Forecast Verification: Part I

Forecast Verification

- Forecast verification is the process of assessing the quality of forecasts.
  - Also called validation, or evaluation.
- We may need different verification metrics.
  - There can be differing views of what constitutes a good forecast.
  - There are many ways for forecasts to go right or go wrong.
- Why are objective evaluations of forecast quality important? Brier and Allen (1951) suggested that they serve three purposes.
  - Administrative use: It pertains to ongoing monitoring of operational forecasts. Forecast verification allows objective comparison of the relative merits of competing forecasting systems.
  - Scientific use: assess specific strengths and weaknesses of forecasting systems and provide information on forecast improvement
  - Economic use: provide users with an objective assessment on the usefulness of forecasts and support better decision making

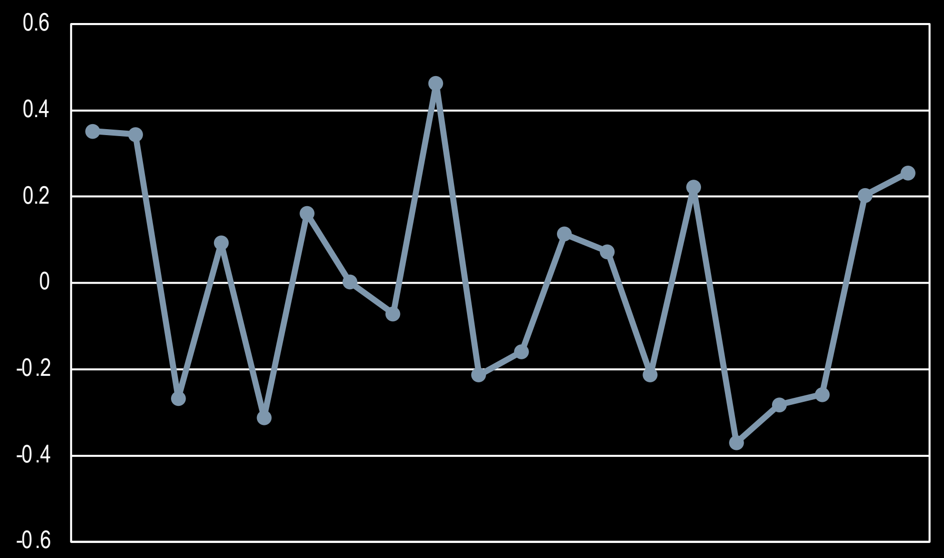
Wilks 2011

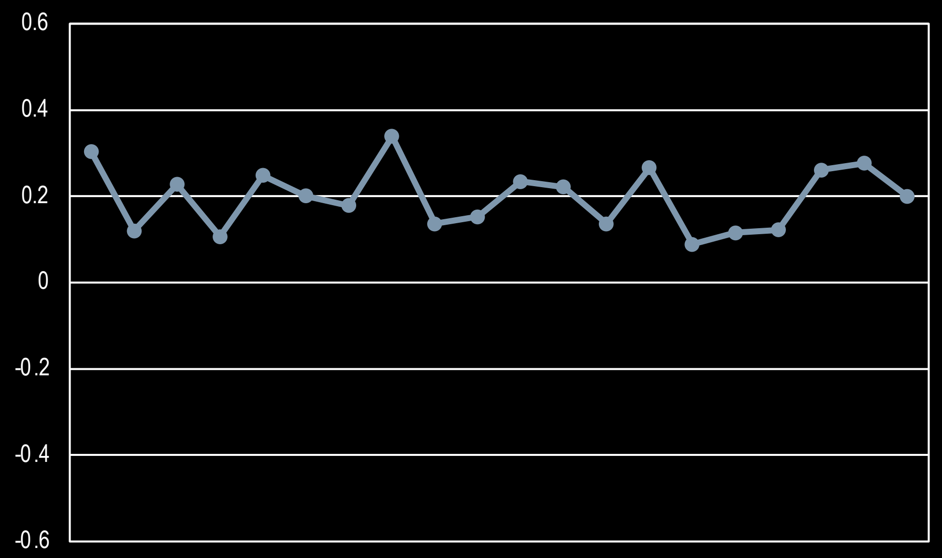
Scalar Attributes of Forecast Performance

