USING STATISTICS IN RESEARCH

David A. Stephens

Department of Mathematics, Imperial College

d.stephens@imperial.ac.uk
stats.ma.ic.ac.uk/~das01/StatsShortCourse/

3rd December, 2003

WEEK 5

ASSOCIATION AND AGREEMENT MEASURES,

SURVIVAL ANALYSIS

&

DISCRIMINATION AND CLASSIFICATION

SECTION 9.

OBSERVER AGREEMENT

Assessing the results of diagnostic procedures and the effects of therapies often involves subjective judgements. Observer agreement studies are conducted to investigate the level of consensus on such assessments.

Typically, several observers make assessments on each of a series of subjects and these assessments are compared.

An important consideration for study design is the presence of both **within-observer** and **between-observer** variation. The apparent disagreement between observers may be due to either one of these components or both. It is important to distinguish between them, as any action taken to reduce disagreement will depend on which type of variation dominates. To do this, we require observations repeated by the same observer.

We might consider any of the following types of observer agreement studies

- studies with binary assessments and designs;
- where each of two observers assesses all subjects once,
- where each observer assesses all subjects twice
- where each observer assesses a proportion of the subjects once and the remainder twice.

and so on.

NOTE: Sample-size calculations are conventionally based upon hypothesis-testing theory Observer-agreement studies, however, are designed to estimate the level of observer agreement. Moreover, unlike clinical trials, there are no obvious hypotheses to test. The hypothesis of perfect agreement between observers is unrealistic and the hypothesis of agreement purely by chance is also unrealistic in most circumstances.

Rejection of such a hypothesis does not provide useful information since the investigator needs to know more than the fact that the observed level of agreement is unlikely to be due to chance.

9.1 CONTINUOUS MEASUREMENTS

9.1.1 THE INTRACLASS CORRELATION

The Intraclass Correlation (ICC) assesses rating reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects. The theoretical formula for the ICC is:

$$\rho = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_\epsilon^2}$$

where

- σ_S^2 is the **between subjects** variability
- σ_ϵ^2 is the **within subjects** variability

These quantities are directly estimable from ANOVA analyses.

In a one-way ANOVA with K groups, we have the ANOVA table

Source	D.F.	Sum of squares	Mean square
Between Samples	$K-1$	FSS	$FSS/(K-1)$
Within Samples	$n-K$	RSS	$RSS/(n-K)$
Total	$n-1$	TSS	

where

$$TSS = \sum_{k=1}^{K}\sum_{j=1}^{n_k}(y_{kj}-\overline{y}_{..})^2 \quad RSS = \sum_{k=1}^{K}\sum_{j=1}^{n_k}(y_{kj}-\overline{y}_{k})^2 \quad FSS = \sum_{k=1}^{K} n_k\,(\overline{y}_{k}-\overline{y}_{..})^2$$

Then

$$\widehat{\sigma}^2_{\epsilon} = \frac{RSS}{n-K} \qquad \widehat{\sigma}^2_{S} = \frac{FSS/(K-1) - RSS/(n-K)}{K}$$

9.1.2 DIFFERENT TYPES OF ICC

In their paper, Shrout and Fleiss (1979) describe three classes of ICC for reliability, which they term Case 1, Case 2 and Case 3. Each Case applies to a different rater agreement study design.

- **Case 1:** Raters for each subject are selected at random
  - This case has a pool of raters. For each subject, one randomly samples from the rater pool k different raters to rate this subject. Therefore the raters who rate one subject are not necessarily the same as those who rate another. This design corresponds to a one-way ANOVA in which Subject is a random effect, and Rater is viewed as measurement error.

- **Case 2:** The same raters rate each case. These are a random sample.
  - The same set of k raters rate each subject. This corresponds to a fully-crossed (rater $\times$ subject) two-way ANOVA design in which both Subject and Rater are separate effects.
  - In Case 2, Rater is considered a **random effect**; this means the k raters in the study are considered a random sample from a population of potential raters.
  - The Case 2 ICC estimates the reliability of the larger population of raters.

