Receiver Operating Characteristic (ROC) Curve Preparation - A Tutorial

Massachusetts Water Resources Authority

Environmental Quality Department
Report ENQUAD 2005-20

Citation

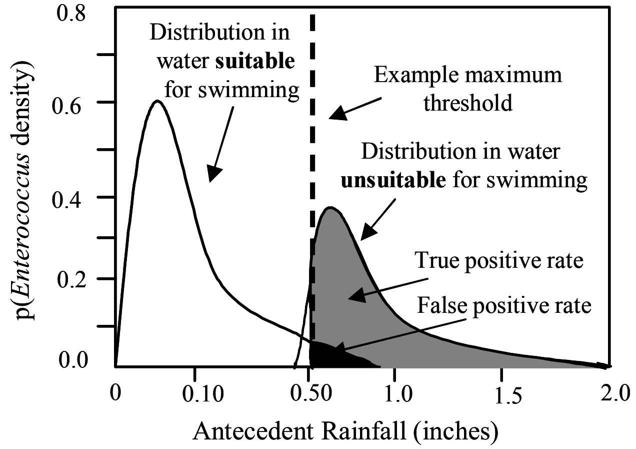
Morrison, Ann Michelle. 2005**. Receiver Operating Characteristic (ROC) Curve Preparation - A Tutorial.** Boston: Massachusetts Water Resources Authority. Report ENQUAD 2005-20. 5 p.

Receiver Operating Characteristic (ROC) Curve Preparation – A tutorial

Receiver Operating Characteristic (ROC) curves were developed in the field of statistical decision theory, and later used in the field of signal detection for analyzing radar images during World War II (1). ROC curves enabled radar operators to distinguish between an enemy target, a friendly ship, or noise. ROC curves assess the value of diagnostic tests by providing a standard measure of the ability of a test to correctly classify subjects. The biomedical field uses ROC curves extensively to assess the efficacy of diagnostic tests in discriminating between healthy and diseased individuals (2). ROC curves can (1) assess the overall discriminatory ability of different potential indicator variables by generating a common metric for comparison and (2) aid in the selection of a specific value of an indicator variable to use as a threshold, or limit, that provides a desired trade-off in the true positive rate and false positive rate. With respect to beach water quality indicator variables, ROC curves can quantify the overall effectiveness of different indicator variables to correctly and incorrectly classify a beach as suitable for swimming and generate a single metric by which the different indicator variables can be compared.

Statistical theory states that there are two populations of water: that which is suitable for swimming and that which is unsuitable for swimming. The distributions of these populations overlap. When you choose an indicator, such as *Enterococcus* density to define the transition between water that is suitable for swimming and water that is unsuitable for swimming, two regions of interest to regulators are formed (Figure 1). The true positive rate (TPR) corresponds to the proportion of water samples unsuitable for swimming that are correctly identified by the indicator variable threshold, which in Figure 1 is antecedent rainfall. The false positive rate (FPR) is the proportion of water samples that are incorrectly identified as unsuitable for swimming by the indicator variable. Recreational water management seeks to protect public health by identifying an indicator variable that maximizes the TPR and minimizes the FPR.

Figure 1. Hypothetical distributions of water suitable and unsuitable for swimming.



The following text describes in detail how to construct an ROC curve using Microsoft Excel (Microsoft Corporation) software.

Data preparation:

Beach data is typically collected from replicate sites along a beach. To minimize pseudoreplication, it is necessary to generate a single value that represents level of the bacterial indicator (*i*.e. *Enterococcus)* for the beach on any given day. This can be done most simply by calculating the arithmetic mean or the geometric mean for the replicate samples. Preliminary work suggested that ROC curves of bacterial data were stronger using a geometric mean. If you chose to calculate a geometric mean, convert it to a real value with the "EXP" function in Microsoft Excel.

1. Load data into a Microsoft Excel spreadsheet. Include a column for beach, date, year, mean *Enterococcus* value, and your indicator variable(s) such as rain.