- Accuracy refers to the average correspondence between individual forecasts and the events they predict.
- Bias, or unconditional bias, or systematic bias, measures the correspondence between the average forecast and the average observed value of the predictand.
  - Bias is different from accuracy, which measures the average correspondence between individual pairs of forecasts and observations.
  - Temperature forecasts that are consistently too warm or precipitation forecasts that are consistently too wet both have a positive bias, whether or not the forecasts are otherwise reasonably accurate or quite inaccurate.
- Reliability, or calibration, or conditional bias, characterize the conditional distributions of the observations (o_j) given the forecasts (y_i): $p(o_j|y_i)$

Wilks 2011

Time Series of Forecast - Observation differences





Scalar Attributes of Forecast Performance (cont'd)

- Resolution is concerned with the properties of the conditional distributions of the observations given the forecasts, $p(o_j|y_i)$, which is similar to reliability. However, resolution pertains to the differences between the conditional averages of the observations for different values of the forecast.
  - If average temperature outcomes following forecasts of 50F and 60F are very different, the forecasts can resolve these different temperature outcomes, and are said to exhibit resolution. If the temperature outcomes following forecasts of 50F and 60F are nearly the same on average, the forecasts exhibit almost no resolution.
- Discrimination is the converse of resolution, in that it pertains to differences between the conditional averages of the forecasts for different values of the observation. $p(y_i|o_j)$.
- Sharpness, or refinement, is an attribute of the forecasts alone. Forecasts that rarely deviate much from the climatological value of the predictand exhibit low sharpness. By contrast, forecasts that are frequently much different from the climatological value of the predictand are sharp.
  - In the extreme, forecasts consisting only of the climatological value of the predictand exhibit no sharpness.

Wilks 2011

Skill Score

- Forecast skill refers to the relative accuracy of a set of forecasts, with respect to some set of reference forecasts.
  - Common choices for the reference forecasts are climatological average values of the predictand, persistence forecasts , or random forecasts
  - When evaluating the performance of a new forecasting system, the old forecasting system can be used as the reference.
- The skill score for forecasts characterized by a basic accuracy measure A is given by

$$SS_{\mathrm{ref}} = \frac{A - A_{\mathrm{ref}}}{A_{\mathrm{perf}} - A_{\mathrm{ref}}} \times 100\%,$$

- SS = 100%: perfect forecasts
- SS = 0: no improvement over the reference forecasts
- SS < 0, inferior to the reference forecasts

Where Aref is the accuracy of reference forecasts, and Aperf is the value of the accuracy measure for perfect forecasts.

