hear more about the linear model and regression in later lectures.

Example – GMLs

- At UCLH prematurely born infants are subjected to noxious stimulation as part of their scheduled treatment.
- Many noxious stimulations are inevitable in the due course of their stay.
- We wish to understand how they respond to stimuli, painful or otherwise.
- The data comes in the form of time-courses $Y_i^{(j)}$ measured at time t_i. We think

$$Y_i^{(j)} = \sum a_{jk} z_k(t_i) + \varepsilon_{ij}. \tag{8}$$

Typical Signals



Delta-brushes

- One of the $z_k(t_i)$ is a delta-brush, e.g. a non-specific neuronal burst.
- It is a significant change from the baseline energy occurring simultaneously in the low frequency 86 band (0.5-1.5 Hz) and the high frequency band (8-25 Hz).
- We detect this from a regression coefficient, which is "present" if exceeds a threshold, based on the statistics of the noise, correcting for multiple testing.
- We now wish to explain the detection in terms of the age of the infant τ_j. How???

GLMs

- We assume that the variable Z_j takes the value of zero or one depending on if a delta-brush was detected.
- We **cannot** assume

$$\mathrm{E}Z_j = \theta_j = \beta_0 + \beta_1 \tau_j. \tag{9}$$

- Instead we take

$$\theta_j = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \tau_j)}}$$

- This can be fitted to the data using the fact that Bernoulli random variables are part of the **exponential** family. Parameter fitting is part of a larger class of algorithms.

Range of Behaviour

Development of Touch and Pain Discrimination
1555

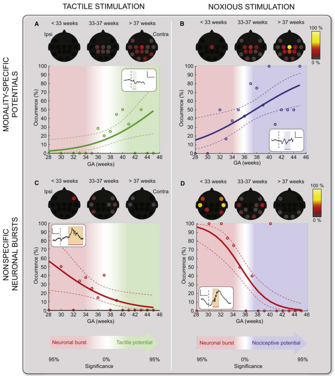


Figure 3. Relationship between Response Type, Nonspecific Neuronal Burst, or Modality-Specific Potentials, Evoked by Tactile and Noxious Stimulation, with Gestational Age

Age dependence of the occurrence and topographical distribution of tactile (A), nociceptive-specific potentials (B), and nonspecific neuronal bursts (C and

Likelihood Inference

- Learning from the data. – "rational degree of belief".
- R.A. Fisher, On the mathematical foundations of theoretical statistics, Philosophical Transactions of the Royal Society, A, 222: 309–368. (1922).
- *'... inference from an experiment should be based only on the likelihood function for the observed data.'*
- For a given observed set of data $\ell(\theta)$ gives the likelihood of that set occurring as a function of θ. The ML principle of estimation is to **choose as the estimate of θ that value for which the observed set of data would have been most likely to occur**. That is

$$\hat{\theta} = \arg \max_{\theta \in \Omega} \ell(\theta).$$

Likelihood Inference

- **If** the likelihood is differentiable and achieves a maximum in Ω, then the MLE will be the solution to the maximum likelihood equation

$$\frac{d}{d\theta}\ell(\theta) = 0, \text{ with } \frac{d^2}{d\theta^2}\ell(\theta) < 0.$$

- If $\hat{\theta}$ is the mle of θ and if $t(\theta)$ is a monotone function of θ then $u(\hat{\theta})$ is an mle of $u(\theta)$.
- The ml estimator with a sample size n, $\hat{\theta}_n$,
  1. exists and is unique,
  2. is a consistent estimator of θ (increasing n),
  3. is nearly normal with approximate mean θ and variance

$$\left[nE\left\{\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]^2\right\}\right]^{-1}.$$

Vector θ

- If θ is a p-parameter vector, then most often with increasing n, $\hat{\theta}_n - \theta$ is nearly zero-mean multivariate Gaussian.
- It has a covariance matrix which can be found from the Hessian matrix of the log-likelihood.
- Properties follow from

$$\nabla L(\theta)|_{\theta=\theta_o} = \nabla L(\theta)_{\hat{\theta}} + \mathbf{H}\left(\theta_o - \hat{\theta}\right)$$

- Most computer packages can maximize the likelihood for you.
- What happens if p is large??? (LARS/LASSO/penalization).

Nuisance parameters

- Unfortunately often not all of $\boldsymbol{\theta}$ are of interest.
- If we only want to estimate ψ where $\boldsymbol{\theta} = [\psi, \phi]$ then we can either use:
  iterated maximization (profile likelihood),
  marginal likelihood methods.
- Profile likelihood is defined by

$$\hat{\phi}(\psi) = \arg_{\psi\, fixed} \max L(\boldsymbol{\psi}, \boldsymbol{\phi}), L_*(\psi) = L(\boldsymbol{\psi}, \hat{\phi}(\psi)). \quad (10)$$

- There is some loss of performance, but there is a well-developed theory.

