Introductory Handout for Information Science

Events and outcomes

Let X be an *event* with N possible mutually exclusive *outcomes* $x_1, x_2, ..., x_N$.

$X \in \{x_1, x_2, ..., x_N\}$ (a discrete distribution)

The probability that $X = x_1$ is p_1, that $X = x_2$ is p_2, that $X = x_i$ is p_1 for $i = 1...N$.
Equivalent expressions: $\Pr(X = x_1) = P(x_1) = p(x_1) = p_1$.

Two important aspects of probability:

Every probability value lies between 0 and 1 inclusive: $0 \le p_i \le 1$, for all i.

The sum of probabilities of all possible outcomes is 1: $\sum_{i=1}^{N} p_i = 1$

Example: an event M with possible outcomes, *gold, silver, bronze,* or *loser*.

$\Pr(M = gold) = 0.05$ $\Pr(M = silver) = 0.10$
$\Pr(M = bronze) = 0.15$ $\Pr(M = loser) = 0.70$

Find the probability that M is *gold*, *silver*, or *bronze*.

Answer:

$\Pr(M$ is *gold*, *silver*, or *bronze* $)$ $= \Pr(M$ is *gold*$) + \Pr(M$ is*silver*$) + \Pr(M$ is*bronze*$)$
$= 0.05 + 0.10 + 015 = 0.30$

OR

$\Pr(M$ is *gold*, *silver*, or *bronze* $)$ $= 1 - \Pr(M$ is *loser*$) = 1 - 0.70 = 0.30$

Random variables

A **random variable** X takes on a value from a given set. Thus, it is an event where the outcomes $x_1, x_2, ..., x_N$ have numerical values.

The **expected value** of X is $\sum_{i=1}^{N} p_i x_i$.

Example: Find the expected value of X if

$\Pr(X = 2) = 0.15$ $\Pr(X = 5) = 0.45$
$\Pr(X = 6) = 0.20$ $\Pr(X = 8) = 0.20$

Answer: $\sum_{i=1}^{N} p_i x_i = (0.15)(2) + (0.45)(5) + (0.20)(6) + (0.20)(8) = 0.3 + 2.25 + 1.20 + 1.60 = 5.35$

More than one random variable

Two random variables X and Y, with $X \in \{x_1, x_2, ..., x_N\}$ and $Y \in \{y_1, y_2, ..., y_M\}$

Let X and Y be two simultaneous events with outcomes x_i and y_j. This joint event has a probability $p(x_i, y_j)$. These probabilities can be written in matrix form. Note that the rows sum to the total probability of the corresponding x_i, and the columns sum to the total probability of the corresponding y_j.

$$\begin{array}{ccccccc}
p(x_1,y_1) & p(x_1,y_2) & \cdots & p(x_1,y_M) & \rightarrow & p(x_1) \\
p(x_2,y_1) & p(x_2,y_2) & \cdots & p(x_2,y_M) & \rightarrow & p(x_2) \\
\cdots & \cdots & \cdots & \cdots & & \cdots \\
p(x_N,y_1) & p(x_N,y_2) & \cdots & p(x_N,y_M) & \rightarrow & p(x_N) \\
\downarrow & \downarrow & & \downarrow & & \downarrow \\
p(y_1) & p(y_2) & \cdots & p(y_M) & \rightarrow & 1
\end{array}$$

The sums of the columns and rows is mathematically expressed as follows:

Rows: $p(x_i) = \sum_{j=1}^{M} p(x_i, y_j)$ Columns: $p(y_j) = \sum_{i=1}^{N} p(x_i, y_j)$

The sum of all the joint probabilities is 1:

$$\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i,y_j) = \sum_{i=1}^{N} p(x_i) = \sum_{j=1}^{M}\sum_{i=1}^{N} p(x_i,y_j) = \sum_{j=1}^{M} p(y_j) = 1.$$

Example: Given the joint probabilities below, find the probability of each X and each Y:

$$p(x_1,y_1) = 0.1 \quad p(x_1,y_2) = 0.0 \quad p(x_1,y_3) = 0.2$$

$$p(x_2,y_1) = 0.1 \quad p(x_2,y_2) = 0.1 \quad p(x_2,y_3) = 0.5$$

Answer: Using the summation formulas we get $p(x_1) = 0.3$, $p(x_2) = 0.7$ for the X values and $p(y_1) = 0.2$, $p(y_2) = 0.1$, $p(y_3) = 0.7$ for the Y values.

Conditional Probabilities (Baye's Theorem)

The probability distribution across the X variables may change, depending on whether the Y value is known (and vice-versa). If one variable does not influence the other's probability, the variables are called *independent*.

