# Statistics II

Richard D. Gill

Fac. of Mathematical and Natural Sciences
Leiden University

Spring 2009
Version: March 13, 2009

**Lecture 1: Intro to intro**

**Idea of course**

Q: is statistics application driven or mathematics driven?
A: wrong question

Statistical Science and Mathematical Statistics
Venn diagram of maths, statistics, computer science

Book: Venables and Ripley:
*Modern Applied Statistics with S-plus* (MASS)

Tool: R language ($\succcurlyeq$ S)

Method: alternate theory and practice
sequence lectures directed by MASS

## Lecture 2: Distributions and graphics

Convention: Lazy notation

Embellishments like sub- and superscripts, "vector" sign, ... are omitted when the context allows

The distinction between *random variables* and *constants* is that random variables are always introduced explicitly as such – everything else is a constant

In the context of matrix operations a list of random variables $\vec{X} = X = (X_1, ..., X_p)$ is treated as a column vector

## Distributions and graphics

*Distribution (law) of a random variable*
$P_X(B) = P(X \in B), \qquad B \in \mathcal{B}(\mathbb{R});$
a probability measure on the Borel sets of the real line

*Distribution function* (c.d.f., d.f.)
$F(x) = F_X(x) = P(X \leq x), \qquad x \in \mathbb{R};$
right continuous with left hand limits, non-decreasing,
$F(-\infty) = 0, \;\; F(+\infty) = 1$

*Quantile function*
$Q(p) = F^{-1}(p) = \inf\{x \in \overline{\mathbb{R}} : F(x) \geq p\}, \qquad p \in [0, 1];$
right continuous with left hand limits, non-decreasing,
$Q(0) = -\infty$, sup support $(P_X) = Q(1) \leq \infty$

Exercise. Check all of this!

## Plots

*PP-plot*
Plot of probabilities against probabilities

Plot $G = F_Y$ (vertical axis) against $F = F_X$ (horizontal)
$(F(x), G(x)) : x \in \mathbb{R}$
Graph of function $(G(F^{-1}(p)) : p \in [0, 1])$

*QQ-plot*
Plot of quantiles against quantiles

Plot $G^{-1}$ (vertical axis) against $F^{-1}$ (horizontal)
$(F^{-1}(p), G^{-1}(p)) : p \in [0, 1]$
Graph of function $(G^{-1}(F(x)) : x \in \mathbb{R})$

Empirical distribution

sample of $n$ observations $X_i, i = 1, \ldots, n$

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

a random probability measure on $\mathbb{R}$
with support the set of observed values

Empirical PP plot, QQ plot

Replace one or both distributions with empirical (sample) counterparts

Normal plot, exponential plot ...

Empirical versus theoretical ..

Two sample PP plot, QQ plot

Sample 2 versus sample 1

Empirical density, mass function

Density estimates

Empirical distribution of a statistic

Bootstrapping

(Cross-validation)

**Lecture 3: Multivariate normal distribution**

Univariate normal distribution:

$X \sim N(\mu, \sigma^2)$ means $X$ is normally distributed,
mean $\mu \in \mathbb{R}$, variance $\sigma^2 \geq 0$

Let $V = \vec{V} = (V_1, ..., V_p)$ denote $p$ i.i.d. random variables,
$V_i \sim N(0, 1)$
Possible values are $v \in \mathbb{R}^p$
Let $X = AV + b$ where $A$ is $n \times p$, $b$ is $n \times 1$.

Definition of MVN: $X \sim N_n(\mu, \Sigma)$ where $\Sigma = AA^\top$, $\mu = b$

Exercise: (i) Definition makes sense: the distribution depends only on mean $\mu$ and covariance $\Sigma$
Hint: characteristic function
(ii) Its support is the affine subspace $\mu + \mathcal{R}(\Sigma)$
(iii) $\Sigma$ nonsingular implies $X$ cts. dist. on $\mathbb{R}^n$ with density

$$\frac{\exp(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu))}{\sqrt{\det 2\pi\,\Sigma}}$$

## Properties

Property I: affine transformations

$X \sim \text{MVN} \implies AX + b \sim \text{MVN}$

Parameters: just compute mean and variance

## Property II: marginalization and conditioning

Partition $\vec{X} \sim N(\mu, \Sigma)$ as $X = (\vec{X}_1, \vec{X}_2)$

Partition $\mu$ and $\Sigma$ accordingly; then we have the *disintegration*

$$X_1 \sim N(\mu_1, \Sigma_{11})$$

$$X_2 \mid X_1 = x_1 \sim N\Big( \mu_2 + \Sigma_{21}\Sigma_{11}^{-}(x_1 - \mu_1) \ , \ \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-}\Sigma_{12} \Big)$$

