Statistische Methoden in der Computerlinguistik

2 More Probability Theory

2.1 Random Variables

Idea:

- Axioms: Ω can be any (non-empty) set.
- It is often easier (sometimes even more natural) to restrict our observations to a set of numbers, such as $\mathbb{R}$.
- *Random variables* are used to map any sample space Ω (partially) to $\mathbb{R}$.

Definition 1 (Random Variable) *Let S be an event space over a set Ω of basic outcomes, and let P be a probability distribution over S. A* random variable (over Ω) *is a function:*

$$X : \Omega \to \mathbb{R}$$

such that for all $x \in \mathbb{R} : \{\omega \in \Omega \mid X(\omega) = x\} \in S$

Example 1 (Coin Toss: Random Variable) Let $\Omega = \{heads, tails\}$. Then, $X = \{(heads, 0), (tails, 1)\}$ is a random variable over Ω:

$$\begin{aligned} X(heads) &= 0 \\ X(tails) &= 1 \end{aligned}$$

Example 2 (Two Dice: Random Variable) For the throw of two dice, let $\Omega = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$. Then, $X = \bigcup_{i=1}^{6} \bigcup_{j=1}^{6} \{((i, j), i + j)\}$ is a random variable over Ω:

$$\begin{aligned} X(1,1) &= 1+1 = 2 \\ X(1,2) &= 1+2 = 3 \\ &\vdots \\ X(6,6) &= 6+6 = 12 \end{aligned}$$

2.1.1 Discrete *vs.* Continuous Random Variables

A *discrete random variable* is one whose image under Ω is finite or countable. *Continuous random variables* are out there too, but they will not concern us in this course.

2.2 Probability Mass Functions

A *probability mass function* (sometimes simply called a *probability function*) is determined by a random variable X and a probability distribution P over $\Omega = dom(X)$.

Definition 2 (Probability Mass Function) *Let X be a random variable over a set Ω of basic outcomes, and let P be a probability distribution over Ω. Then, the* probability mass function (pmf) *associated with X is given by:*

$$p_X(x) = p(X = x) = P(\{\omega \in \Omega \mid X(\omega) = x\})$$

It can be shown that p_X is always a probability distribution over $\mathbb{R}$. Often, the subscript X is omitted from the probability mass function p_X when the random variable concerned is clear from the context.

Example 3 (Coin Toss: Probability Function) Assuming the coin is fair, $P(heads) = P(tails) = \frac{1}{2}$. Therefore, $p_X(0) = p_X(1) = 0.5$, and $\forall x \in \mathbb{R} - \{0, 1\} : p_X(x) = 0$.

Example 4 (Two Dice: Probability Function) Probability mass is distributed according to the following table:

x	$X^{-1}(x)$	$p_X(x)$
2	$\{(1,1)\}$	1/36
3	$\{(1,2),(2,1)\}$	2/36
4	$\{(1,3),(2,2),(3,1)\}$	3/36
5	$\{(1,4),(2,3),(3,2),(4,1)\}$	4/36
6	$\{(1,5),(2,4),(3,3),(4,2),(5,1)\}$	5/36
7	$\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$	6/36
8	$\{(2,6),(3,5),(4,4),(5,3),(6,2)\}$	5/36
9	$\{(3,6),(4,5),(5,4),(6,3)\}$	4/36
10	$\{(4,6),(5,5),(6,4)\}$	3/36
11	$\{(5,6),(6,5)\}$	2/36
12	$\{(6,6)\}$	1/36

2.3 Some Properties of Random Variables

2.3.1 Sample Space

Definition 3 (Sample Space of a Random Variable) *The* sample space of a random variable X *is just the image of X under Ω, and is written Ω_X*

$$\Omega_X := X(\Omega) = \bigcup_{\omega \in \Omega} \{X(\omega)\}$$

2.3.2 Expectation Value

Definition 4 (Expectation Value) *The* expectation value $E(X)$ *of a random variable* X *is simply the* mean *or* average value *of that variable, computed as a weighted sum of the variable's sample space:*

$$E(X) = \sum_{x \in \Omega_X} x \cdot p_X(x)$$

One common convention uses the Greek letter μ to denote the expectation value of a random variable, when the particular variable is clear from the surrounding context: $\mu = E(X)$.