The probability that $X=x_i$ given that we know $Y=y_j$ is written $p(x_i|y_j)$. This is also called the probability of x_i conditioned on y_j. Baye's theorem expresses the conditional probability as the quotient of the joint probability and the probability of the condition. Equivalently, the joint probability is the product of the conditional probability and the probability of the condition. Note that since $p(x_i, y_j) = p(y_j, x_i)$, there are two such products (either x_i or y_j can be the condition):

$$p(x_i|y_j) = \frac{p(x_i, y_j)}{p(y_j)} \quad \text{OR} \quad p(x_i, y_j) = p(x_i|y_j)p(y_j) = p(y_j|x_i)p(x_i)$$

The example below computes conditional probabilities from the joint distribution on the previous page:

Example: Given the joint probabilities below, find the conditional probabilities of each X given each Y:

$$p(x_1, y_1) = 0.1 \quad p(x_1, y_2) = 0.0 \quad p(x_1, y_3) = 0.2$$

$$p(x_2, y_1) = 0.1 \quad p(x_2, y_2) = 0.1 \quad p(x_2, y_3) = 0.5$$

Answer: The conditional probabilities are found by dividing each probability by the corresponding $p(y_j)$:

$$p(x_1|y_1) = 0.5 \quad p(x_1|y_2) = 0.0 \quad p(x_1|y_3) = \frac{2}{7}$$

$$p(x_2|y_1) = 0.5 \quad p(x_2|y_2) = 1.0 \quad p(x_2|y_3) = \frac{5}{7}$$

Note that the columns all sum to 1 since they are all probability distributions across the X values, given different values of Y. The conditional probabilities of Y given X can be found by dividing by $p(x_i)$. In this case, the *rows* should sum to 1.

Independent variables

Two random variables X and Y are *independent* if knowledge of one has no effect on the probability distribution of the other one.
That is, if $p(x_i|y_1) = p(x_i|y_2) = \text{K} = p(x_i|y_M) = p(x_i)$, then X and Y are independent. Using Baye's Theorem, we can show that X and Y are independent if and only if the joint probabilities are always equal to the product of the individual values:
That is, when $p(x_i, y_j) = p(x_i)\, p(y_j)$ for all i and j.

Logarithms

The logarithm is the inverse of the exponential function:

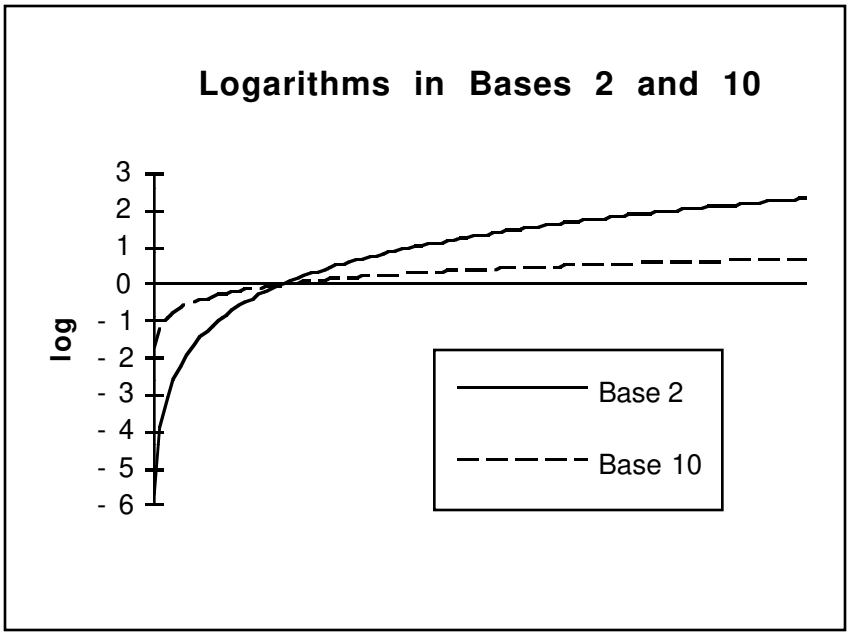
Definition: $y = \log_b x$ is equivalent to $x = b^y$.

Identities about logarithms:

$$\log_b xy = \log_b x + \log_b y$$

$$\log_b \frac{x}{y} = \log_b x - \log_b y$$

$$\log_b \frac{1}{x} = -\log_b x$$

$$\log_b 1 = 0$$

$$\log_b x = \frac{\log_c x}{\log_c b}$$