- **Case 3:** The same raters rate each case. These are the only raters.
  - This is similar to Case 2; a fully-crossed, two-way ANOVA design. But here one estimates the ICC that applies only to the k raters in the study. Since this does not permit generalization to other raters, the Case 3 ICC is not often used.

Shrout and Fleiss (1981) also show that for each of the three Cases above, one can use the ICC in two ways:

- To estimate the reliability of a single rating, or
- To estimate the reliability of a mean of several ratings.

For each of the Cases, then, there are two forms, producing a total of 6 different versions of the ICC.

9.2 DISCRETE MEASUREMENTS

9.2.1 THE KAPPA STATISTIC

Options for discrete data observer agreement analysis are rather more limited; one simple measure of agreement between two raters is the **Kappa Statistic.**

For a $K \times K$ table of results for the observer assessments of two observers on a categorical scale,let n_{ij} is the number of times rater 1 accords a measure i whilst Rater 2 accords a measure j, for $i, j = 1, ..., K$.

- "Considerable Agreement": Diagonal elements "large"
- "Low Agreement": Off-diagonal elements "large"

EXAMPLE: Assessment of xeromammograms by two radiologists

	Radiologist B				
Radiologist A	Normal	Benign	Suspected	Cancer	Total
Normal	21	12	0	0	33
Benign	4	17	1	0	22
Suspected	3	9	15	2	29
Cancer	0	0	0	1	1
Total	28	38	16	3	85

Proportion of Agreements:

$$p_A = \frac{n_A}{n} = \frac{(21 + 17 + 15 + 1)}{85} = 0.64$$

However, this does not take into account/quantify the probability of "chance" agreements; this can be measured by the expected number of chance agreements

$$\widehat{n}_A = \frac{33 \times 28}{85} + \frac{22 \times 38}{85} + \frac{29 \times 16}{85} + \frac{1 \times 3}{85} = 26.2$$

which gives a proportion

$$\widehat{p}_A = \frac{\widehat{n}_A}{n} = \frac{26.2}{85} = 0.31$$

Hence the "excess agreement" in the observed data is

$$\kappa = \frac{p_A - \widehat{p}_A}{1 - \widehat{p}_A}$$

which is termed the **Kappa Statistic.** Guidelines for interpretation of κ are

- $\kappa \leq 0.20 \Longrightarrow$ Poor Agreement
- $0.20 < \kappa \leq 0.40 \Longrightarrow$ Fair Agreement
- $0.40 < \kappa \leq 0.60 \Longrightarrow$ Moderate Agreement
- $0.60 < \kappa \leq 0.80 \Longrightarrow$ Good Agreement
- $0.80 < \kappa \leq 1.00 \Longrightarrow$ Very Good Agreement

Standard errors for κ are also available

9.2.2 WEIGHTED KAPPA

A weighted version of the Kappa Statistic can be used to reflect the ordinal nature of many observation scales (e.g. Normal⟶Benign⟶Suspected⟶Cancer)

Each off-diagonal element in the agreement table is given a weight reflecting how "severe" the disagreement is; usually the weights are proportional to the distance from the diagonal. This gives a **weighted kappa**, κ_W

$$\kappa_W = \frac{p_A^{(W)} - \widehat{p}_A^{(W)}}{1 - \widehat{p}_A^{(W)}}$$

where

$$p_A^{(W)} = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij} n_{ij} \qquad \widehat{p}_A^{(W)} = \frac{1}{n^2}\sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij} n_{i.} n_{.j}$$

where $n_{i.}, n_{.j}$ are the row an column totals for row i and column j respectively.

SECTION 10.

SURVIVAL ANALYSIS

Survival (or lifetime, or time-to-event) analysis is a special type of regression modelling that explains the observed variability in a **response** variable Y via consideration of **predictors** $X = (X_1, ..., X_K)$. The principal difference between survival analysis and conventional regression is that account is taken of potential **censoring** in the response variable

- we may observe some actual responses (survival, failure) times, but also some censored responses where we do not observe an actual failure but rather only that the failure occurs after a **censoring time** (the end of study) – this is called **right-censoring**

- the response data is thus bivariate (Y, Z) where Y is the time at which the response is measured, and

$$Z = \begin{cases} 1 & \text{Failure is observed} \\ 0 & \text{Censored} \end{cases}$$

- occasionally, we observe **left-censoring** or **interval-censoring**

The potential presence of censoring fundamentally changes how we view the modelling process; previously we have looked at probability densities and Expected responses.

