Probability

Random Experiment a process or course of action resulting in outcome(s) which cannot be predicted with certainty

Sample Space set of all possible outcomes for a random experiment; outcomes must be *mutually exclusive* and *exhaustive*

Simple Event an outcome in a sample space

Probability Value one of a set of values which satisfy the following three conditions

$$0 \leq P(E_i) \leq 1$$

$$\sum P(E_i) = 1 \qquad \text{(discrete)}$$

$$\int P(E_i)dx = 1 \qquad \text{(continuous)}$$

$$P(E_i \text{ or } E_j) = P(E_i) + P(E_j) \qquad E_i, E_j \text{ mutually exclusive}$$

the $\sum$ and $\int$ expressions are also known as the *Law of Total Probability.*

Complement $P(A) = 1 - P(\bar{A})$

Addition (General) $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Addition (Mutually Exclusive Events) $P(A \text{ or } B) = P(A) + P(B)$

Conditional Probability $P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$

Bayes Rule given k categories for event A, $P(B) \neq 0$, and $1 \leq i \leq k$, then

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{\sum_{j=1}^{k} P(A_j)P(B \mid A_j)}$$

which can also be stated as

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)}$$

given that, by the *Law of Total Probability*

$$P(B) = \sum_{j=1}^{k} P(A_j)P(B \mid A_j)$$

Prior Probability the probability of an event *before* new information has been taken into account; the initial value of a *prior* is the unconditioned probability of the event in question.

Posterior Probability the probability of an event *after* new information has been taken into account.

Classification assigning group membership to each observation in a set of observations

Minimum Bayes Risk Classification classification procedure which assigns group membership based on the highest value from a set of group membership probabilities (*i.e.*, the *most likely* category)

Misclassification assignment of an observation to an incorrect category

True Positive (TP) correct identification of the presence of a condition

False Positive (FN) incorrect identification of the presence of a condition; *a.k.a.*: type I error rate (α); can be calculated as $1 - specificity$

True Negative (TN) correct identification of the absence of a condition

False Negative (FN) incorrect identification of the absence of a condition; *a.k.a.*: type II error rate (β); can be calculated a $1 - sensitivity$

Positive Predictive Value calculated as $= TP/(TP + FP)$; *a.k.a.*: precision

Negative Predictive Value calculated as $= TN/(FN + TN)$ *a.k.a.*: recall

Sensitivity the proportion of instances where a condition actually exists and a classification (diagnostic) procedure indicates the presence of that condition (positive result); the probability of a positive test result, given that the condition does exist; can be interpreted as the probability of a positive test result given an underlying true positive condition; calculated as $= TP/(TP + FN)$

Specificity the proportion of instances where a condition does not exist and a classification (diagnostic) procedure indicates the absence of that condition (negative result); the probability of a negative test result, given that the condition does not exist; calculated as the number of

true negatives divided by the total of true negatives + false positives; can be interpreted as the probability of a negative test result given an underlying true negative condition; calculated as $= TN/(FP + TN)$

Power *a.k.a.*: sensitivity; calculated as $1 - \beta$

Conditional Independence if $P(A \mid B) = P(A)$ or $P(B \mid A) = P(B)$

Intersection (General)

$$\begin{aligned} P(A \text{ and } B) &= P(A)P(B \mid A) \\ &= P(B)P(A \mid B) \end{aligned}$$

Intersection (Independent Events) $P(A \text{ and } B) = P(A)P(B)$

Random Variable (r.v.) a function which assigns a numerical value to each possible outcome in a sample space

Stochastic Table a table of numbers where all entries are in the range 0.0 to 1.0 and each row (or column) sums to 1.0; a table where the rows *and* columns both sum to 1.0 is known as *doubly stochastic*

Example 4-1

Conditional Probabilities in a Dice Game

A single die is a cube with the number 1 on one of the sides, the number 2 on another side, and so forth for all six sides. When a die is "rolled" or "thrown" the side which is on top displays the "value" of that roll or throw. If a particular die if "fair" (unbiased), the probability of each possible outcome is the same as for any other possible outcome, *i.e.*, $\frac{1}{6}$. It should be recognized that a die has no memory and will not "compensate" for the failure to deliver a particular value during a set of rolls.

Craps is a game played with a pair of dice. The pair is thrown together and the sum of their values is used. The following table shows the possible outcome values from throwing a pair of dice. The labels D1 and D2 are for the first and second die, respectively.

