

Receiver Operating Characteristics Curves

DANIEL B. WRIGHT

Volume 4, pp. 1718–1721

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Receiver Operating Characteristics Curves

The receiver operating characteristics curve, better known as ROC and sometimes called *receiver operating curve* or *relative operating characteristics*, is the principal graphical device for **signal detection theory (SDT)**. While ROCs were originally developed in engineering and psychology [3], they have become very common in medical diagnostic research (see [6] for a recent textbook, and also journals such as *Medical Decision Making*, *Radiology*, *Investigative Radiology*).

Consider an example: I receive about 100 emails a day, most of which are unsolicited junk mail. When each email arrives, it receives a ‘spamicity’ value from an email filter. Scores close to 1 mean that the filter predicts the email is ‘spam’; scores close to 0 predict the email is nonspam (‘ham’). I have to decide above what level of spamicity, I should automatically delete the emails. For illustration, suppose the data for 1000 emails and two different filters are those shown in Table 1.

The two most common approaches for analyzing such data are **logistic regression** and SDT. While in their standard forms, these are both types of the **generalized linear model**, they have different origins and different graphical methods associated with them. However, the basic question is the same: what is the relationship between spamicity and whether the email is spam or ham? ROCs are preferred when the decision criterion is to be determined and when the decision process itself is of interest, but when discrimination is important researchers should choose the logistic regression route. Because both

these approaches are based on similar methods [1], it is sometimes advisable to try several graphical methods and use the ones which best communicate the findings.

There are two basic types of curves: empirical and fitted. Empirical curves show the actual data (sometimes with slight modifications). Fitted curves are based on some model. The fitted curves vary on how much they are influenced by the data versus the model. In general, the more parameters estimated in the model, the more influence the data will have. Because of this, some ‘fitted’ models are really just empirical curves that have been smoothed (*see Kernel Smoothing*) (see [5] and [6] for details of the statistics underlying these graphs).

The language of SDT is explicitly about accuracy and focuses on two types: sensitivity and specificity. Sensitivity means being able to detect spam when the email is spam; and specificity means just saying an email is spam if it is spam. In psychology, these are usually referred to as hits (or true positives) and correct rejections. The opposite of specificity is the false alarm rate: the proportion of time that the filter predicts that real email. Suppose the data in Table 1 were treated as predicting ham if spamicity is 0.5 or less and predicting spam if it is above. Table 2 shows the breakdown of hits and false alarms by whether an email is or is not spam, and whether the filter declassifies it as spam or ham. Included also are the calculations for hit and false alarm rates. The filters only provide a spamicity score. I have to decide above what level of spamicity the email should be deleted. The choice of criterion is important. This is dealt with later and is the main advantage of ROCs over other techniques.

It is because all the values on one side of a criterion are classified as either spam or ham that ROCs are cumulative graphs. Many statistical packages

Table 1 The example data used throughout this article. Each filter had 1000 emails, about half of which were spam. The filter gave each email a spamicity rating

	Spamicity ratings										
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Filter one											
Email is ham	47	113	30	122	104	41	21	20	4	1	0
Email is spam	0	2	3	10	78	66	27	114	3	28	166
Filter two											
Email is ham	199	34	0	44	11	40	17	8	21	66	63
Email is spam	51	37	3	20	21	45	21	11	17	74	197

2 Receiver Operating Characteristics Curves

Table 2 The decision table if the criterion to delete emails was that the spamicity scores be 0.6 or above. While the two filters have similar hit rates, the first filter has a much lower false alarm rate (i.e., better specificity)

	Filter 1			Filter 2			Total
	≤ 0.5	> 0.6	Rate	≤ 0.5	> 0.6	Rate	
Is ham	457 correct rejection	46 false alarm	9%	328 correct rejection	175 false alarm	35%	503
Is spam	159 miss	338 hit	68%	177 miss	320 hit	64%	497
Total	616	384		505	495		1000

have cumulative data transformation functions (for example, in SPSS the CSUM function), so this can be done easily in mainstream packages. In addition, good freeware is available (for example, ROCKIT, RscorePlus), macros have been written for some general packages (for example, SAS and S-Plus), and other general packages contain ROC procedures (for example, SYSTAT). For each criterion, the number of hits and false alarms is divided by the number of spam and ham emails, respectively.