We now take an alternative view, and examine **survivor** and **hazard** functions.

10.1 THE SURVIVOR FUNCTION

The probability **density function** for response variable Y is f_Y, and the expectation, likelihood function and so on that are required for regression modelling are formed from f_Y. The **distribution function** F_Y is

$$F_Y(y) = P[Y \leq y] = \int_0^y f_Y(t)\, dt$$

In conventional regression modelling, the likelihood contribution for data point i with response y_i is $f_Y(y_i)$. For right-censored data with censoring at y_i, however, the likelihood contribution is

$$P[Y > y_i] = 1 - F_Y(y_i)$$

(i.e. we have "observed" that $Y_i > y_i$, the survival was at least y_i). This motivates consideration of the **survivor (reliability) function**

$$S_Y(y) = 1 - F_Y(y)$$

The likelihood function is thus

$$\left\{ \prod_{i:Z_i=1} f_Y(y_i) \right\} \times \left\{ \prod_{i:Z_i=0} S_Y(y_i) \right\}$$

that is

LIKELIHOOD FOR UNCENSORED DATA

$\times$

LIKELIHOOD FOR CENSORED DATA

and the role of the predictors can be introduced via the parameters of f_Y and F_Y.

10.2 THE HAZARD FUNCTION

As a further alternative method of specification, we consider the **hazard function**

$$h_Y(y) = P\left[\text{Failure at } y | \text{Survival } \geq y\right]$$

$$= \frac{f_Y(y)}{S_Y(y)}$$

and the **integrated hazard**

$$H_Y(y) = \int_0^y h_Y(t)\, dt$$

and it can be shown that

$$S_Y(y) = \exp\left\{-H_Y(y)\right\}$$

10.3 THE KAPLAN-MEIER CURVE

The **Kaplan-Meier curve** is a non-parametric estimate of the survivor function; it takes into account the censored data and produces a decreasing "step-function" curve, where the downward steps take place at the times of the failures, giving the estimated survival function at the jth failure/censoring time as

$$\widehat{S}_j = \prod_{i=1}^{j} \left(1 - \frac{z_i}{n - i + 1}\right)$$

This curve can be used to report an estimated survival probability at a given time (1 year, 5 years etc.).

Standard errors for these estimated survival probabilities are also available.

10.4 THE COX REGRESSION MODEL

The **Cox** (or **Proportional Hazards**) model provides a simple way of introducing the influence of predictors into the survival model. The basic components are a **baseline hazard** function, h_0 and a linear predictor and (positive) link function g (similar to the GLM modelling of previous chapters). Then for an experimental unit with observed predictor values $X_1 = x_1$, $X_2 = x_2, ..., X_K = x_K$, the hazard function takes the form

$$h_Y(y; x) = g(x^T \beta) h_0(y)$$

that is, the hazard is modified in a multiplicative fashion by the linked-linear predictor.

Typically, g is the exponential function.

From the previously established relationships,

$$S_Y(y;x) = \exp\left\{-\int_0^y h_Y(t)\ dt\right\} = \exp\left\{-\int_0^y g(x^T\beta)h_0(y)\ dt\right\}$$

If a coefficient β_k is positive, the hazard is **increased**, and the expected failure time **decreased.**

The relevance/significance of a particular predictor is assessed using a **Wald** test based on the magnitude of

$$\frac{\widehat{\beta}}{s.e.\left(\widehat{\beta}\right)}$$

10.5 THE ACCELERATED LIFE MODEL

The **Accelerated Life** model provides another way of introducing the influence of predictors into the survival model. The basic components now are a **baseline survivor** function, S_0 and a linear predictor and (positive) link function g as above. Then for an experimental unit with observed predictor values $X_1 = x_1$, $X_2 = x_2, ..., X_K = x_K$, the survivor function takes the form

$$S_Y(y; x) = S_0(g(x^T\beta)y)$$

that is, the time scale is modified in a multiplicative fashion by the linked-linear predictor.

