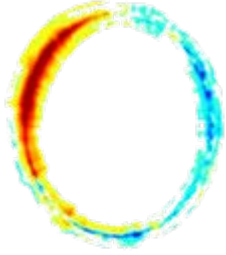
Confidence Intervals and Hypothesis Tests

Simon J. Mason

simon@iri.columbia.edu



International Research Institute for Climate and Society
The Earth Institute of Columbia University

Third International Verification Methods Workshop

Reading, England, 29 January – 2 February, 2007

Introduction

There are three fundamental questions to ask when interpreting a forecast verification score:

Is the score good?

Hypothesis testing – could we have achieved the same or a better score by chance?

Is the score correct?

Confidence intervals – is the sample score an accurate indication of the “true” score?

What does the score mean anyway?

Outline

- What does the area under the ROC graph mean?
- Hypothesis testing and the ROC
- Confidence intervals and the ROC

Two-Alternative Forced Choice Test

If we give deterministic forecasts for a multi-category outcome accuracy scores can be misleading.

For example, if we forecast for 3 equiprobable categories, the probability of a correct forecast is 33%, which would suggest that if our forecasts are correct 45% of the time then they are skilful.

But to many users anything less than 50% sounds hopeless!

Two-Alternative Forced Choice Test

It would be useful to have a score for which 50% represents no skill regardless of the number of categories.

An option is a two-alternative forced choice (2AFC) test.

A 2AFC test is a test to correctly identify which of two options has a characteristic of interest. (Normally, one, and only one, of these choices would have the characteristic.)

Two-Alternative Forced Choice Test

The commonest 2AFC tests are comparisons:

- does *A* score higher (or lower) on some characteristic than *B*?

But 2AFC tests can also involve identifying the presence of a characteristic that is present in only one of the two options …

Two-Alternative Forced Choice Test

Which of these two years was a “wet” year In Lusaka, Zambia (DJF rainfall was more than the climatological upper quartile)?

Year
1993/94
1998/99

What is the probability of getting the answer correct?

50% (assuming that you do not have inside information about Lusaka’s rainfall history or knowledge of ENSO teleconnections).

Two-Alternative Forced Choice Test

Which of these two years was a “wet” year in Lusaka, Zambia (DJF rainfall was more than the climatological upper quartile)?

Year	Forecast
1993/94	547
1998/99	728

What is the probability of getting the answer correct now?

That depends on whether we can believe the forecasts.

Two-Alternative Forced Choice Test

A simple score can be defined to indicate how good we are at 2AFC tests: calculate the proportion of tests for which we gave the correct answer.

One advantage of this test is that the interpretation of the test's score is intuitive:

A score of 100% indicates a perfect set of answers.
A score of 0% indicates a perfectly bad set of answers.
A score of 50% would be expected by guessing the answer every time.
A score of more than 50% suggests we have some skill in answering the questions.

Interpreting the ROC area

Year	Forecast
1984/85	661
1985/86	658
1986/87	573
1987/88	512
1988/89	707
1989/90	692
1990/91	621
1991/92	532
1992/93	584
1993/94	547
1994/95	496
1995/96	713
1996/97	623
1997/98	386
1998/99	728
1999/00	712
2000/01	682
2001/02	671
2002/03	571
2003/04	597

Retroactive forecasts of DJF seasonal rainfall totals for Lusaka.

In which years would we expect “wet” conditions (wettest 25%) to occur?

Interpreting the ROC area

Forecast	Year
728	1998/99
713	1995/96
712	1999/00
707	1988/89
692	1989/90
682	2000/01
671	2001/02
661	1984/85
658	1985/86
623	1996/97
621	1990/91
597	2003/04
584	1992/93
573	1986/87
571	2002/03
547	1993/94
532	1991/92
512	1987/88
496	1994/95
386	1997/98

The most sensible strategy would be to list the years in order of decreasing forecast rainfall.

If the forecasts are good, the "wet" years should be at the top of the list.

Note that the precise values of the forecasts are irrelevant – it is only the ordering of the forecasts that is important.