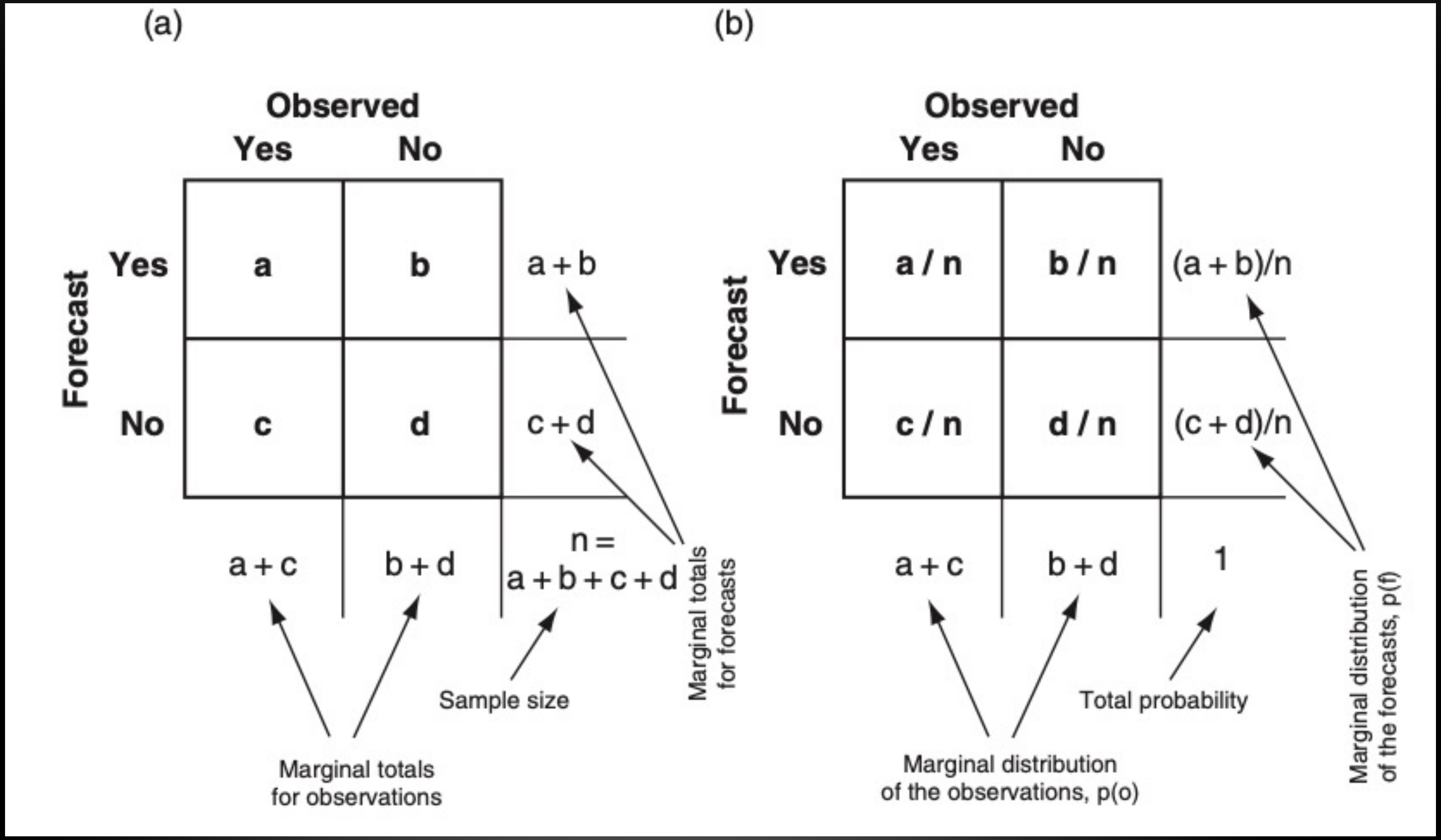
Wilks 2011

Some Commonly Used Verification Metrics

- Non-probabilistic forecasts of discrete predictands
  - The 2X2 contingency table
  - Skill scores for 2X2 contingency tables
- Non-probabilistic Forecasts of Continuous Predictands
  - Mean absolute error
  - Mean squared error
  - Correlation
- Probabilistic forecasts of Discrete Predictands: such as the occurrence of extreme precipitation
  - The Brier Score
  - Reliability Diagrams
- Non-probabilistic forecasts of fields
  - Anomaly correlation coefficients (ACC)
  - Root mean squared error (RMSE)