Again, typically, g is the exponential function. This model allows direct modelling of the influence of predictors on survival.

10.6 THE LOG-RANK TEST

The **log-rank** test is a standard test for significant differences between two (or more) survivor functions that differ because of the influence of the different levels of a discrete predictor.

$$\begin{aligned} H_0 &: S_1 = S_2 \\ H_1 &: S_1 \neq S_2 \end{aligned}$$

It is a non-parametric test based on ranks of samples for the two or more subgroups.

Asymptotic or exact versions of the test can be carried out; SPSS and other packages give further alternatives.

10.7 PARAMETRIC MODELLING

It is possible to fit and compare **parametric** survival models to such data. Parametric densities, survivor functions, hazards etc. can be readily used in the formation of a likelihood, potentially within the proportional hazards/accelerated life framework.

Typical models used are

- Weibull
- Gamma
- Log-Logistic
- Log-Normal
- Pareto

SECTION 11.

CLASSIFICATION AND DISCRIMINATION

Classification is another special type of regression modelling that explains the observed variability in a **response** variable Y via consideration of **predictors** $X = (X_1, ..., X_K)$. The principal difference between classification and conventional regression is that the response variable is a **nominal categorical variable**, that is, for data item i

$$Y_i \in \{1, 1, 2, ...K\}$$

so that the value of Y_i is a **label** rather than a numerical value, where the label represents the **group** or **class** to which that item belongs.

We again wish to use the **predictor** information in X to allocate Y to one of the classes

Thus, there are two main goals:

- to partition the observations into two or more labelled classes. The emphasis is on **deriving a rule** that can be used to **optimally assign** a new object to the labeled classes.
  - This is the process of **CLASSIFICATION**
- to describe either graphically or algebraically, the different features of observations from several known collections. We attempt to find **discriminants** whose numerical values are such that the collections are separated as much as possible.
  - This is the process of **DISCRIMINATION**

Both are special cases of what we have previously termed **MULTIVARIATE ANALYSIS**

Typically, the exercise of classification will be **predictive**, that is,

- we have a set of data available where both the **response** and **predictor** information is known
  - these data are the **training** data
- we also have a set of data where only the **predictor** information is known, and the **response** is to be predicted
  - these data are the **test** data
- often we will carry out an exercise of **model-building** and **model-testing** on a given data set by extracting a **training set**, building a model using the training data, whilst holding back a proportion (the **test set**) for model-testing.

11.1 CLASSIFICATION FOR TWO CLASSES ($K = 2$)

Let $f_1(x)$ and $f_2(x)$ be the probability functions associated with a (vector) random variable X for two populations 1 and 2. An object with measurements x must be assigned to either class 1 or class 2. Let $\mathbb{X}$ denote the sample space. Let $\mathcal{R}_1$ be that set of x values for which we classify objects into class 1 and $\mathcal{R}_2 \equiv \mathbb{X} \backslash \mathcal{R}_1$ be the remaining x values, for which we classify objects into class 2.

The **conditional probability**, $P(2|1)$, of classifying an object into class 2 when, in fact, it is from class1 is:

$$P(2|1) = \int_{\mathcal{R}_2} f_1(x)\ dx.$$

Similarly, the conditional probability, $P(1|2)$, of classifying an object into class 1 when, in fact, it is from class 2 is:

$$P(1|2) = \int_{\mathcal{R}_1} f_2(x)\ dx$$

Let p_1 be the *prior* probability of being in class 1 and p_2 be the *prior* probability of 2, where $p_1 + p_2 = 1$. Then,

$$\begin{aligned} P(\text{Object correctly classified as class 1}) &= P(1|1)p_1 \\ P(\text{Object misclassified as class 1}) &= P(1|2)p_2 \\ P(\text{Object correctly classified as class 2}) &= P(2|2)p_2 \\ P(\text{Object misclassified as class 2}) &= P(2|1)p_1 \end{aligned}$$