Interpreting the ROC area

Forecast	Year	Observed
728	1998/99	880
713	1995/96	401
712	1999/00	681
707	1988/89	929
692	1989/90	685
682	2000/01	813
671	2001/02	323
661	1984/85	588
658	1985/86	754
623	1996/97	615
621	1990/91	652
597	2003/04	749
584	1992/93	538
573	1986/87	538
571	2002/03	559
547	1993/94	269
532	1991/92	274
512	1987/88	537
496	1994/95	297
386	1997/98	563

For the first guess:

$$\text{Hit rate} = \frac{\text{number of hits}}{\text{number of events}}$$

$$= \frac{1}{5}$$

$$\text{FAR} = \frac{\text{number of false alarms}}{\text{number of non-events}}$$

$$= \frac{0}{15}$$

Repeat for all forecasts.

Interpreting the ROC area

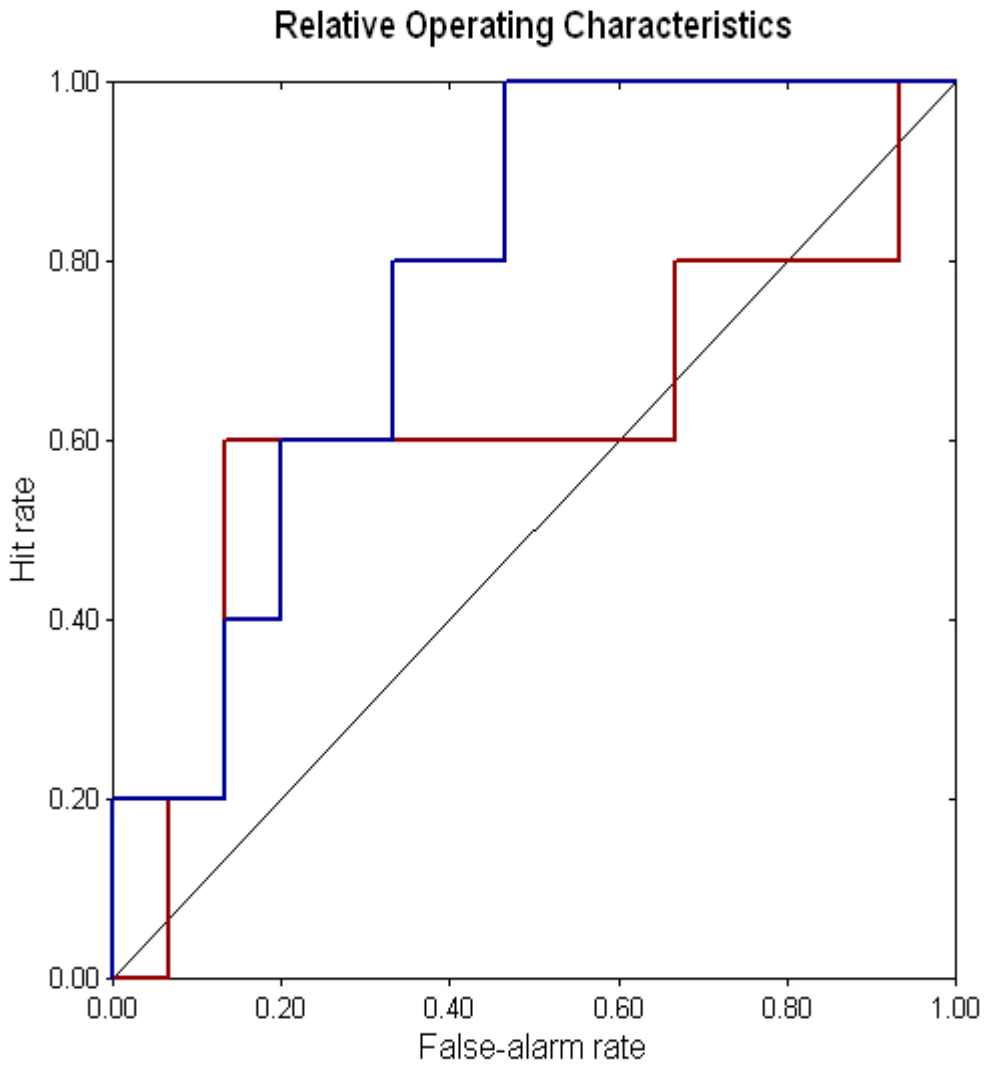
Forecast	Year	Correct	Incorrect
728	1998/99	1 of 5	0 of 15
713	1995/96	1 of 5	1 of 15
712	1999/00	1 of 5	2 of 15
707	1988/89	2 of 5	2 of 15
692	1989/90	2 of 5	3 of 15
682	2000/01	3 of 5	3 of 15
671	2001/02	3 of 5	4 of 15
661	1984/85	3 of 5	5 of 15
658	1985/86	4 of 5	5 of 15
623	1996/97	4 of 5	6 of 15
621	1990/91	4 of 5	7 of 15
597	2003/04	5 of 5	7 of 15
584	1992/93	5 of 5	8 of 15
573	1986/87	5 of 5	9 of 15
571	2002/03	5 of 5	10 of 15
547	1993/94	5 of 5	11 of 15
532	1991/92	5 of 5	12 of 15
512	1987/88	5 of 5	13 of 15
496	1994/95	5 of 5	14 of 15
386	1997/98	5 of 5	15 of 15