Figures 1(a) and 1(b) are the standard empirical ROCs and the fitted binormal curves (using RscorePlus [2]). The program assumes ham and spam vary on some dimension of spaminess (which is related

to, but not the same as, the variable spamicity) and that these distributions of ham and spam are normally distributed on this dimension. The normal distribution assumption is there for historical reasons though nowadays researchers often use the logistic distribution, which yields nearly identical results and is simpler mathematically; the typical package offers both these alternatives plus others. In psychology, usually normality is assumed largely because Swets [3] showed that much psychological data fits with this assumption (and therefore also with the logistic distribution). In other fields like medicine, this assumption is less often made [6]. ROCs usually show the concave pattern of Figures 1(a) and

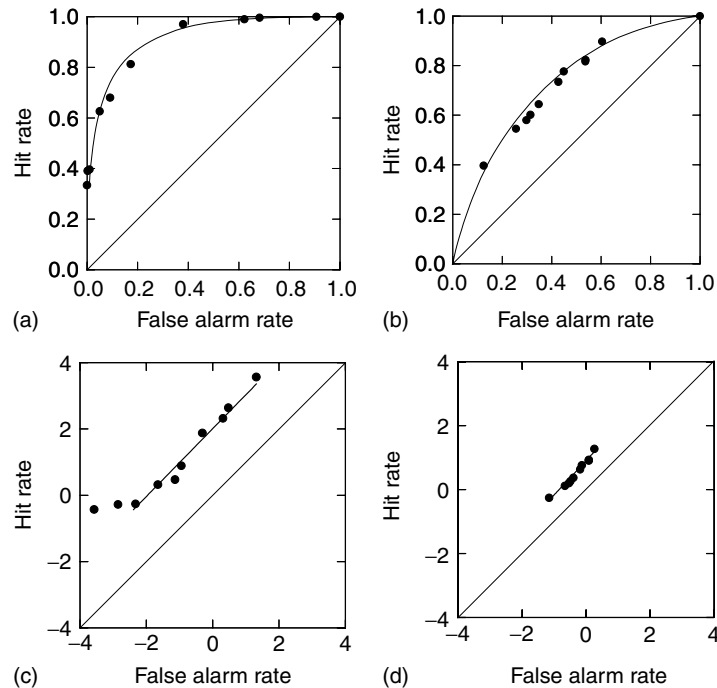


Figure 1 (a) and (b) plot the empirical hit rate and false alarm rates with fitted binormal model. Figures (c) and (d) show these graphs after they have been normalized so that the fitted line is straight

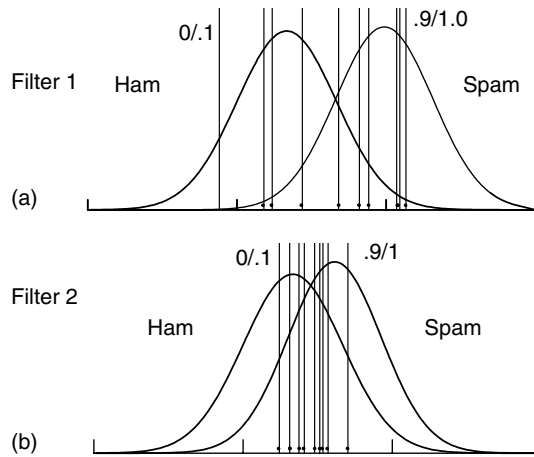


Figure 2 The models for filters (a) and (b) based on the fitted ROCs. The vertical lines show the 10 different response criteria

1(b). The diagonal lines stand for chance responding, and the further away from the diagonal the more diagnostic the variable spamicity is. Thus, comparing these two ROCs shows that the first filter is better.

The data are more easily seen if they are transformed so that the predicted lines are straight. These are sometimes called *normalized*, *standardized*, or *z-ROCs*. They can be calculated with the inverse normal function in most statistics packages and are available in most SDT software. These are shown in Figures 1(c) and 1(d). The straight lines in these graphs provide two main pieces of information. First, if the slopes are near 1, which they are for both these graphs, then they suggest the distributions for ham and spam have the same variance. The distance from the line to the diagonal is a measure of diagnosticity. If the slope of the line is 1, then this distance is the same at all points on the line and it corresponds to the SDT statistic d' . If the slope is not 1, then the distance between the diagonal and the line depends on the particular decision criterion. Several statistics are available for this (see [5] and [6]).

The fitted models in Figures 1(a–d) can be shown as normal distributions. Figures 2(a) and 2(b) indicate how well the ham and the spam can be separated by the filters. For the first filter, the spam distribution is about two standard deviations away from the ham distribution, while it is only about one standard deviation adrift for the second filter. The decision criteria are also included on these graphs.