Hypothesis testing

- Hypothesis testing can be implemented by comparing the likelihood optimized under different hypotheses of the parameters.
- The ratio of the likelihood follows a known distribution if hypotheses are *nested*.
- Other methods include the score test.

Bayesian Inference

- We have assumed that there is some parameter θ with some unknown constant value.
- We could think of the unknown parameter θ as being a realisation from random variable Θ where Θ has some supposed distribution $p(\Theta = \theta)$.
- The previous approach is a special case of this method with $p(\Theta = \theta_0) = 1$ and $p(\Theta \neq \theta) = 0$.

Bayesian Inference II

- We write

$$p(\mathcal{D},\theta) = p(\mathcal{D}|\theta)p(\theta) = p(\theta|\mathcal{D})p(\mathcal{D}),$$

  where $\mathcal{D} = (X_1,\ldots,X_n)$ and $p(\cdot)$ is either a pmf or pdf, giving us

$$\begin{aligned} p(\theta|D) &= \frac{p(D|\theta)p(\theta)}{p(D)} \propto p(D|\theta)p(\theta) \\ \text{Posterior} &= \text{Likelihood} \times \text{Prior}. \end{aligned}$$

- By Bayes' Theorem and note that $p(D)$ is **not** a function of θ and it is given by

$$p(D) = \int_\Theta p(D|\theta)p(\theta)d\theta.$$

Bayesian Inference III

- We write the likelihood with a conditional sign rather than a semi-colon to reflect the fact that θ is a random variable rather than a constant.
- Using Bayes theorem allows us to determine a posterior distribution for Θ which gives us all the available information about it after we have seen the data, D.
- Usually it is impossible to do all of the integrals analytically, unless the distributions are chosen to be conjugate.
- Winbugs is a practical programme for implementing Bayesian analysis.

Bayesian Inference III

- We may report a single value for each parameter, such as a maximum aposteriori estimate.
- We may want an interval which will contain Θ with probability that include the most concentrated areas of $p(\theta|D)$.
- We can do this by determining the $100\gamma\%$ credible interval which is an interval which contains $100\gamma\%$ of the total density in the posterior distribution.
- Let $l(\mathbf{x})$ and $u(\mathbf{x})$ be some functions of the observed data then a $100\gamma\%$ credible interval satisfies

$$\begin{aligned} P(l(\mathbf{x}) < \Theta < u(\mathbf{x})|D) &= \int_{l(x)}^{u(x)} p(\theta|D)d\theta \\ &= \gamma. \end{aligned}$$

Approximate Bayesian Computation

- Often the posterior distribution is not readily available. Computational methods such as Metropolis Hastings (Robert and Casella) and Gibbs sampling (Casella and George), can give samples from the posterior.
- When we have large sets of data, this may be because the likelihood cannot be computed quickly.
- We follow Pritchard et al (1999). We can however **simulate** from $f(y|\boldsymbol{\theta})$.
- We sample a vector $\boldsymbol{\theta}^*$ from some proposal density $\pi(\boldsymbol{\theta})$.
- We simulate $f(y|\boldsymbol{\theta}^*)$.
- If $d(\boldsymbol{Y}, \boldsymbol{Y}_0) < \varepsilon$, for some tolerance level – accept $\boldsymbol{\theta}^*$ as a sample from the posterior.
- Choosing $d()$ is a **very** strong statement. Can replace $\boldsymbol{Y}$ by some well chosen summary statistics.

References I

- George Casella and Edward I. George.
  Explaining the Gibbs sampler
  The American Statistician, 46:167–174, 1992.
- Cox, D. R.,
  Principles of Statistical Inference,
  CUP, Cambridge, 2008.
- A. C. Davison,
  Statistical Models,
  CUP, Cambridge, 2003.

References II

- Hastie, T., Tibshirani, R. and Friedman, J.
  The Elements of Statistical Learning: Data Mining, Inference, and Prediction.
  Springer, New York, 2009.
- Pawitan, Y.,
  In All Likelihood,
  OUP, Oxford, 2001.
- Pritchard, J. K.; Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. T.
  Population Growth of Human Y Chromosomes: A Study of Y Chromosome Microsatellites,
  Mol. Biol. Evol., 16 (12): 17911798, 1999.

References III

- C.P. Robert and G. Casella
  Monte Carlo Statistical Methods (second edition)
  New York: Springer-Verlag, 2004.
- Wasserman, L.,
  All of Statistics, Springer, New York, 2003.
- Wasserman, L.,
  All of Nonparametric Statistics, Springer, New York, 2005.