Interpreting the ROC area

We could do the same thing for the “dry” years, but the years are sorted in ascending order:

Forecast	Year	Correct	Incorrect
386	1997/98	0 of 5	1 of 15
496	1994/95	1 of 5	1 of 15
512	1987/88	1 of 5	2 of 15
532	1991/92	2 of 5	2 of 15
547	1993/94	3 of 5	2 of 15
571	2002/03	3 of 5	3 of 15
573	1986/87	3 of 5	4 of 15
584	1992/93	3 of 5	5 of 15
597	2003/04	3 of 5	6 of 15
621	1990/91	3 of 5	7 of 15
623	1996/97	3 of 5	8 of 15
658	1985/86	3 of 5	9 of 15
661	1984/85	3 of 5	10 of 15
671	2001/02	4 of 5	10 of 15
682	2000/01	4 of 5	11 of 15
692	1989/90	4 of 5	12 of 15
707	1988/89	4 of 5	13 of 15
712	1999/00	4 of 5	14 of 15
713	1995/96	5 of 5	14 of 15
728	1998/99	5 of 5	15 of 15

Interpreting the ROC area



From the trapezium rule, the areas beneath the curves are 0.77 for wet (blue) and 0.61 for dry (red).

But we do not have to construct the graphs to calculate these areas …

Interpreting the ROC area

Forecast	Year
728	1998/99
713	1995/96
712	1999/00
707	1988/89
692	1989/90
682	2000/01
671	2001/02
661	1984/85
658	1985/86
623	1996/97
621	1990/91
597	2003/04
584	1992/93
573	1986/87
571	2002/03
547	1993/94
532	1991/92
512	1987/88
496	1994/95
386	1997/98

A simple way to calculate the ROC area is to count the proportion of times the forecasts for when it is “wet” exceed the forecasts for when it is not “wet” …

(If there is a tie, which may often be the case if probabilities are used to rank the forecasts rather than actual forecast values, count as a half.)

Interpreting the ROC area

Forecast	Year
728	1998/99
713	1995/96
712	1999/00
707	1988/89
692	1989/90
682	2000/01
671	2001/02
661	1984/85
658	1985/86
623	1996/97
621	1990/91
597	2003/04
584	1992/93
573	1986/87
571	2002/03
547	1993/94
532	1991/92
512	1987/88
496	1994/95
386	1997/98

The forecast for 1998/98 exceeds all 15 of the forecasts when not-wet.

The forecast for 1988/89 exceeds 13 of the forecasts when not-wet.

The forecast for 2000/01 exceeds 12 of the forecasts when not-wet.

The forecast for 1985/86 exceeds 10 of the forecasts when not-wet.

