Probability Basics (Lecture 20)

Analysis of Software Artifacts

Applications of Probability and Statistics

- Decision Making
  - should I go with an off-the-shelf component or develop it in-house
  - tradeoffs between cost and reliability
  - should I test more or release a possibly “buggy” software now
  - tradeoff between reliability and time to market

Applications

- Cost models
  - linear regression used in software cost models
  - such as `COCOMO`
  - Bayesian statistics used in cost models as well
  - any more?

Applications

- metrics for software reliability
  - entirely based on statistics
  - *mean time between failures* or *MTBF*
  - expected time between two failure events
- testing
  - random sampling used in testing
  - Markov chains also used to generate test cases

Role of Probability Theory

- probability used as a means for quantifying uncertainties
- probability is dependent on time
  - reference time will be denoted by τ
- known quantities about the software
  - denoted by $\mathcal{H}$
  - amount of testing already done
  - composition of the team
  - the cost of producing it

Random quantities

- quantities that are unknown we will call *random quantities*
  - MTBF
  - number of bugs remaining
  - reliability

Types of Random Quantities

- Random Variables
  - realization of these variables are numbers
  - could real numbers or integers
  - realizations denoted by smaller case letters
  - for random variables T and X, realizations denoted by t and x

Random Events

- Random events
  - distinguishing feature is that a random event only takes two values
  - random events will be generally propositions, e.g. true or false
  - $MTBF$ is greater than a specified time Z
  - any more?

Notation

- $P^{\tau}(E|\mathcal{H})$
  - probability at time τ that event E happens
  - given the past history $\mathcal{H}$
- why is $P^{\tau+\gamma}(E|\mathcal{H})$ not same as $P^{\tau}(E|\mathcal{H})$

How should we interpret probability?

- probability of a head occurring when a coin is flipped
- flip a coin N times (N is large, say a million)
- let us say coin comes up head h number of times
- probability of a head is $\frac{h}{N}$
- probability interpreted as the frequency of a *repeatable event*

Subjective View

- $P(E|\mathcal{H})$ interpreted as the belief of a person given that he/she knows the history $\mathcal{H}$ that E will occur
- interpret $P(E|\mathcal{H})$ as a *betting coefficient*
- how much a person is willing to bet that event E will happen in exchange of one dollar?
- we call this view *subjective probability*
- which view is better for software engineering?

Let us start

- X a random variable and $X = x$ denote the event that X realizes the value x
- $P(X = x|\mathcal{H})$ is abbreviated as $P_X(x|\mathcal{H})$
- $P_X(x|\mathcal{H}) > 0$ then X is said to have a *point mass* at x
- $P(X \leq x|\mathcal{H})$ is called the *distribution function* of X
- distribution function denoted by $F_X(x|\mathcal{H})$
- the derivative of $F_X(x|\mathcal{H})$ at x is called the *probability density function* and denoted by $f_X(x|\mathcal{H})$

Some questions

- let X be uniformly distributed between $[0, 1]$
- what is $F_X(x)$?
- what is $f_X(x)$?
- is $F_X(x)$ always smooth?
- if $F_X(x)$ jumps at x_1, what does it mean?

Multiple Random Variables

- interpret $F_{X_1,X_2}(x_1, x_2|\mathcal{H})$ as $P(X_1 \leq x_1$ and $X_2 \leq x_2|\mathcal{H})$
- $f_{X_1,X_2}(x_1, x_2|\mathcal{H})dx_1dx_2$ approximates
  - $P(x_1 \leq X_1 \leq x_1 + dx_1$ and $x_2 \leq X_2 \leq x_2 + dx_2)$

Example

- consider an unit square and
  X and Y be the two coordinates
- let $F_{X,Y}(x, y)$ be $x \cdot y$
- what is $f_{X,Y}(x, y)$?

Conditional Probabilities and Independence

- consider two random variables X_1 and X_2
- suppose you know the value of X_2
- this knowledge affects your judgement about X_1

Conditional Probabilities

- $P_{X_1|X_2}(x_1|x_2, \mathcal{H})$ is the probability that X_1 realizes the value x_1 *given that* X_2 has the value x_2
- this is called the *conditional probability of* X_1 *given* X_2
- $P(X_1 \leq x_1|X_2 = x_2, \mathcal{H})$ is abbreviated by $F_{X_1|X_2}(x_1|x_2, \mathcal{H})$
- known as the *conditional distribution* of X_1 given X_2

Independence

- suppose the following equation is true

$$P(X_1 = x_1 | X_2 = x_2, \mathcal{H}) \ = \ P(X_1 = x_1 | \mathcal{H})$$

- what does it mean?
- realization of X_2 does not affect the distribution of X_1
- X_1 is said to be *independent* of X_2
- are X and Y independent in our unit square example?

