Confusion Matrix Terms explanation and spreadsheet to accompany the "Introduction to Binary Classification" video

The spreadsheet titled, "Confusion Matrix Terms" is a reference that gives definitions for all the most important metrics related to the performance of binary classification systems – and relates each to the information provided in a "Confusion Matrix" - the standardized format for binary classification measures.

Every Confusion Matrix contains eight numbers.

In this course the eights numbers are consistently labeled a,b,c,d,e,f,g,h as shown on the *Confusion Matrix Terms* spreadsheet..

Every Confusion Matrix shows three different discrete probability distributions.

The cells labeled "a" and "b" in positions E9 and E11 show the probability that, for each event, the Condition is *actually* Present or Absent. By convention presence of the condition being studied is called a "Positive" and its absence a "negative." The probability that the condition is actually present, a, is also called condition "incidence." As a discrete probability distribution, a + b must sum to 1.

The two cells labeled "c" and "d" in positions G6 and I6 show the probability that the classification method used will *classify* an item as positive or negative. These classifications are not always correct – they are an attempt to assign a label "positive or negative" with partial information, acknowledging that errors in labeling will persist. The probability that an event is assigned a positive classification, c, is sometimes called the classification incidence to test incidence. As a discrete probability distribution, c + d must sum to 1.

The four cells labeled e, f, g, and h in positions G9, I9, G11 and H11 show the joint probabilities that each pair of the two Conditions and two Classifications occurs together. Cells labeled e and h show correct classifications, while cell f shows the probability of Positive events falsely classified negative – "False Negatives" (FN) and g shows the probability of negative events falsely classified positive – "False Positives" (FP). As a discrete probability distribution, e + f + g + h must sum to 1.

The confusion matrix provides many metrics for analyzing the performance of a binary classification method. These metrics, listed in Column Q of the Spreadsheet, rows 10-18, each have different uses, and are explained and discussed throughout this course.

Simply note for now that each metric can be thought of as a *conditional* probability – the conditional probability of a certain Test Classification, *given* a certain actual Condition [Q10-Q13] – or the conditional probability of a certain actual Condition, *given* a certain test classification [Q16-Q19].