The forecast for 2003/04 exceeds 8 of the forecasts when not-wet.

$$\text{ROC area} = \frac{(15+13+12+10+8)}{(15\times 5)} = 0.77$$

Interpreting the ROC area

So the ROC can be calculated in terms of 2AFC tests:

$$\text{ROC area} = \frac{\sum_{i=1}^{e}\sum_{j=1}^{e'} I\left(f_i, g_j\right)}{e'e}$$

where e = number of events

e' = number of non-events

f_i = forecast for i^{th} event

g_j = forecast for j^{th} non-event

$$I\left(f_i, g_j\right) = \begin{cases} 1 & \text{if } f_i > g_j \\ 0 & \text{if } f_i < g_j \\ \frac{1}{2} & \text{if } f_i = g_j \end{cases}$$

Outline

- What does the area under the ROC graph mean?
- Hypothesis testing and the ROC
- Confidence intervals and the ROC

Hypothesis testing and the ROC

A commonly used method to assess whether a verification score is “good” is to calculate the probability that a scores at least as good as that observed could have been achieved given completely useless forecasts (i.e., forecasts that have no discriminatory power).

This probability is called a *p*-value. There are a number of ways of calculating the *p*-value:

Hypothesis testing and the ROC

What is the probability of achieving an ROC area at least as large as observed given forecasts with no discriminatory power?

1. Calculate the ROC areas for all possible rankings of the forecasts identify the proportion of times the ROC area is at least as large as observed.

2. For some statistics the distribution of scores follows a theoretical distribution. For the ROC, this proportion is related to the U-distribution.

$$p\text{-value} = 0.040$$

Hypothesis testing and the ROC

What is the probability of achieving an ROC area at least as large as observed given forecasts with no discriminatory power?

3. Use an approximate distribution.

   The ROC area approximates a gaussian distribution for "large" samples.

$$U \sim N\left(0.5, \frac{(e + e' + 1)}{12e'e}\right)$$

$$U \sim N\left(0.5, \frac{(5 + 15 + 1)}{12 \times 5 \times 15}\right)$$

$$\sim N(0.5, 0.023)$$

$$z = \frac{0.77 - 0.5}{\sqrt{0.023}} \approx 1.79$$

$$p\text{-value} = 0.037$$

Hypothesis testing and the ROC

What is the probability of achieving an ROC area at least as large as observed given forecasts with no discriminatory power?

3. Approximate the distribution by generating a large number of random rankings.

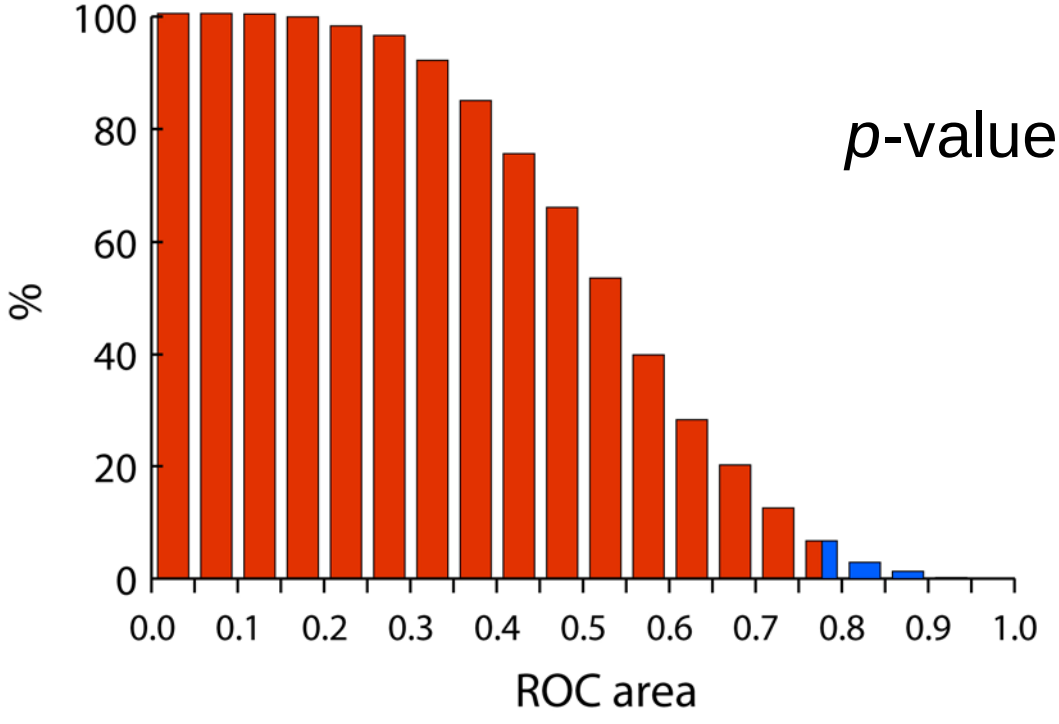
   Perhaps the easiest approach is to randomly redefine when events occurred.