Independence

- if X_1 and X_2 are independent, then what is $P(X_1 = x_1 \text{ and } X_2 = x_2|\mathcal{H})$?
- suppose a software is developed by two teams A and B
- X_A first time software developed by A fails
- similarly for X_B

Independence

- analyst assesses $P(X_A \geq \tau | \mathcal{H})$ and $P(X_B \geq \tau | \mathcal{H})$ as p_A and p_B respectively
- what does it mean to say that X_A and X_B are independent?

Why independence?

- generally the independence assumption is not true
- code developed from same specification
- many experiments performed which refute the independence assumption
- independence assumption makes calculations easier
- generally a very idealistic assumption

Laws of Probability

- **Convexity:** For any event E

$$0 \leq P(E|\mathcal{H}) \leq 1$$

- **Additivity:** If both E_1 and E_2 cannot occur simultaneously (they are *mutually exclusive*), then

$$P(E_1 \text{ or } E_2|\mathcal{H}) \ = \ P(E_1|\mathcal{H}) + P(E_2|\mathcal{H})$$

- **Multiplicativity:**

$$P(E_1 \text{ and } E_2|\mathcal{H}) \ = \ P(E_1|E_2, \mathcal{H})P(E_2|\mathcal{H})$$

Generalizations

- consider n events $E_1, \cdots, E_n$ that are mutually exclusive

$$P(E_1 \text{ or } E_2 \text{ or } \cdots E_n|\mathcal{H}) = \sum_{i=1}^{n} P(E_i|\mathcal{H})$$

- multiplicative law takes the following

$$\begin{aligned} P(E_1 \text{ and } E_2 \text{ and } \cdots E_n|\mathcal{H}) &= P(E_1|E_2, \cdots, E_n, \mathcal{H}) \times \\ &\quad P(E_2|E_3, \cdots, E_n, \mathcal{H}) \times \cdots \\ &\quad \times P(E_n|\mathcal{H}) \end{aligned}$$

More equations

- suppose E_1 and E_2 are not mutually exclusive

$$P(E_1 \text{ or } E_2|\mathcal{H}) = P(E_1|\mathcal{H}) + P(E_2|\mathcal{H}) - P(E_1 \text{ and } E_2|\mathcal{H})$$

- if E_1 and E_2 are independent

$$P(E_1 \text{ or } E_2|\mathcal{H}) = P(E_1|\mathcal{H}) + P(E_2|\mathcal{H}) - P(E_1|\mathcal{H})P(E_2|\mathcal{H})$$

An Example

- consider a system made up of a hardware and software component
- E_H denote the event that the hardware experiences a fault within the next day
- E_S denote the event that the software experiences a fault within the next day
- the system fails if either hardware or software fail

An Example

- the system reliability is given by

$$P(E_H \text{ or } E_S|\mathcal{H})$$

- given by the following expression (suppressing the history)

$$P(E_H) + P(E_S) - P(E_H \text{ and } E_S)$$

- if E_H and E_S are independent, then the expression simplifies to

$$P(E_H) + P(E_S) - P(E_H)P(E_S)$$

Example Extended

- suppose the hardware has a backup system
- the probability that the hardware component fails is

$$P(E_H \text{ and } E_B)$$

- the probability given above evaluates to

$$P(E_H|E_B)P(E_B)$$

- the probability that the system will fail is

$$P((E_H \text{ and } E_B) \text{ or } E_S)$$

The Law of Total Probability

- suppose X_1 and X_2 are two discrete random variables
- $P(X_1 = x_1, X_2 = x_2)$ is their *joint probability*
- the marginal of X_1 alone is

$$P(X_1 = x_1) \; = \; \sum_{x_2} P(X_1 = x_1 | X_2 = x_2)$$

Bayes Law

- compute

$$P(X_1 = x_1 | X_2 = x_2) \quad = \quad \frac{P(X_1 = x_1, X_2 = x_2)}{\sum_{x_1} P(X_1 = x_1, X_2 = x_2)}$$

- applying the multiplicative rule we get

$$\frac{P(X_2 = x_2 | X_1 = x_1) P(X_1 = x_1)}{\sum_{x_1} P(X_2 = x_2 | X_1 = x_1)}$$