$\Sigma_{11}^{-}$ is a *generalized inverse* of $\Sigma_{11}$: specifically, the one defined by

$$\Sigma_{11}^{-}\Sigma_{11} = \Sigma_{11}\Sigma_{11}^{-} = \Pi_{\mathcal{R}(\Sigma_{11})} = \Pi_{\mathcal{N}(\Sigma_{11})^{\perp}}$$

## Lecture 4: Maximum likelihood estimation

Suppose $X \sim p(x; \theta_0)$, density w.r.t. $\mu$, where $\theta_0 \in \Theta \subseteq \mathbb{R}^p$

Define:

- $\ell(\theta) = \log p(X; \theta)$, the log likelihood
- $U(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}\ell(\theta)$, the score function
- $I(\theta) = -\frac{\mathrm{d}}{\mathrm{d}\theta}U(\theta)$, the observed information

Verify:

- $E_0\ell(\theta) \leq E_0\ell(\theta_0)$, by Jenssen (equality iff $p(x;\theta) = p(x;\theta_0)$ a.e.)
- $E_0U(\theta_0) = 0$, by interchange of integration over $x$ and differentiation w.r.t. $\theta$
- $E_0I(\theta_0) = \mathrm{var}_0U(\theta_0)$, idem

Interesting (semi)-example: the location family based on the Laplace or double exponential distribution, $p(x;\theta) = \frac{1}{2}\exp(|x-\theta|)$ w.r.t. Lebesgue

Now suppose $X = (X_1, \ldots, X_n)$ where the $X_i$ are i.i.d.

For each $\theta$, $l_n(\theta)/n \to E_0\ell_1(\theta) \leq E_0\ell_1(\theta_0)$ by l.l.n.

$\widehat{\theta_n} = !\max_\theta \ell_n(\theta)/n$ maximizes a random function of $\theta$, converging pointwise in $\theta$ to a deterministic function of $\theta$, with unique maximum at $\theta = \theta_0$

Smoothness of model and (effective) compactness of parameter space gives control of oscillations of $\ell_n$ and of its "behaviour at infinity"

Pointwise convergence then extends to uniform convergence hence gives consistency of $\widehat{\theta_n}$

By mean value theorem

$$0 \;=\; U_n(\widehat{\theta}_n) \;=\; U_n(\theta_0) - I_n(\widetilde{\theta}_n)\,\left(\widehat{\theta}_n - \theta_0\right)$$

where $\widetilde{\theta}_n$ (different for each component of this vector equation) lies on the line-segment between $\widehat{\theta}_n$ and $\theta_0$

Suppose $I_n(\theta)/n$ converges not only pointwise (l.l.n.) but also uniformly in $\theta$ to $j_0(\theta) \;=\; E_0 I_1(\theta)$
Define $i_0 = i(\theta_0) = j_0(\theta_0)$
Suppose $j_0$ is continuous in $\theta$ and suppose $i_0$ is non-singular

$$n^{-1/2} U_n(\theta_0) \;=\; n^{-1} I_n(\widetilde{\theta}_n)\; n^{1/2}(\widehat{\theta}_n - \theta_0)$$

$$n^{-1/2} U_n(\theta_0) \to N_p(0, i_0) \qquad \text{as } n \to \infty, \text{ by c.l.t.}$$

$$n^{1/2}(\widehat{\theta}_n - \theta_0) \;\to\; i_0^{-1} N_p(0, i_0) \;=\; N_p(0, i_0^{-1})$$

Similarly,

$$\ell_n(\widehat{\theta}_n) \;=\; \ell_n(\theta_0) + U_n(\theta_0)^\top(\widehat{\theta}_n - \theta_0) \;-\; \tfrac{1}{2}(\widehat{\theta}_n - \widehat{\theta}_0)^\top I_n(\widetilde{\theta}_n)\,(\widehat{\theta}_n - \widehat{\theta}_0)$$

Thus

$$\ell_n(\widehat{\theta}_n) \;-\; \ell_n(\theta_0) \;\approx\; \tfrac{1}{2}(\widehat{\theta}_n - \widehat{\theta}_0)^\top I_n(\widetilde{\theta}_n)\,(\widehat{\theta}_n - \widehat{\theta}_0)$$

$$\to\; N_p(0, i_0^{-1})\, i_0\, N_p(0, i_0^{-1}) \;=\; \chi^2_p$$

The first result permits us to pretend

$$\widehat{\theta}_n - \theta_0 \;\approx\; N_p(0, I_n(\widehat{\theta}_n)^{-1})$$

The obvious (and asymptotically equivalent) tests of $\theta = \theta_0$ based on the large sample distributions of $U_n(\theta_0)$, $\widehat{\theta}_n - \theta_0$, and of $2(\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0))$, replacing *expected* Fisher information by *observed* Fisher information where convenient, are called the *Rao, Wald* and *Wilks'* tests respectively.