   The exact distribution is obtained from the results for all $\binom{e' + e}{e}$ possible redefinitions of events.

Hypothesis testing and the ROC

What is the probability of achieving an ROC area at least as large as observed assuming that the forecasts have no discriminatory power?

3. Approximate the distribution by generating a large number of random rankings.



$$p\text{-value} = \frac{56}{1000} = 0.056$$

Hypothesis testing and the ROC

3. Approximate the distribution by generating a large number of random rankings.

   A *permutation* procedure is used to obtain the random rankings

Year	Obs.	Year	For.
2001	-0.23	2001	0.28
2002	1.59	2002	0.77
2003	0.41	2003	0.44
2004	0.92	2004	0.59
2005	-0.55	2005	0.37

Permutation 1

Year	Obs.	Year	For.
2001	-0.23	2005	0.37
2002	1.59	2004	0.59
2003	0.41	2001	0.28
2004	0.92	2003	0.44
2005	-0.55	2002	0.77

Permutation 2

Year	Obs.	Year	For.
2001	-0.23	2004	0.59
2002	1.59	2001	0.28
2003	0.41	2003	0.44
2004	0.92	2005	0.37
2005	-0.55	2002	0.77

Permutation 3

Year	Obs.	Year	For.
2001	-0.23	2003	0.44
2002	1.59	2001	0.28
2003	0.41	2005	0.37
2004	0.92	2002	0.77
2005	-0.55	2004	0.59

Outline

- What does the area under the ROC graph mean?
- Hypothesis testing and the ROC
- Confidence intervals and the ROC

Confidence intervals

If we had a different set of forecasts the calculated score will vary from the sample score even if the skill of the forecasts is unchanged. The calculated score is therefore only an estimate of the ‘real’ score. It would be helpful to know how sensitive the score is to the sample; if the score is sensitive the uncertainty in the estimate will be high.

A recommended way of indicating uncertainty is to calculate confidence intervals.