Wilks 2011

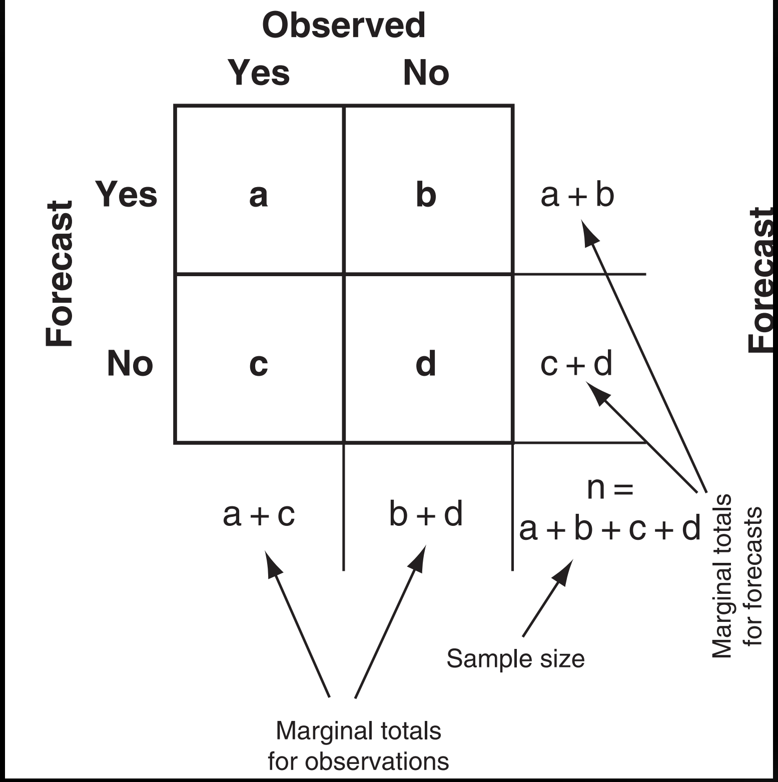
Non-probabilistic forecasts of Discrete Predictands: 2X2 Contingency Table



Wilks 2011

the average correspondence between individual pairs of forecasts and observations

Accuracy: Proportion Correct



$$PC = \frac{a+d}{n}.$$

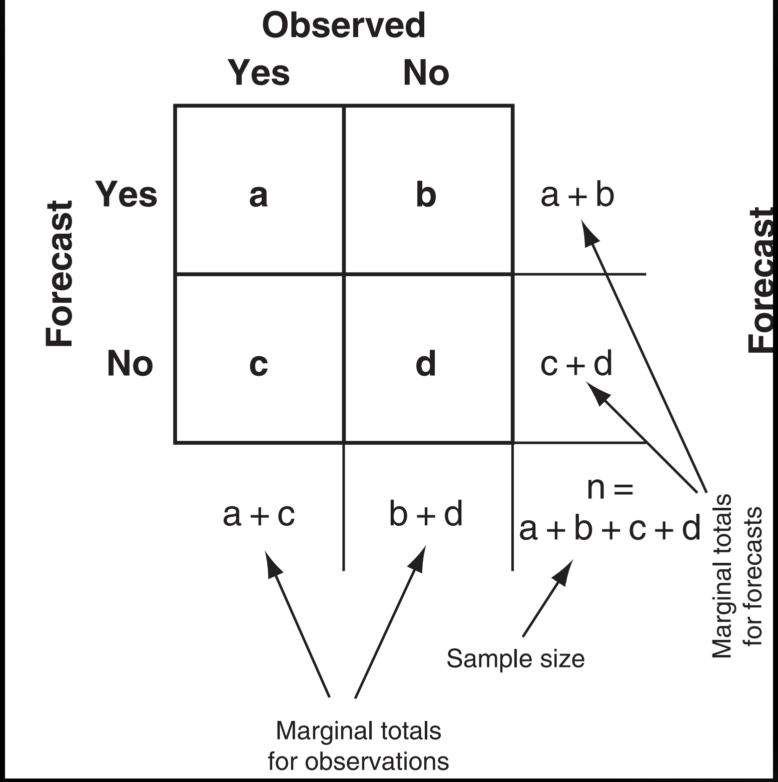
- the fraction of the n forecast occasions for which the forecast correctly anticipated the yes or no event.
- It credits correct yes and no forecasts equally. However, this is not always a desirable attribute, particularly when the yes event is rare.