.	$D1 = 1$	$D1 = 2$	$D1 = 3$	$D1 = 4$	$D1 = 5$	$D1 = 6$
$D2 = 1$	2	3	4	5	6	7
$D2 = 2$	3	4	5	6	7	8
$D2 = 3$	4	5	6	7	8	9
$D2 = 4$	5	6	7	8	9	10
$D2 = 5$	6	7	8	9	10	11
$D2 = 6$	7	8	9	10	11	12

Counting the unique outcome values gives us the following frequency distribution.

Value	2	3	4	5	6	7	8	9	10	11	12	*Total*
Frequency	1	2	3	4	5	6	5	4	3	2	1	36

Based on this frequency distribution, we can see that the *probability* of throwing a 7 is $\frac{6}{36}$ which simplifies to $\frac{1}{6}$.

The rules for craps start with the definitions of *craps*, which is the throwing of a 2, 3, or 12 on the initial roll of a game. This is an automatic *lose*. The throwing of a 7 or 11 on the initial roll of a game is an automatic *win*. Throwing anything else on the initial roll of a game defines what is known as a *point*. Thus,

$$\begin{aligned} P(lose) &= P(2,3,12) \\ &= \frac{1+2+1}{36} \\ &= \frac{4}{36} \\ P(win) &= P(7,11) \\ &= \frac{6+2}{36} \\ &= \frac{8}{36} \\ P(point) &= P(4,5,6,8,9,10) \\ &= \frac{3+4+5+5+4+3}{36} \\ &= \frac{24}{36} \end{aligned}$$

Once a *point* has been established, the goal is to throw the dice again until either the *point* value is obtained, in which case the player *wins* or the value 7 is obtained, in which case the player *loses*. All other possible outcomes from a roll of the dice are irrelevant at this stage of the game.

Once a *point* has been established, the probability of wining depends on the value of the *point*. For example, if the *point* is 4, then the outcome values 2, 3, 5, 6, 8, 9, 10, 11, and 12 are irrelevant to outcome of the game. Thus we have the following table to work with:

$Value$	$Frequency$	$Percent$
4	3	0.333
7	6	0.667
$Total$	9	1.000

This *conditional frequency table* shows us that there is now a $\frac{3}{9}$ chance of winning the game, or we could write $P(win \mid point = 4) = \frac{3}{9}$. The game can be started again once it has been won or lost.

NOTE: This is basic description of the game for teaching purposes and does *not* suggest that the game be played, nor presents any kind of strategy or suggestions on how to place wagers.

Example 4-2

Conditional Probabilities in Diagnosis/Classification

Suppose a disease occurs in a population at a rate of 8 per 1,000 population. There is a quick and inexpensive test for this disease and 7 out of 8 individuals who have the disease (true positive) will be identified correctly by this test. Of the remaining population who do not have the disease, an additional 70 individuals will be incorrectly identified as having the disease (false positive).

Suppose a person takes this screening test for this disease, what is the probability that an individual with a *positive* test result really has the disease? If the same person gets a *negative* test result, what is the probability that they have the disease?

Organizing this information into tabular form, with the actual condition across rows and the test results down columns, we have

	Tested Positive	*Tested Negative*	*Actual*
Disease	7	1	8
No Disease	70	922	992
Tested	77	923	1000

from this table we can identify the number of *true positives* ($tp = 7$), the number of *false positives* ($fp = 1$), the number of *false negatives* ($fn = 70$) and the number of *true negatives* ($tn = 922$). The *sensitivity* of the classification would be found as

$$\begin{aligned} sensitivity &= \frac{tp}{tp + fn} \\ &= \frac{7}{7 + 70} \\ &= 0.090909 \end{aligned}$$

and the *specificity* of the classification would be found as

$$\begin{aligned} specificity &= \frac{tn}{tn + fp} \\ &= \frac{922}{922 + 1} \\ &= 0.998917 \end{aligned}$$

Using this information, we can see that of the people with positive results, 7 have the disease and 70 do not have the disease. Thus, we know that a *positive* test result really says they have a $\frac{7}{70+7}$ chance of having the disease. Again, from the table, we see that of the people with negative results, 1 has the disease and 922 do not. Using this information we know that a *negative* test result really says they have a $\frac{1}{922+1}$ chance of having the disease.

Using *Bayes Rule*, we can calculate these same numbers as follows:

Initial Probability Estimate

Before any diagnostic procedure, diagnostic test or additional information is considered, we are given the *prior* probability

$$\begin{aligned} P(disease) &= \frac{disease}{tested} \\ &= \frac{8}{1000} \\ &= 0.008 \end{aligned}$$

based completely on the *actual* observed count of the disease compared to the actual number tested.