A	B	C	D	E	F	G	H	I	J	K	L
1	**Beach**	**Date**	**Year**	**gmEC**	**Rain48**						
2	CARS	7/2/04	2004	1	0						
3	CONS	9/1/04	2004	3.1	0.52						
4	TEN	8/2/04	2004	534.2	0.15						
5	WOLL	6/2/04	2004	127.4	0.08						

2. Select all of the data and on the Microsoft Excel tool bar under "Data" choose "Sort."

3. Select the indicator variable (*i.e.* rain) and sort "ascending."

A	B	C	D	E	F	G	H	I	J	K	L
1	**Beach**	**Date**	**Year**	**gmEC**	**Rain48**						
2	CARS	7/2/04	2004	1	0						
3	WOLL	6/2/04	2004	127.4	0.08						
4	TEN	8/2/04	2004	534.2	0.15						
5	CONS	9/1/04	2004	3.1	0.52						

ROC preparation:

1. You now need to create two new columns of data: "Exceed" and "No Exceed." To do this, use an "if, then" function. In the following example, column E contains the geometric mean *Enterococcus* data (gmEC). You want to assign a value of 1 to the "Exceed" cell if the *Enterococcus* value in cell E2 is greater than 104 CFU (or whatever regulatory limit is used). To calculate the "No Exceed" value, use an "If, then" statement that says if the "Exceed" cell is 1, then the "No Exceed" cell is 0.

Exceed=IF(E2>104,1,0)   No Exceed =IF(G=1,0,1)

A	B	C	D	E	F	G	H	I	J	K	L
1	**Beach**	**Date**	**Year**	**gmEC**	**Rain48**	**EX**	**NOEX**				
2	CARS	7/2/04	2004	1	0	0	1				
3	WOLL	6/2/04	2004	127.4	0.08	1	0				
4	TEN	8/2/04	2004	534.2	0.15	1	0				
5	CONS	9/1/04	2004	3.1	0.52	0	1				

2. The next columns that need to be created are "True Positive Rate" (TPR) and "False Positive Rate" (FPR). With respect to bacterial beach indicator variables, the "True Positive Rate" is the proportion of times that the *Enterococcus* density exceeds the single sample maximum (*i.e.* 104 CFU) if you regulated at a particular threshold value of the indicator variable (*i.e.* rain). The "False Positive Rate" corresponds to the proportion of times that the *Enterococcus* density did not exceed the single sample maximum (*i.e.* 104 CFU) if you regulated at a particular threshold value of the indicator variable (*i.e.* rain). In the spreadsheet the equation to create the TPR and FPR for row 2 are as follows:

TPR =sum(G3:$G$9)/sum($G$2:$G$9)

FPR =sum(H3:$H$9)/sum($H$2:$H$9)

The easiest way to think about this is to think, "If I used the rain value in row 2 as my limit for closing the beach, what proportion of *Enterococcus* exceedances occur above this rainfall volume." By summing the number of exceedances below row 2 (corresponding to larger rainfall volumes) and dividing by the total number of exceedances, you are able to calculate the TPR. Likewise for the FPR, if you think, "If I used the rain value in row 2 as my limit for closing the beach, what proportion of *Enterococcus* densities below 104 CFU (or whatever regulatory limit you have chosen) occur above this rainfall volume. By summing the number of "No exceedances" above the rainfall volume in row 2 and dividing by the total number of "No exceedances", you are able to calculate the FPR.

Here is an example:

A	B	C	D	E	F	G	H	I	J	K	L
1	**Beach**	**Date**	**Year**	**gmEC**	**Rain48**	**EX**	**NOEX**	**TPR**	**FPR**		
2	CARS	7/2/04	2004	1	0	0	1	1	0.75		
3	WOLL	6/2/04	2004	127.4	0.08	1	0	0.75	0.75		
4	TEN	8/2/04	2004	534.2	0.15	1	0	0.5	0.75		
5	CONS	9/1/04	2004	7.5	0.15	0	1	0.5	0.5		
6	CARS	7/2/04	2004	1	0.15	0	1	0.5	0.25		
7	WOLL	6/2/04	2004	210.5	0.37	1	0	0.25	0.25		
8	TEN	8/2/04	2004	425.2	0.40	1	0	0	0.25		
9	CONS	9/1/04	2004	3.1	0.52	0	1	0	0		