What is the value of PC for perfect forecasts?

Wilks 2011

Accuracy: threat score (TS) or critical success index (CSI)

the average correspondence between individual pairs of forecasts and observations



$$TS = CSI = \frac{a}{a+b+c}.$$

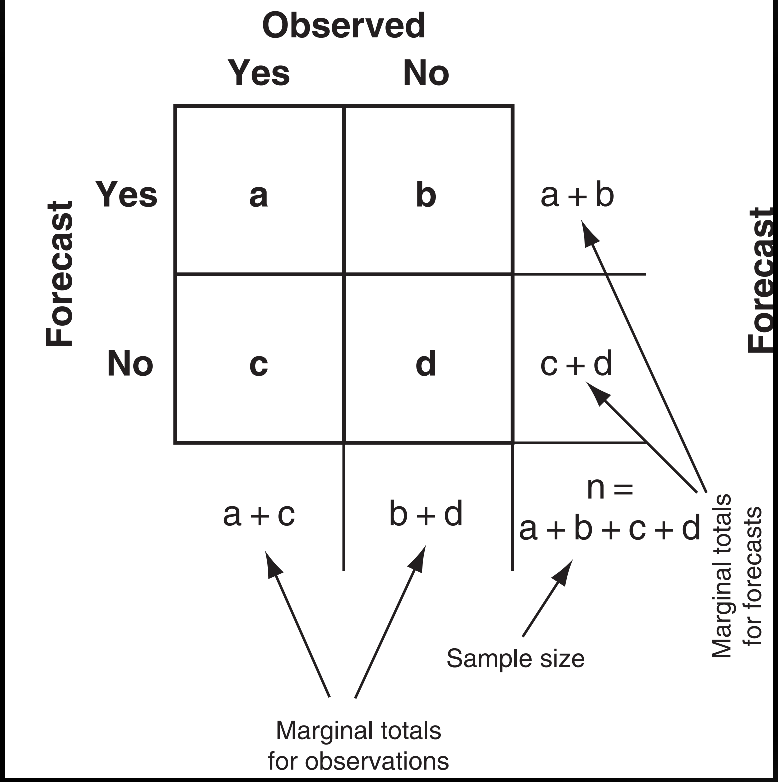
- the number of correct yes forecasts divided by the total number of occasions on which that event was forecast and/or observed
- A useful metric when the yes event is rare

What is the value of TS for perfect forecasts?

Wilks 2011

diff. between the conditional averages of the forecasts for different values of the observation

Discrimination: Hit Rate



$$H = \frac{a}{a + c}.$$

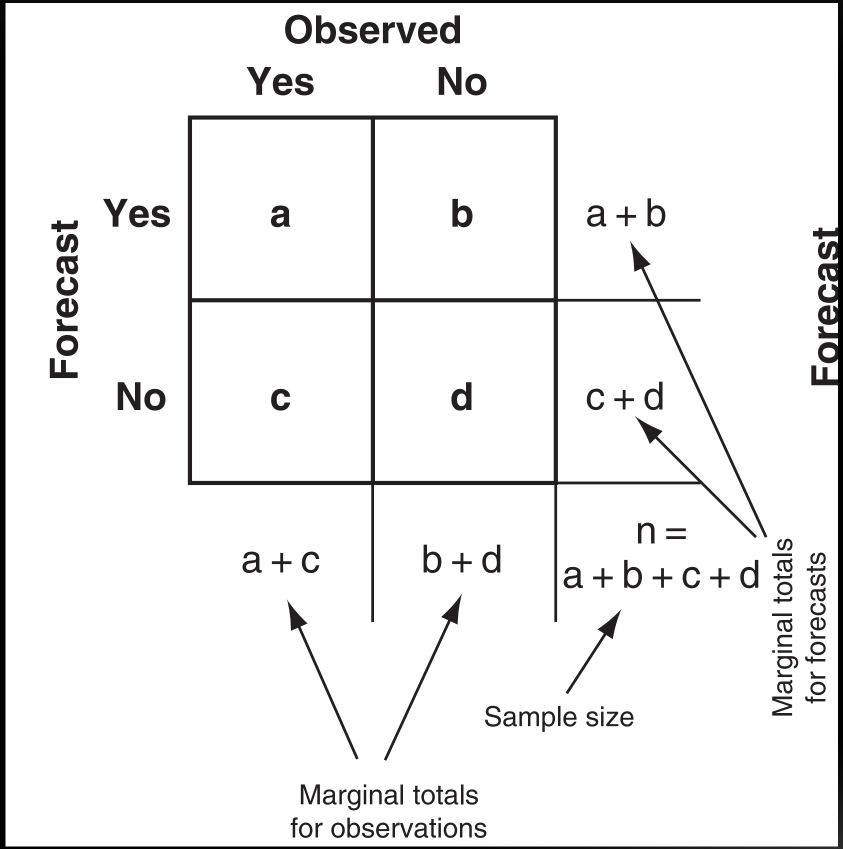
- the ratio of correct yes forecasts to the number of times this event occurred
- It is also called the probability of detection (POD).