For example, the far left criterion is at the cut-off between 0.0 and 0.1 on the spamicity variable. As can be seen, for the second filter the criteria are all close together, which means the users would have less choice about where along the dimension they could choose a criterion. These graphs are useful ways of communicating the findings.

An obvious question is whether the normal distribution assumption is valid. There are statistical tests that look at this, but it can also be examined graphically. Note, however, that as these graphs are cumulative, the data are not independent. Consequently, conducting standard regressions for the data in Figures 1(c) and 1(d) is not valid, and in these cases, the analysis should be done on the noncumulative data or using specialized packages like those cited earlier.

More advanced and less restrictive techniques are discussed in [6], but are relatively rare. The more common procedure is simply to draw straight lines between each of the observed points. This is called the *trapezoid method* because the area below the line between each pair of points is a trapezoid. Summing these trapezoids gives a measure called A . This tends to underestimate the area under real ROC curves. The values of A can range between 0 and 1 with chance discrimination as 0.5. The first filter has $A = 0.92$ and the second has $A = 0.73$.

An important characteristic of SDT is how the relative values of sensitivity and specificity are calculated and used to determine a criterion. In the

4 Receiver Operating Characteristics Curves

spam example, having low specificity would mean many real emails (ham) are labelled as spam and deleted. Arguably, this is more problematic than having to delete a few unsolicited spam. Therefore, if my filter automatically deletes messages, I would want the criterion to be high because specificity is more important than sensitivity.

To decide the relative value of deleting ham and not deleting spam, an equation from expected utility theory needs to be applied (slightly adapting the notation from [3], p.127):

$$S_{\text{opt}} = \frac{P(\text{ham})}{P(\text{spam})} \times \frac{V_{\text{CR}} - V_{\text{FA}}}{V_{\text{Hit}} - V_{\text{miss}}} \quad (1)$$

where S_{opt} is the optimal slope, $P(\text{ham})$ is the probability that an item will be ham, $P(\text{spam})$ is $1 - P(\text{ham})$, V_{CR} is the value of correctly not identifying ham as spam (this will be positive), V_{FA} is the value of incorrectly identifying ham as spam (negative), V_{Hit} is the value of correctly identifying spam (positive), and V_{miss} is the value of not identifying spam (negative). (It is worth noting here that a separate study is needed to estimate these utilities unless the minimum sensitivity or specificity is set externally; in such cases, simply go to this value on the ROC.) Thus, the odds value (see **Odds and Odds Ratios**) of ham is directly proportional to the slope. It is important to realize how important this baseline is for deciding the decision criterion. Often, people do not consider the baseline information when making decisions (see [4]).

Once S_{opt} is found, if one of the standard fitted ROC curves is used, then the optimal decision point is where the curve has this slope. For more complex

fitted curves and empirical curves, start in the upper left-hand corner of the ROC with a line of slope S_{opt} and move towards the opposite corner. The point where the line first intersects the ROC shows where the optimal decision criterion should be. Because there are usually only a limited number of possible decision criteria, the precision of this method is usually adequate to identify the optimal criterion.

This discussion only touches the surface of an exciting area of contemporary statistics. This general procedure has been expanded to many different experimental designs (see [2] and [5]), and has been generalized for **meta-analyses**, correlated and biased data, robust methods, and so on [6].

References

- [1] DeCarlo, L.T. (1998). Signal detection theory and generalized linear models, *Psychological Methods* **3**, 186–205.
- [2] Harvey Jr, L.O. (2003). *Parameter Estimation of Signal Detection Models: RscorePlus User's Manual*. Version 5.4.0, (<http://psych.colorado.edu/~lharvey/>, as at 17.08.04).
- [3] Swets, J.A. (1996). *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*, Lawrence Erlbaum.
- [4] Tversky, A. & Kahneman, D. (1980). Causal schemes in judgements under uncertainty, in *Progress in Social Psychology*, M. Fishbein, ed., Lawrence Erlbaum, Hillsdale.
- [5] Wickens, T.D. (2002). *Elementary Signal Detection Theory*, Oxford University Press, New York.
- [6] Zhou, X.-H., Obuchowski, N.A. & McClish, D.K. (2002). *Statistical Methods in Diagnostic Medicine*, John Wiley & Sons, New York.

DANIEL B. WRIGHT