Test with Positive Result

We now assume that a diagnostic procedure or test has been performed to evaluate a single individual or entity. Given a *positive test result* (which we will simply designate *positive*), this additional information, together with the original table, would allow us to find the *conditional probability* of the disease as

$$\begin{aligned} P(disease \mid positive) &= \frac{disease}{disease + no\ disease} \\ &= \frac{7}{7+70} \\ &= 0.090909 \end{aligned}$$

using the information from the *Tested Positive* column.

Using *Bayes theorem*, we would come up with the same final result by starting with the definition

$$P(disease \mid positive) = \frac{P(disease)P(positive \mid disease)}{P(disease)P(positive) + P(no\ disease)P(positive)}$$

and then finding values for each of the expressions used in that definition

$$\begin{aligned} P(disease) &= 8/1000 \\ &= 0.008 \\ P(no\ disease) &= 1 - P(disease) \\ &= 1 - 0.008 \\ &= 0.992 \\ P(positive \mid disease) &= 7/8 \\ &= 0.875 \\ P(positive) &= 77/1000 \\ &= 0.077 \end{aligned}$$

from which we can do the final calculations, substituting back into the original definition, as

$$\begin{aligned} P(disease \mid positive) &= \frac{P(disease)P(positive \mid disease)}{P(disease)P(positive) + P(no\ disease)P(positive)} \\ &= \frac{(0.008)(0.875)}{(0.008)(0.077) + (0.992)(0.077)} \\ &= 0.090909 \end{aligned}$$

Note that we could have arrived at the same final result by the alternate form of

$$\begin{aligned} P(disease \mid positive) &= \frac{P(disease)P(positive \mid disease)}{P(positive)} \\ &= \frac{(8/1000)(7/8)}{77/1000} \\ &= \frac{7/1000}{77/1000} \\ &= \frac{7}{77} \\ &= 0.090909 \end{aligned}$$

because we can see that

$$P(positive) = P(disease)P(positive) + P(no\ disease)P(positive)$$

using the *Law of Total Probability.*

Test with Negative Result

Again, we assume that a diagnostic procedure or test has been performed to evaluate a single individual or entity. Given a *negative result* (which we will simply designate *negative*), this additional information, together with the original table, would allow us to find the *conditional probability* of the disease as

$$\begin{aligned} P(disease \mid negative) &= \frac{disease}{disease + no\ disease} \\ &= \frac{1}{1 + 922} \\ &= 0.001083 \end{aligned}$$

using the information from the *Tested Negative* column.

Using *Bayes theorem*, we would, again, come to the same final result by starting with the definition

$$P(disease \mid negative) = \frac{P(disease)P(negative \mid disease)}{P(disease)P(negative) + P(no\ disease)P(negative)}$$

and then finding values for each of the expressions used in that definition

$$\begin{aligned} P(disease) &= 8/1000 \\ &= 0.008 \\ P(no\ disease) &= 1 - P(disease) \\ &= 1 - 0.008 \\ &= 0.992 \\ P(negative) &= 923/1000 \\ &= 0.923 \\ P(negative \mid disease) &= 1/8 \\ &= 0.125 \end{aligned}$$

from which we can do the final calculations, substituting back into the original definition, as

$$\begin{aligned} P(disease \mid negative) &= \frac{P(disease)P(negative \mid disease)}{P(disease)P(negative) + P(no\ disease)P(negative)} \\ &= \frac{(0.008)(0.125)}{(0.008)(0.923) + (0.992)(0.923)} \\ &= 0.001083 \end{aligned}$$

Note that we could have arrived at the same final result by the alternate form of

$$\begin{aligned} P(disease \mid negative) &= \frac{P(disease)P(negative \mid disease)}{P(negative)} \\ &= \frac{(8/1000)(1/8)}{923/1000} \\ &= \frac{1/1000}{923/1000} \\ &= 1/923 \\ &= 0.001083 \end{aligned}$$

because, again, we can see that

$$P(negative) = P(disease)P(negative) + P(no\ disease)P(negative)$$

using the *Law of Total Probability*.

Suggested reading:

Calculated Risks: How To Know When Numbers Deceive You, Gerd Gigerenzer, 2002, Simon and Schuster, ISBN: 978-0-7432-5423-6,

The theory that would not die, Sharon Bertsch McGrayne, 2001, Yale University Press, ISBN: 978-0-300-16969-0

Example 4-3

Minimum Bayes Risk Classification

A particular kind of item (the infamous *widget*) has four distinct types into which each instance can be classified: **A**, **B**, **C**, or **D**.

A sample of 200 of these items is obtained, and each instance is manually classified using a difficult, time-consuming process. The results of this manual classification are presented in the table below, together with the *percentages* which we will use as estimates of group membership probability.