Now suppose that the *costs* of misclassification of a class 2 object as a class 1 object, and vice versa are, respectively.

$$c(1|2) \qquad \text{and} \qquad c(2|1).$$

Then the expected cost of misclassification is therefore

$$c(2|1)P(2|1)p_1 + c\,(1|2)\,P(1|2)p_2.$$

The idea is to choose the regions $\mathcal{R}_1$ and $\mathcal{R}_2$ so that this expected cost is minimized. This can be achieved by comparing the predictive probability density functions at each point x

$$\mathcal{R}_1 \equiv \left\{ x : \frac{f_1\,(x)}{f_2\,(x)} \frac{p_1}{p_2} \geq \frac{c\,(1|2)}{c\,(2|1)} \right\}$$

or, equivalently

$$\mathcal{R}_2 \equiv \left\{ x : \frac{f_1\,(x)}{f_2\,(x)} \frac{p_1}{p_2} < \frac{c\,(1|2)}{c\,(2|1)} \right\}$$

or indeed minimizing the total probability of misclassification

$$p_1 \int_{\mathcal{R}_2} f_1\,(x)\;\; dx + p_2 \int_{\mathcal{R}_1} f_2\,(x)\;\; dx$$

If $p_1 = p_2$, then

$$\mathcal{R}_1 \equiv \left\{ x : \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)} \right\}$$

and if $c(1|2) = c(2|1)$, equivalently

$$\mathcal{R}_1 \equiv \left\{ x : \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1} \right\}$$

and finally if $p_1 = p_2$ and $c(1|2) = c(2|1)$ then

$$\mathcal{R}_1 \equiv \left\{ x : \frac{f_1(x)}{f_2(x)} \geq 1 \right\} \equiv \{ x : f_1(x) \geq f_2(x) \}$$

11.2 CLASSIFICATION FOR TWO NORMAL SAMPLES

Suppose that we have two (multivariate) normal classes (in d dimensions), that is where

- **class 1**: $X \sim N_d(\mu_1, \Sigma_1)$

$$f_1(x) = \left(\frac{1}{2\pi}\right)^{d/2} \frac{1}{|\Sigma_1|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)\right\}$$

- **class 2**: $X \sim N_d(\mu_2, \Sigma_2)$

$$f_2(x) = \left(\frac{1}{2\pi}\right)^{d/2} \frac{1}{|\Sigma_2|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2)\right\}$$

We sometimes assume that $\Sigma_1 = \Sigma_2 = \Sigma$ (*homogeneity of variances).* Using the previous formula, we identify the following **classification rule**; we allocate an observation with predictor variable x_0 to class 1 if

$$(\mu_1 - \mu_2)^T \Sigma^{-1} x_0 - \frac{1}{2}(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 + \mu_2) \geq \log\left[\frac{c(1|2)\, p_2}{c(2|1)\, p_1}\right]. \quad (1)$$

More generally, if $\Sigma_1 \neq \Sigma_2$, we allocate an observation with predictor variable x_0 to class 1 if

$$-\frac{1}{2} x_0^T \left(\Sigma_1^{-1} - \Sigma_2^{-1}\right) x_0 + \left(\mu_1^T \Sigma_1^{-1} - \mu_2^T \Sigma_2^{-1}\right) x_0 - k \geq \log\left[\frac{c(1|2)\, p_2}{c(2|1)\, p_1}\right] \quad (2)$$

where

$$k = \frac{1}{2}\log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}\left(\mu_1^T \Sigma_1^{-1} \mu_1 - \mu_2^T \Sigma_2^{-1} \mu_2\right)$$

The parameters μ_1, μ_2 and Σ, Σ_1 and Σ_2 may be estimated from training data.