What is the value of H for perfect forecasts?

Wilks 2011

diff. between the conditional averages of the forecasts for different values of the observation

Discrimination: False Alarm Rate



$$F = \frac{b}{b+d},$$

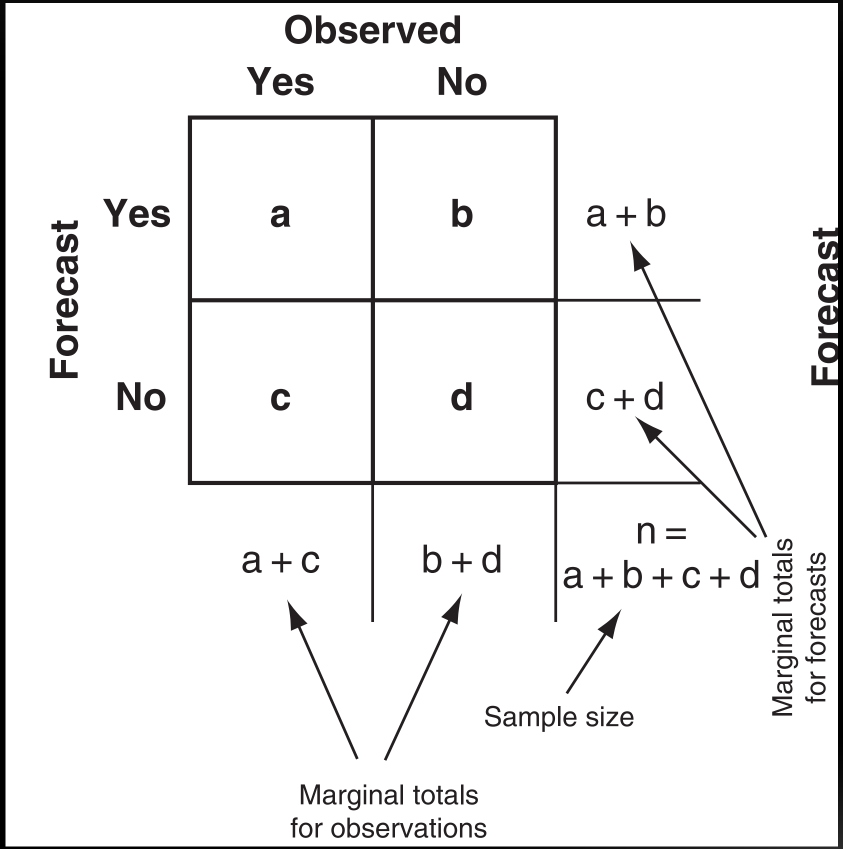
- the ratio of false alarms to the total number of non-occurrences of the event

What is the value of F for perfect forecasts?