Category	*Frequency*	*Percentage*
A	50	0.250
B	75	0.375
C	45	0.225
D	30	0.150
Total	200	1.000

The *Minimum Bayes Risk Classification* minimizes misclassification risk by assigning an observation to the classification which has the *highest* estimated probability. *Bayes Rule* can be applied to use known features of the observation under consideration to improve the estimated probability.

Given this table, the *Minimum Bayes Risk Classification* for each item is category **B**. Assigning category **B** to each item would be correct 37.5% of the time ($\frac{3}{8}$), which is correct more often than any of the other categories, but not particularly good. The value of the *Bayes Risk* for each classification is $0.250 + 0.225 + 0.150 = 0.625$ or $\frac{5}{8}$. Now we find that there are three independent characteristics of these items which can be readily observed and that each of these characteristics each have three possible values. We will refer to these characteristics and their values as Factor-1 (E, F, or G), Factor-2 (H, I, or J), and Factor-3 (K, L, or M). The *contingency tables* showing the joint incidence of each characteristic and the classified category are listed below.

$Factor-1$	E	F	G	$Total$
A	15	10	25	50
B	5	20	50	75
C	10	15	20	45
D	10	5	15	30
$Total$	40	50	110	200

$Factor-2$	H	I	J	$Total$
A	10	25	15	50
B	25	5	45	75
C	5	15	25	45
D	5	10	15	30
$Total$	45	55	100	200

$Factor-3$	K	L	M	$Total$
A	10	25	15	50
B	5	32	38	75
C	8	22	15	45
D	15	8	7	30
$Total$	38	87	75	200

These three tables can now be used to obtain estimates of the *conditional probabilities* of each classification category *given* the information observed about a given characteristic. These (column) stochastic matrices are listed below (printed to three decimal places).

$Factor-1$	E	F	G
A	0.375	0.200	0.227
B	0.125	0.400	0.455
C	0.250	0.300	0.182
D	0.250	0.100	0.136
$Total$	1.000	1.000	1.000

$Factor-2$	H	I	J
A	0.222	0.455	0.150
B	0.556	0.091	0.450
C	0.111	0.273	0.250
D	0.111	0.182	0.150
$Total$	1.000	1.000	1.000

$Factor-3$	K	L	M
A	0.263	0.287	0.200
B	0.132	0.368	0.507
C	0.211	0.253	0.200
D	0.395	0.092	0.093
$Total$	1.000	1.000	1.000

The question at this point is how to use the information we have collected. We will now go through the steps of using *Bayes Rule* to take into account the conditional probabilities and give us better information about the probability of group membership for some unclassified observations.

We have an unclassified observation with characteristics (E, I, L). We will use *Bayes Rule* to update the classification probabilities based on the observable information in the three characteristics. The process can have any number of steps but always involves three distinct sets of probability values: (1) the *prior probability* which takes into account all the information used till the current processing step, (2) the *conditional probability* which shows the effect of a given value of a characteristic on the distribution of classification, and (3) the *posterior probability* which is the revised probability estimate *after* the conditional probability information has been taken into account.

The process we will use will go through step by step, using information which has been presented earlier in this handout.

Case 1: Observation with characteristics (E, I, L)

Step 1a: Use Factor-1=E.

Classification	*Prior*	*Conditional*	*Product*	*Posterior*
A	0.250	0.375	0.094	0.400
B	0.375	0.125	0.047	0.200
C	0.225	0.250	0.056	0.240
D	0.150	0.250	0.038	0.160
Total	1.000	1.000	0.234	1.000

As can be seen, the column labeled *Product* does *not* sum to 1.0, as probabilities should. The column total (0.234) is used as a *scale factor* to give us the values seen in the column labeled *Posterior*. Notice that the *Minimum Bayes Risk Classification* at this point has changed from category **B** to category **A**.

Step 1b: Use Factor-2=I.

Classification	*Prior*	*Conditional*	*Product*	*Posterior*
A	0.400	0.455	0.182	0.617
B	0.200	0.091	0.018	0.062
C	0.240	0.273	0.065	0.222
D	0.160	0.182	0.029	0.099
Total	1.000	1.000	0.295	1.000

Step 1c: Use Factor-3=L.

$Classification$	$Prior$	$Conditional$	$Product$	$Posterior$
A	0.617	0.287	0.177	0.668
B	0.062	0.368	0.023	0.086
C	0.222	0.253	0.056	0.212
D	0.099	0.092	0.009	0.034
$Total$	1.000	1.000	0.265	1.000

We have now used all the additional information available to us, and see our final estimate of the *Posterior Probability* telling us that the most likely classification is **A**.