Example 5 (Coin Toss: Expectation Value)

$$\begin{array}{rcccccccc} E(X) & = & X(heads) & \cdot & p_X(X(heads)) & + & X(tails) & \cdot & p_X(X(tails)) \\ & = & 0 & \cdot & 0.5 & + & 1 & \cdot & 0.5 \\ & = & 0.5 & & & & & & \end{array}$$

Example 6 (Two Dice: Expectation Value)

$$\begin{array}{rcl} E(X) & = & \frac{1}{36} \cdot 2 + \frac{2}{36} \cdot 3 + \frac{3}{36} \cdot 4 + \frac{4}{36} \cdot 5 + \frac{5}{36} \cdot 6 + \frac{6}{36} \cdot 7 + \\ & & \frac{5}{36} \cdot 8 + \frac{4}{36} \cdot 9 + \frac{3}{36} \cdot 10 + \frac{2}{36} \cdot 11 + \frac{1}{36} \cdot 12 \\ & = & 7 \end{array}$$

Every function $g : \mathbb{R} \to \mathbb{R}$ can be used to map a random variable X to a new random variable $Y = g(X)$. The expectation value of such a functionally composed random variable is given by:

$$E(Y) = E(g(X)) = \sum_{x \in \Omega_X} g(x) \cdot p(x)$$

In particular, it is interesting (and useful) to note that ...

Theorem 1 (Expectation Value of Linear Functions) *If* $g : \mathbb{R} \to \mathbb{R}$ *is a linear function – that is, if* $g(x) = ax + b$ *for some constants* $a, b \in \mathbb{R}$ *and for all* $x \in \mathbb{R}$*, then* $E(g(X))$ *can be computed as a function of* $E(X)$*:*

$$E(g(X)) = E(aX + b) = aE(X) + b$$

Theorem 2 (Sum of Expectation Values) *The sum of linear combinations of arbitrary random variables can be computed in terms of the expectation values of the variables themselves:*

$$E(aX + bY) = aE(X) + bE(Y)$$

Theorem 3 (Product of Independent Expectation Values) *If X and Y are independent random variables, then:*

$$E(X \cdot Y) = E(X) \cdot E(Y)$$

2.3.3 Variance

Definition 5 (Variance) *The* variance *of a random variable X is a measure of how widely that variable's values are distributed, computed as the average square difference between the variable's values its mean:*

$$\begin{aligned} Var(X) &:= E((X - E(X))^2) \\ &= E(X^2) - E(X)^2 \end{aligned}$$

It is common to write σ^2 to refer to the variance of a random variable, when the random variable in question is clear from the surrounding context: $\sigma^2 = Var(X)$. This is largely due to the fact that the *standard deviation* – commonly written σ – is defined as the square root of the variance: $\sigma = \sqrt{Var(X)}$.

Example 7 (Coin Toss: Variance)

$$\begin{aligned} Var(X) &= E((X - 0.5)^2) \\ &= \sum_{x \in \Omega_X} p_X(x) \cdot (x - 0.5)^2 \\ &= 0.5 \cdot (0 - 0.5)^2 + 0.5 \cdot (1 - 0.5)^2 \\ &= 0.5 \cdot -0.5^2 + 0.5 \cdot .5^2 \\ &= 0.5 \cdot 0.25 + 0.5 \cdot .25 \\ &= 0.125 + 0.125 \\ &= 0.25 \end{aligned}$$

Example 8 (Two Dice: Variance)

$$\begin{aligned} Var(X) &= E(X^2) - E(X)^2 \\ &= 2^2 \cdot \tfrac{1}{36} + 3^2 \cdot \tfrac{2}{36} + 4^2 \cdot \tfrac{3}{36} + 5^2 \cdot \tfrac{4}{36} + 6^2 \cdot \tfrac{5}{36} + 7^2 \cdot \tfrac{6}{36} + \\ &\quad 8^2 \cdot \tfrac{5}{36} + 9^2 \cdot \tfrac{4}{36} + 10^2 \cdot \tfrac{3}{36} + 11^2 \cdot \tfrac{2}{36} + 12^2 \cdot \tfrac{1}{36} \\ &\quad -7^2 \\ &= 4 \cdot \tfrac{1}{36} + 9 \cdot \tfrac{2}{36} + 16 \cdot \tfrac{3}{36} + 25 \cdot \tfrac{4}{36} + 46 \cdot \tfrac{5}{36} + 49 \cdot \tfrac{6}{36} + \\ &\quad 64 \cdot \tfrac{5}{36} + 81 \cdot \tfrac{4}{36} + 100 \cdot \tfrac{3}{36} + 121 \cdot \tfrac{2}{36} + 144 \cdot \tfrac{1}{36} \\ &\quad -49 \\ &= 35/6 \\ &\approx 5.83 \end{aligned}$$