Examining the data in row 3, TPR is interpreted as follows:
If a beach is posted as unsuitable for swimming at 48 hour rain values greater than 0.08, the rate that *Enterococcus* exceedences of 104 CFU will be correctly flagged is 0.75. However, because 0.08 is not a lot of rain, the rate that the beach will be posted as unsuitable for swimming when it is in fact safe (<104 CFU) is also 0.75. The objective with beach management is to balance the TPR and the FPR. Ideally a beach manager would like to have an indicator variable threshold associated with a high TPR (close to 1) and a low FPR (close to 0). This means that beach is closed correctly most of the time and open correctly most of the time as well.

3. To construct the ROC curve, you need a TPR, FPR pair for each unique value of the indicator variable (*i.e.* rain). In our previous example there are 3 TPR, FPR pairs corresponding to 0.15” of rainfall. To avoid this problem, create two new columns, TPRu and FPRu, for the TPR and FPR values associated with each unique rain value. The equation used to do this involves an “If, then” statement.

TPRu = IF(F3>F2,I3, “”)   FPRu =IF(F3>F2,J3,””)

These statements read, “If the rain fall value in the cell is greater than the rainfall value in the cell above, then use the TPR or FPR rate in this row, otherwise leave blank.”

A	B	C	D	E	F	G	H	I	J	K	L
1	**Beach**	**Date**	**Year**	**gmEC**	**Rain48**	**EX**	**NOEX**	**TPR**	**FPR**	**TPRu**	**FPRu**
2	CARS	7/2/04	2004	1	0	0	1	1	0.75	1	0.75
3	WOLL	6/2/04	2004	127.4	0.08	1	0	0.75	0.75	0.75	0.75
4	TEN	8/2/04	2004	534.2	0.15	1	0	0.5	0.75	0.5	0.75
5	CONS	9/1/04	2004	7.5	0.15	0	1	0.5	0.5		
6	CARS	7/2/04	2004	1	0.15	0	1	0.5	0.25		
7	WOLL	6/2/04	2004	210.5	0.37	1	0	0.25	0.25	0.25	0.25
8	TEN	8/2/04	2004	425.2	0.40	1	0	0	0.25	0	0.25
9	CONS	9/1/04	2004	3.1	0.52	0	1	0	0	0	0

4. Select columns K and L and copy the values only (Paste Special) into a new worksheet within the Microsoft Excel workbook.

5. Within the new worksheet, add a new data point, 1, to the top of both the TPR and FPR columns. This is done to anchor the shape of the curve.

6. Now, remove the empty spaces. Select both columns of data, and under the Microsoft Tool Bar, select "Data." Select "Filter" and "Autofilter." Then choose "Blanks." Delete the blank rows and return to the autofilter drop down menu and choose "All" in both columns. Turn Autofilter off.

7. Select the chart function and plot the data using an x-y scatter. TPR is on the y axis and FPR is on the x axis.

8. To calculate the area under the curve, use the trapezoid rule for each row. Create a new column called area. The formula for these cells is as follows:

   area = ((A1+A2)/2)*(B1-B2)
   The "A" column is the TPR and the "B" column is the FPR.

9. After the area has been calculated for each row, sum all of the areas to determine the area under the curve (AUC), which is the common metric by which you can compare different indicator variables. An AUC close to 1 indicates a strong indicator variable, and AUC close to 0.5 indicates that the variable has little discriminatory power.

10. To determine an appropriate threshold value for the indicator variable (rain) consider the TPR and FPR you desire. With these points in mind, search the data to find the value of rain associated with that TPR, FPR pair.

References:

1. **Collison, P.** 1998. Of bombers, radiologists, and cardiologists: time to ROC. Heart **80:**215-217.
2. **Metz, C.** 1978. Basic principles of ROC analysis. Seminars in Nuclear Medicine **8:**283-298.

Massachusetts Water Resources Authority
Charlestown Navy Yard
100 First Avenue
Boston, MA 02129
(617) 242-6000
http://www.mwra.state.ma.us