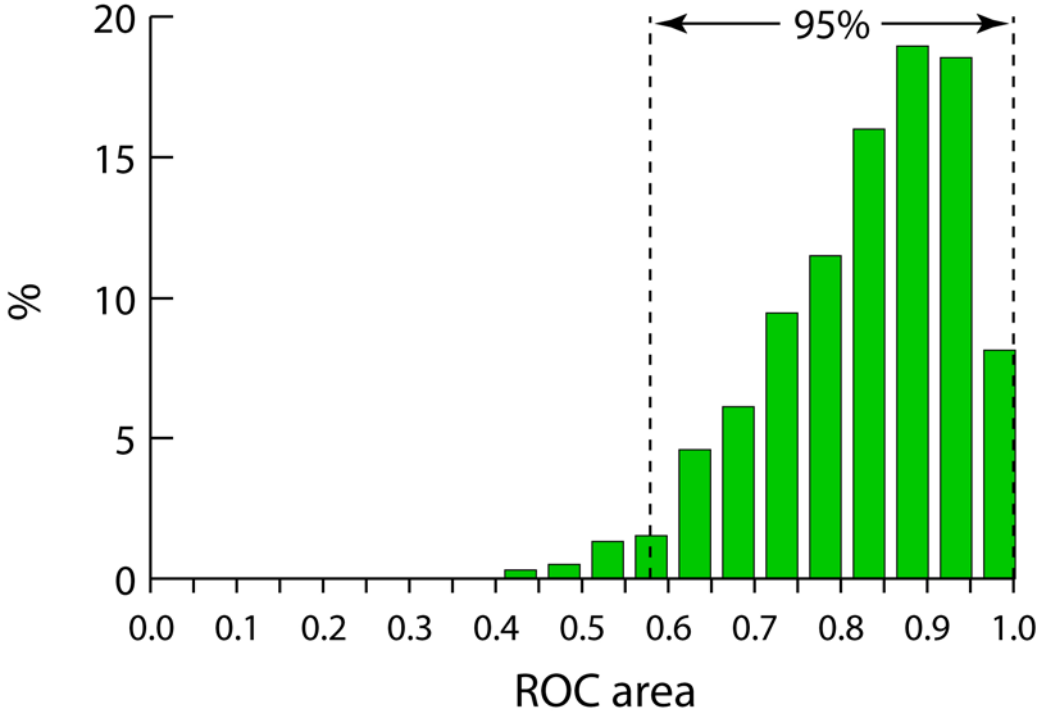
Confidence intervals

There are many ways of calculating confidence intervals. Some of the most commonly used procedures include:

1. Assume gaussian errors (e.g. mean)
2. Use exact distribution (e.g., hit rate)
3. Resampling

Confidence intervals and the ROC

The best way to obtain confidence intervals for the ROC area and individual points on the graph is by bootstrapping – randomly re-sampling the forecast and observed *pairs* with replacement.



Note:

1. the sample score can be biased;
2. the distribution of skill scores generally will be skewed.

Confidence intervals and the ROC

A *bootstrap* procedure is used to obtain the resamples (compare with the permutation procedure)

Year	Obs.	Year	For.
2001	-0.23	2001	0.28
2002	1.59	2002	0.77
2003	0.41	2003	0.44
2004	0.92	2004	0.59
2005	-0.55	2005	0.37

Bootstrap 1

Year	Obs.	Year	For.
2001	-0.23	2001	0.28
2001	-0.23	2001	0.28
2002	1.59	2002	0.77
2004	0.92	2004	0.59
2005	-0.55	2005	0.37

Permutation 2

Year	Obs.	Year	For.
2002	1.59	2001	0.28
2002	1.59	2001	0.28
2003	0.41	2003	0.44
2003	0.41	2003	0.44
2004	0.92	2005	0.37

Permutation 3

Year	Obs.	Year	For.
2002	1.59	2001	0.28
2002	1.59	2001	0.28
2002	1.59	2001	0.28
2004	0.92	2002	0.77
2005	-0.55	2004	0.59

Confidence intervals and the ROC

Why calculate confidence intervals?

1. Indicates sampling uncertainty in the score.
2. More informative than *p*-values.
3. Facilitates comparison of scores.

Exercises

- Construct an ROC graph using the following forecasts of the Nino3 index from the DEMETER project.
- Calculate the ROC area for forecasts of “El Nino” using the 2AFC method, and confirm that this area is exactly the same as using the trapezium rule.
- Calculate the probability of obtaining by chance an area at least as large using:
  - the exact tail area of the Mann-Whitney U-statistic
  - the normal approximation to the U-statistic
  - permutation tests

Exercises

- Calculate confidence intervals for the ROC area using:
  - the normal approximation to the Mann-Whitney U-statistic
  - bootstrapping
- Repeat for “La Nina”.
- Is there a significant difference in the skill of the “El Nino” and “La Nina”-onset forecasts?