Wilks 2011

the conditional distributions of the observations given the forecasts

Reliability and Resolution: False Alarm Ratio



$$\mathrm{FAR} = \frac{b}{a+b}.$$

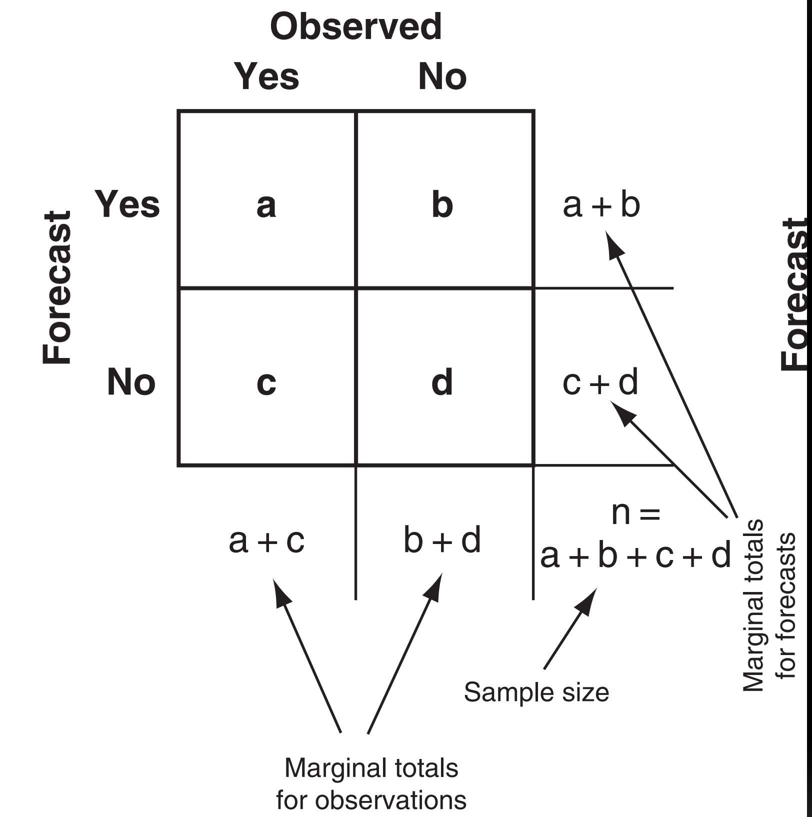
- the fraction of yes forecasts that turn out to be wrong.

What is the value of FAR for perfect forecasts?

Wilks 2011

the correspondence between the average forecast and the average observations

Bias



$$B = \frac{a+b}{a+c}.$$

- the ratio of the number of yes forecasts to the number of yes observations.
- B=1: a perfect forecast
- B>1: overforecasting, the event was forecast more often than observed.
- B<1: underforecasting, the event was forecast less often than observed,

What is the value of B for perfect forecasts?

Wilks 2011

Skill Score

- The skill score for forecasts characterized by a basic accuracy measure A is given by

$$SS_{\text{ref}} = \frac{A - A_{\text{ref}}}{A_{\text{perf}} - A_{\text{ref}}} \times 100\%,$$

- If we use the proportion correct as the basic accuracy measure (i.e., A=(a+d)/n) and use random forecasts as the reference, we have the commonly used Heidke Skill Score (HSS)

$$\begin{aligned} \mathrm{HSS} &= \frac{\boxed{(a+d)/n} - \boxed{[(a+b)(a+c)+(b+d)(c+d)]/n^2}}{1 - \boxed{[(a+b)(a+c)+(b+d)(c+d)]/n^2}} \\ &= \frac{2(ad-bc)}{(a+c)(c+d)+(a+b)(b+d)}, \end{aligned} \tag{7.15}$$

Wilks 2011

References

- Wilks, 2011, “Statistical Methods in the Atmospheric Sciences”, Chapter 7