Note that "$n$" does not appear in the resulting approximate statistical inference (confidence regions, tests): we just need to know the log likelihood function

Since we have l.l.n. and c.l.t. in the non-identically distributed case; for martingales and Markov processes and stationary time series ... the results which we use in practice have much wider validity...

In general:
the number $n$ is replaced by a sequence of scaling constants

Consistency of the m.l.e. is proven using global considerations and the relationship with the Kulback-Leibler divergence

Asymptotic normality follows from local considerations

Must show
(i) the scaled score function is asymptotically Gaussian
(ii) the scaled observed information is asymptotically deterministic

See Aad van der Vaart's book for actual theorems, making use of *empirical process theory* (lln and clt for families of functions of the observations)

Lucien LeCam's theory of local asymptotic normality gives a deeper explanation of what is going on as well as of the asymptotic optimality (in appropriate sense) of the likelihood based inference

Interesting examples: the Laplace distribution; the Cauchy distribution (location families)
cf. Reeds' result: number of inconsistent roots of Cauchy likelihood equation is asymptotically Poisson $(1/\pi)$

**Lecture 5: Density estimation**

## Histogram as a density estimator

For $x$ in a certain bin (interval) of width $h$
the histogram estimates the density $f(x)$ by
$\widehat{f}(x) = \#\{\text{observations in bin}\}/nh$

So with $p = \int_{\text{bin}} f(y)\mathrm{d}y \approx hf(x)$, for large $n$ and small $h$, and for a "typical" $x$ half way between the midpoint of the bin and the boundary of the bin

$$\operatorname{var}\widehat{f}(x) = \frac{np(1-p)}{(nh)^2} \approx \frac{f(x)}{nh}$$

$$\operatorname{bias}\widehat{f}(x) = \frac{p}{h} - f(x) \approx \pm\frac{f'(x)h}{4}$$

The mean square error = variance + (bias)$^2$ is therefore of the order

$$\frac{f(x)}{nh} + \frac{f'(x)^2 h^2}{16}$$

This is a convex function of $h$, minimal where its derivative is zero, i.e., where

$$-\frac{f(x)}{nh^2} + \frac{f'(x)^2 h}{8} = 0$$

Optimal bin-width is therefore

$$h_{\text{opt}} = \left(\frac{8 f(x)}{n f'(x)^2}\right)^{\frac{1}{3}}$$

Suppose instead we use the frequency polygon as an estimator of $f$, i.e., join the midpoints of the histogram bars with straight lines

For $x$ on the boundary of two bins

$$\operatorname{var}\widehat{f}(x) = \frac{1}{4}\frac{np_1(1-p_1)}{(nh)^2} + \frac{1}{4}\frac{np_2(1-p_2)}{(nh)^2} \approx \frac{f(x)}{2nh}$$

$$\operatorname{bias}\widehat{f}(x) = \frac{1}{2}\frac{p_1}{h} + \frac{1}{2}\frac{p_2}{h} - f(x) \approx \frac{f''(x)h^2}{6}$$

**Exercise**. Check that for general $x$ one has a similar expansion, only the constants differ.

The mean square error = variance + (bias)$^2$ is therefore of the order

$$\frac{f(x)}{2nh} + \frac{f''(x)^2 h^4}{36}$$

This is a convex function of $h$, minimal where its derivative is zero, i.e., where

$$-\frac{f(x)}{2nh^2} + \frac{f''(x)^2 h^3}{9} = 0$$

Optimal bin-width is therefore

$$h_{\text{opt}} = \left(\frac{9 f(x)}{2n f''(x)^2}\right)^{\frac{1}{5}}$$

## Lecture 6: Linear models

$$y = X\beta + \varepsilon$$

$$(n \times 1) = (n \times p)(p \times 1) + (n \times 1)$$

- $E(y) = X\beta$
- $\text{var}(y) = \sigma^2 I_n$
- $\varepsilon \sim$ multivariate normal

- $y$: vector of observed responses
- $X$: design matrix, each column represents one covariate, fixed and known
- $\beta$ and $\sigma^2$: unknown parameters

The least squares estimator of $\beta$ is a minimizer of
$(y - Xb)^\top (y - Xb) = \|y - Xb\|^2$.
With some matrix algebra one finds that it is unique and equal to $\widehat{\beta} = (X^\top X)^{-1} X^\top y$ *iff* the matrix $X^\top X$ is nonsingular.
Otherwise it is not-unique, but one choice is found by substituting the generalized inverse $(X^\top X)^-$ for the inverse.
Define also $\widehat{y} = X\widehat{\beta}$ and $\widehat{\varepsilon} = y - \widehat{y}$;
these are called the fitted values and the residuals respectively (and are unique, even if $\widehat{\beta}$ isn't).