- if the covariance matrices are presumed **equal** then we have a total of

$$2d + \frac{1}{2}d\,(d+1)$$

  parameters to estimate

- if the covariance matrices are presumed **unequal** then we have a total of

$$2d + d\,(d+1)$$

  parameters to estimate

Thus with limited data in d dimensions, we may be limited in the type of analysis can be done. In fact, we may have to further restrict the type of covariance structure that we may assume; for example, we might have to restrict attention to

- **diagonal** covariance matrices ($2d$ parameters in total),
- or an assumption of **sphericity** ($2\,(d+1)$ parameters in total)

Despite their simplicity, such models often work well in practice.

11.3 DISCRIMINATION

Discriminant analysis works in a very similar fashion; from equations (1) and (2) we note that the boundary between regions $\mathcal{R}_1$ and $\mathcal{R}_2$ takes one of two forms

- **Equal covariances**: we have a **straight line/plane** defined by an equation of the form

  $$A_1 x + a_0$$

  where A_1 is a $d \times d$ matrix

- **Unequal covariances**: we have a **quadratic surface** defined by an equation of the form

  $$x^T B_2 x + B_1 x + b_0$$

  where B_1 and B_2 are $d \times d$ matrices.

11.4 ASSESSMENT OF CLASSIFICATION ACCURACY

The performance of a classification rule can be achieved in a number of ways: we can examine

- the **within-sample** classification error: the proportion of elements in the training sample that are misclassified by the rule
- the **leave-one-out** classification error: the proportion of elements in the training sample when the model is built (that is, the parameters are estimated) on a training sample that omits a single data point, and then attempts to classify that point on the trained model

- an m-**fold cross-validation** : the data are split into m subsamples of equal size, and one is selected at random to act as a **pseudo-test** sample. The remaining data are used as **training** data to build the model, and the prediction accuracy on the pseudo-test sample is computed. This procedure is repeated for all possible splits, and the prediction accuracy computed as a average of the accuracies over all of the splits.

- accuracy using **bootstrap resampling** to achieve the **cross-validation** based estimates of accuracy from above.

The theory behind the assessment of classification accuracy is complex.

11.5 ROC CURVES

Receiver Operating Characteristic (ROC) curves can also be used to compare the classification performance classifiers. We consider the results of a particular classifier for two populations, say one population with a disease, the other population without the disease. Suppose that a single characteristic, x, is to be used to classify individuals.

The classification procedures above reduce to a simple rule; we classify an individual to class 1 if

$$x < t_0$$

for some threshold t_0, and to class 2 otherwise. We then consider the following quantities:

- **Sensitivity**: probability that a test result will be positive when the disease is present (true positive rate, expressed as a percentage).
- **Specificity**: probability that a test result will be negative when the disease is not present (true negative rate, expressed as a percentage).
- **Positive likelihood ratio**: ratio between the probability of a positive test result given the presence of the disease and the probability of a positive test result given the absence of the disease

$$\frac{\text{True Positive Rate}}{\text{False Positive Rate}}$$

- **Negative likelihood ratio:** ratio between the probability of a negative test result given the presence of the disease and the probability of a negative test result given the absence of the disease

$$\frac{\text{False Negative Rate}}{\text{True Positive Rate}}$$

- **Positive predictive value**: probability that the disease is present when the test is positive (expressed as a percentage).
- **Negative predictive value:** probability that the disease is not present when the test is negative (expressed as a percentage).

		Disease Class		
		1	2	Total
Predicted Class	1	a	c	$a+c$
	2	b	d	$b+d$
	Total	$a+b$	$c+d$	$a+b+c+d$

- Sensitivity:/Specificity:

$$Sensitivity : \frac{a}{a+b} \qquad Specificity : \frac{d}{c+d}$$

- Likelihood Ratios

$$PLR = \frac{Sensitivity}{1 - Specificity} \qquad NLR = \frac{1 - Sensitivity}{Specificity}$$

