

APPENDIX B

Methodology of Approximating Probability from Histograms

This appendix explains how the histogram data have been processed. It also defines and symbolizes certain operations associated with the process.

The basic data portrayed in the histograms are counts of how often the data falls within a defined subset (sometimes referred to as a “bin”) of a larger set of data made up of two or more bins. The ratio of the counts within a particular bin to the total number of counts in all of the bins corresponding to a given variable is an approximation to the chance, likelihood, or probability that the variable takes on a value within that particular bin. Consider a two-column, n-row array of data a_{ij} as represented in Table B-1.

Table B-1. Hypothetical data for two histograms: one for R when E is true and one for R when E is false

variable \ sort	E	not E	$P(E R_i)$	$P(R_i E)$
R1	a_{11}	a_{12}	$a_{11} / (a_{11}+a_{12})$	$a_{11} / S_i(a_{i1})$
R2	a_{21}	a_{22}	$a_{21} / (a_{21}+a_{22})$	$a_{21} / S_i(a_{i1})$
R3	a_{31}	a_{32}	$a_{31} / (a_{31}+a_{32})$	$a_{31} / S_i(a_{i1})$
---	---	---	---	---
R_i	a_{i1}	a_{i2}	$a_{i1} / (a_{i1}+a_{i2})$	$a_{i1} / S_i(a_{i1})$
---	---	---	---	---
R_n	a_{n1}	a_{n2}	$a_{n1} / (a_{n1}+a_{n2})$	$a_{n1} / S_i(a_{i1})$

The counts in the bins of these two histograms are represented by a_{i1} from $i = 1$ to n when E is true and by a_{i2} from $i = 1$ to n when E is false. R is a variable binned into n bins (“ R_i ” stands for its i th bin). (Although R stands for range in the main body of the report, as shown in Figure 3, in this discussion R may represent any variable that has been sampled and sorted to make a histogram.) The symbol E stands for a logical variable that is either true or false. For example, in most applications E represents “engaged” which means that the counts for various values of R (that is the counts corresponding to each R_i) were obtained when the system was engaged (CCC driving on the first week, or ACC on the second week). “Not E” means that the data is for manual driving. In this sense the logical variable E stands for a variable that is used in a sorting operation to split the data into histograms that are very useful for comparing ACC driving with manual driving or CCC driving with manual driving.

The column labeled "P (E I Ri)" is like the conditional probability for being engaged given that row i (that is, R falls in the Ri bin) is true. This approximate probability is used, for example, to answer questions like: If a trip is approximately 10 miles long, what is the chance that the ACC system would be engaged during that trip? For answering this example question, the variable listed in the first column represents the length of trips and the entries in the second column are the counts of trips in various length categories (bins) when the control system is engaged. The fourth column, labeled P (E I Ri), gives the approximate probabilities for all lengths of trips. The answer to the example question will be found in the fourth column in the row corresponding to trips that are approximately 10 miles long.

The last column is an approximation to the probability density function for R when E is true. (Although not illustrated, the probability density function for R when E is not true is defined similarly using ai2 in place of ail.) The symbol "Si (ail)" represents the sum of all the counts for all of the bins constituting R. By plotting and comparing, P(Ri I E) with P(Ri I not E), one can compare ACC driving with manual driving with respect to the variable R.

This discussion is tedious but fundamental. It may be easier to understand after examining the results presented later. The ideas behind having a sample space as defined in probability theory may be useful for visualizing the reasoning. We are simply counting the number of members (samples) in various subsets and using these counts to estimate probabilities and conditional probabilities.

The symbol P(· I ·) may be viewed as an operator that performs the operation as defined above on the sets indicated as inputs to the operator. For example P(E I Ri) is the fraction of the set Ri for which E is true. The symbol P(Ri I R) would mean the fraction of the complete set R for which R falls in the Ri bin. This is cumbersome when there are many bins (i is large) and may be shortened to Pd(R I E) to indicate an approximation to the probability density function for R when E is true.

In general, when we are addressing questions of the form "When is ACC likely to be used?", we will be comparing P(E I Si) with P(notE I Si) for various values of i across the set S. In contrast Pd(S I E) is used to answer questions concerning performance with respect to the variable represented by the subset of S defined by E being true. For example if R represents the set of range counts for the range variable, one can examine Pd(R I E) to determine the chance that range will be short given that the ACC system is in operation.