By its definition,
$\widehat{y}$ is the orthogonal projection of $y$ onto $\mathrm{col}(X)$, the column space of $X$
$\widehat{\varepsilon}$ is the orthogonal projection of $y$ onto $\mathrm{col}(X)^{\perp}$.
Since $Ey = X\beta$ lies in $\mathrm{col}(X)$,
$\widehat{y}$ equals $X\beta$ plus the orthogonal projection of $\varepsilon$ onto $\mathrm{col}(X)$,
$\widehat{\varepsilon}$ is the orthogonal projection of $\varepsilon$ onto $\mathrm{col}(X)^{\perp}$.
Moreover, $\widehat{\beta}$ is a minimizer of $\|\widehat{y} - Xb\|^2$,
and equivalently a solution of $\widehat{y} - Xb = 0$,
unique if and only if the columns of $X$ are linearly independent.

Under the complete set of model assumptions, $\varepsilon \sim N_n(0, \sigma^2 I)$.
Choose an orthonormal basis of $\mathbb{R}^n$ such that the first
$q = \text{rank}(X) \leq p$ elements span $\text{col}(X)$,
the remaining elements span $\text{col}(X)^{\perp}$.
Express $\varepsilon$ in this basis, partitioned as $(\varepsilon_1, \varepsilon_2)$.
By the rotational symmetry of the $N_n(0, \sigma^2 I)$ distribution,
$\varepsilon_1$ and $\varepsilon_2$ are independent and
$N_q(0, \sigma^2 I)$ and $N_{n-q}(0, \sigma^2 I)$ distributed respectively.
$\widehat{y}$ and $\widehat{\beta}$ are affine functions of $\varepsilon_1$,
while (in this basis) $\widehat{\varepsilon} = (0_1, \varepsilon_2)$.

The projector onto $\mathrm{col}(X)$ is $X(X^\top X)^- X^\top$,
the projector onto $\mathrm{col}(X)^\perp$ is $I - X(X^\top X)^- X^\top$.
Recall: a projector $\Pi$ is characterized by $\Pi^2 = \Pi = \Pi^\top$;
$I - \Pi$ is also a projector and $\mathrm{trace}(\Pi) = \dim(\mathrm{range}(\Pi))$.
We find that $\widehat{y} \sim N_n(X\beta, \sigma^2 X(X^\top X)^- X^\top)$,
$\widehat{\varepsilon} \sim N_n(0, \sigma^2(I - X(X^\top X)^- X^\top))$,
and the two are independent.

If $\widehat{\beta}$ is unique, it is independent of $\widehat{\varepsilon}$, and
$\widehat{\beta} \sim N_p(\beta, \sigma^2(X^\top X)^{-1})$.
It follows that $\|\widehat{\varepsilon}\|^2 = \|\varepsilon_2\|^2 \sim \sigma^2 \chi^2_{n-q}$, independent of $\widehat{\beta}$, and

$$\frac{\widehat{\beta}_i - \beta_i}{\sqrt{\left(\widehat{\sigma}^2(X^\top X)^{-1}\right)_{ii}}} \sim t_{n-q}$$

**Exercise**. What does $\widehat{\beta} = (X^\top X)^- X^\top y$ estimate when $\text{rank}(X) < p$? Show that $\widehat{\beta}$ is the solution of the least squares problem which furthermore minimizes $\|b\|^2$

Consider now a (linear) submodel of our linear model $y = X\beta + \varepsilon$.
That is a linear model $y = Y\gamma + \varepsilon$ such that $\mathrm{col}(Y)$ is a subspace of $\mathrm{col}(X)$.
Let the dimension of this subspace be $r < q$.
We refer to the models as the small and the large model.
Choose an orthonormal basis of $\mathbb{R}^n$ such that the first $r$ elements span $\mathrm{col}(Y)$ and together with the next $q - r$ span $\mathrm{col}(X)$.
Partition $\varepsilon$, expressed in this basis, as $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$.

We see that the residual sums of squares in the two models are $\|\varepsilon_3\|^2$ (big model) and
$\|\varepsilon_2\|^2 + \|\varepsilon_3\|^2$ (small model)
and hence that, if the submodel is true,

$$\frac{\text{reduction in SSR}/(q-r)}{\text{SSR of large model}/(n-q)} \sim F_{q-r,n-q}$$

If the larger model is true, but the submodel is not true, then we have a non-central chi-square distribution (**exercise**: which?).

**Exercise**. I separated the model assumptions into assumptions on the expectation, variance, and distribution of $y$. Investigate what remains true when we drop the distributional assumption; and when we also drop the variance assumption.