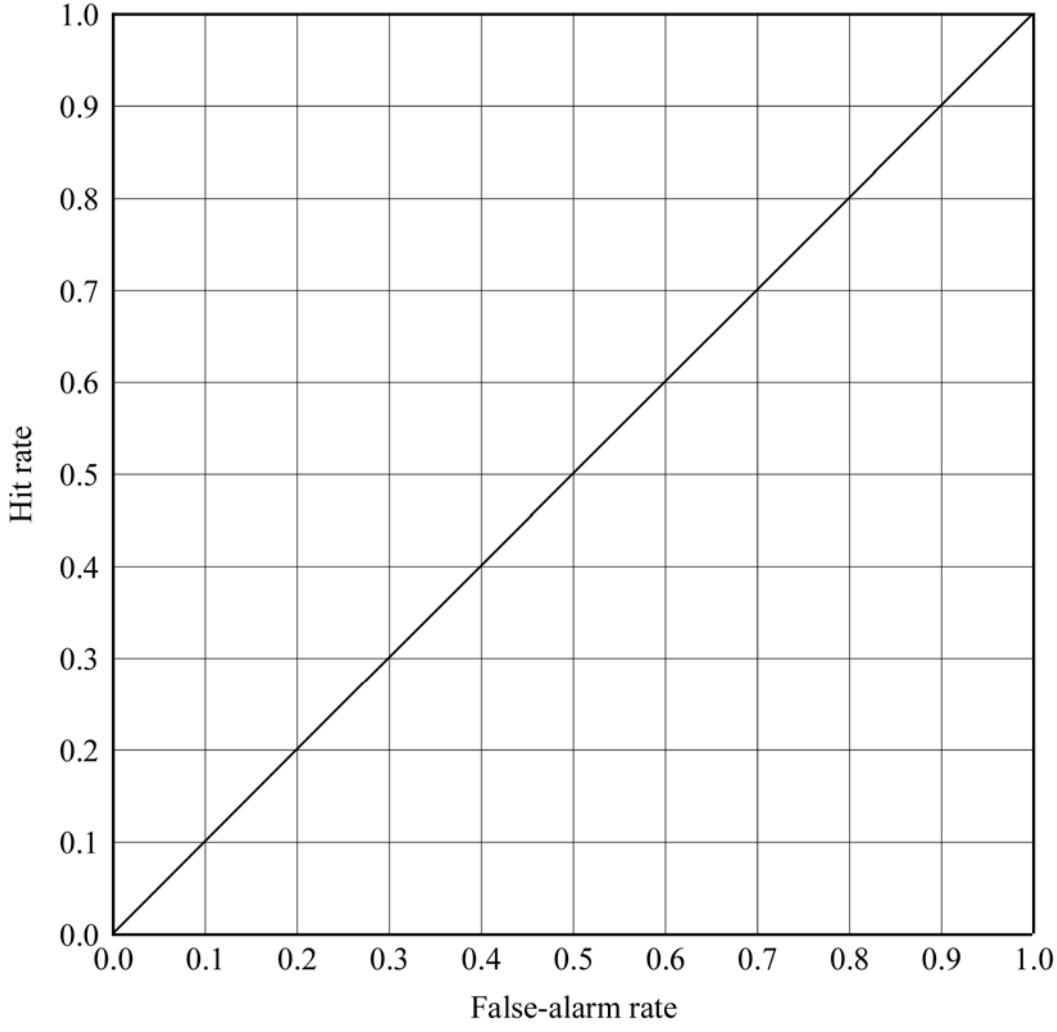
Data

Year	ENSO	Forecast
1981	0	0.18
1982	1	0.96
1983	1	-0.13
1984	-1	-0.66
1985	-1	-0.27
1986	0	0.41
1987	1	0.95
1988	-1	-0.92
1989	0	0.45
1990	0	0.67
1991	1	-0.11
1992	0	-0.07
1993	0	0.16
1994	0	-0.50
1995	0	-0.64
1996	0	0.12
1997	1	0.61
1998	0	-0.52
1999	-1	-0.68
2000	-1	0.00

ENSO phases are:

1 = El Nino
0 = neutral
-1 = La Nina

Multi-model (DEMETER) ensemble-mean predictions of July Nino3 anomalies from February initial conditions.


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Hit rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
False-alarm rate