- Predictive Values

$$PPV = \frac{a}{a+c} \qquad NPV = \frac{b}{b+d}$$

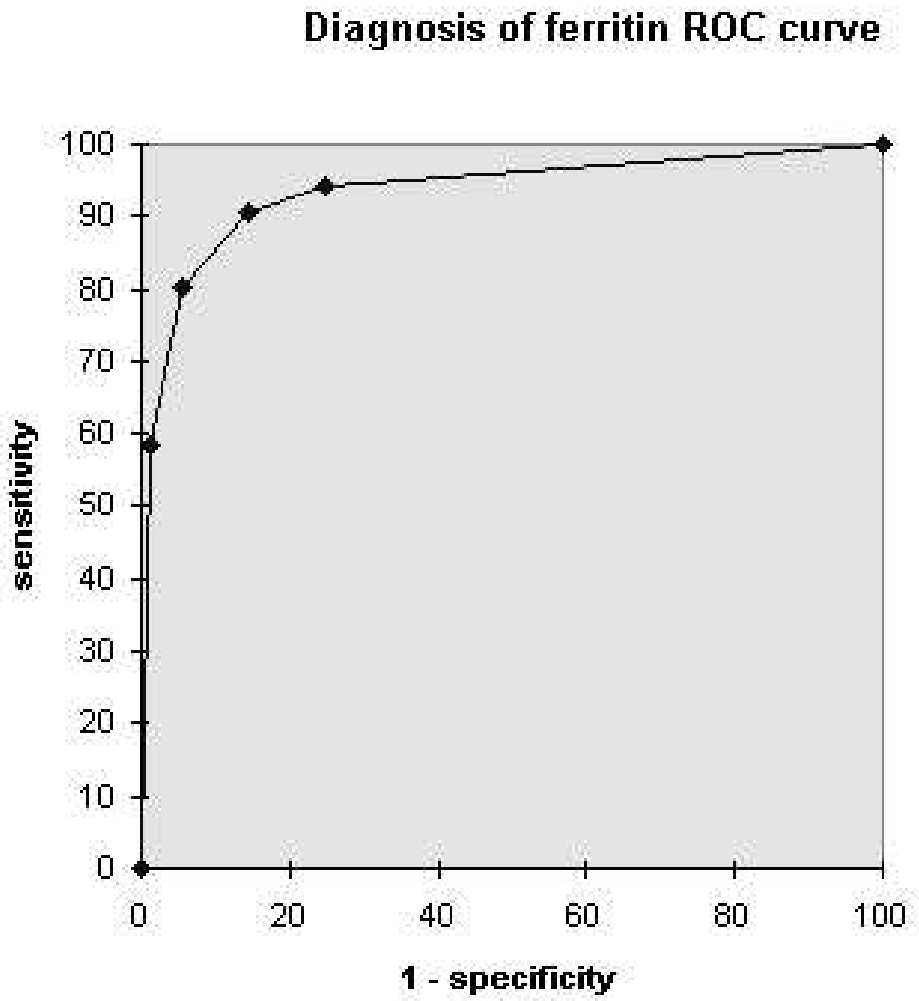
As the classifier producing the predicted class depends on the threshold t_0, we can produce a plot of how these quantities change as t_0 changes.

If we plot

$$\begin{array}{lcll} y(t_0) & : & \textit{Sensitivity} \text{ at } t_0 & \text{(CORRECTED)} \\ x(t_0) & : & 1 - \textit{Specificity} \text{ at } t_0 & \end{array}$$

then we obtain an **ROC curve**;

- for a good classifier would rise steeply and then flatten off ; such a curve would have a large area underneath it on the unit square (the domain of $(x(t_0), y(t_0))$)
- for a poor classifier would be have an ROC curve near the line $y = x$.


Diagnosis of ferritin ROC curve
sensitivity
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
1 - specificity

ROC Curve

11.6 GENERAL CLASSIFICATION SCHEMES

The general exercise of classification can be seen as a exercise in **regression modelling** for a nominal categorical variable. Previously, we studied **regression**, and more briefly **generalized linear regression.**

- For a **binary response,** or a **two-class** problem, we can use **logistic** or **binary** regression
- For a **multinomial response,** or a **multi-class** problem we can use **multinomial** regression

Because of this regression context, we can use all the previous tools for analysis in regression models that we have used previously.