Statistics for Computing

Lecture 1A

Kevin O'Brien

Kevin.obrien@ul.ie

Dept. of Mathematics & Statistics,
University *of* Limerick

Autumn Semester 2013

About this module

- Lecture 1: Monday 16-17 in Room P1033
- Lecture 2: Wednesday 12-13 in Room FB028
- No classes on Bank Holiday Monday and during the open days.
- Tutorials - Tuesdays and Friday.
  (You may attend any tutorial you like.)
- Tutorials will take place from Week 2 to Week 13 inclusive.
- Review classes will take place during week 13 if requested.

Today's Class

- Description of module syllabus
- Details on the assessment
- Teaching materials
  - Lecture notes : Website (bit.ly/ma4413) and available on SULIS system.
  - `R` code, a statistical programming language, will also be included in some parts of the course.
  - My e-mail address is on the lecture slides.

Syllabus

The syllabus is made up of the following subject areas:

- Probability theory,
- Descriptive statistics and graphical methods,
- Probability distributions,
- Information theory and data compression,
- Confidence intervals and hypothesis testing.

Learning Outcomes

On successful completion of this module, students should be able to:

1. Apply probability theory to problem solving.
2. Employ the concepts of random variables and probability distributions to problem solving.
3. Apply information theory to solve problems in data compression and transmission.
4. Analyse rates and proportions.
5. Perform hypothesis tests for a variety of statistical problems.

Module Assessment

The grading of the module is as follows:

- End of semester examination (70%).
- Two mid-term examinations (15% each).
- The mid-terms are provisionally scheduled for Week 5 and Week 9.
- I will re-arrange the date if needs be.
- Sample papers and past papers will be made available through the SULIS system or by email.

Probability

Introduction to Probability *Probability is the very guide of life* - Cicero

1. Probability
2. Random Experiment
3. Outcome
4. Sample Space
5. Event
6. Conditional Probability
7. Independent Events
8. Mutually Exclusive Events
9. Addition Rule

Probability

- Probability theory is the mathematical study of randomness. A probability model of a random experiment is defined by assigning probabilities to all the different outcomes.
- Probability is a numerical measure of the likelihood that an event will occur. Thus, probabilities can be used as measures of degree of uncertainty associated with outcomes of an experiment. Probability values are always assigned on a scale from 0 to 1.
- A probability of 0 means that the event is impossible, while a probability near 0 means that it is highly unlikely to occur.
- Similarly an event with probability 1 is certain to occur, whereas an event with a probability near to 1 is very likely to occur.

Experiments and Outcomes

- In the study of probability any process of observation is referred to as an experiment.
- The results of an experiment (or other situation involving uncertainty) are called the outcomes of the experiment.
- An experiment is called a random experiment if the outcome can not be predicted.
- Typical examples of a random experiment are
    - a role of a die,
    - a toss of a coin,
    - drawing a card from a deck.

  If the experiment is yet to be performed we refer to possible outcomes or possibilities for short. If the experiment has been performed, we refer to realized outcomes or realizations.

Sample Spaces and Events

- The set of all possible outcomes of a probability experiment is called a ***sample space***, which is usually denoted by ***S***.
- The sample space is an exhaustive list of all the possible outcomes of an experiment. We call individual elements of this list ***sample points***.
- Each possible outcome is represented by one and only one sample point in the sample space.

Sample Spaces: Examples

For each of the following experiments, write out the sample space.

- Experiment: Rolling a die once
    - Sample space $S = \{1,2,3,4,5,6\}$
- Experiment: Tossing a coin
    - Sample space $S = \{Heads, Tails\}$
- Experiment: Measuring a randomly selected persons height (cms)
    - Sample space $S =$ The set of all possible real numbers.

Events

- An event is a specific outcome, or any collection of outcomes of an experiment.
- Formally, any subset of the sample space is an event.
- Any event which consists of a single outcome in the sample space is called an ***elementary*** or ***simple event***.
- Events which consist of more than one outcome are called ***compound events.***
- For example, an elementary event associated with the die example could be the “die shows 3”.
- An compound event associated with the die example could be the “die shows an even number”.

The Complement Event

- The complement of an event A is the set of all outcomes in the sample space that are not included in the outcomes of event A.
- We call the complement event of A as A^c.
- The complement event of a die throw resulting in an even number is the die throwing an odd number.
- Question: if there is a 40% chance of a randomly selected student being male, what is the probability of the selected student being female?

Set Theory : Union and Intersection

Set theory is used to represent relationships among events.

Union of two events:
The union of events A and B is the event containing all the sample points belonging to A or B or both. This is denoted $A \cup B$, (pronounce as “A union B”).

Intersection of two events:
The intersection of events A and B is the event containing all the sample points common to both A and B. This is denoted $A \cap B$, (pronounce as “A intersection B”).

More Set Theory

In general, if A and B are two events in the sample space S, then

- $A \subseteq B$ (A is a subset of B) = 'if A occurs, so does B
- $\varnothing$ (the empty set) = an impossible event
- S (the sample space) = an event that is certain to occur

Examples of Events

Consider the experiment of rolling a die once. From before, the sample space is given as $S = \{1,2,3,4,5,6\}$. The following are examples of possible events.

- A = score < 4 = $\{1,2,3\}$.
- B = 'score is even' = $\{2,4,6\}$.
- C = 'score is 7' = 0
- $A \cup B$ = 'the score is < 4 or even or both' = $\{1,2,3,4,6\}$
- $A \cap B$ = 'the score is < 4 and even = $\{2\}$
- A^c ='event A does not occur' = $\{4,5,6\}$

Probability

If there are n possible outcomes to an experiment, and m ways in which event A can happen, then the probability of event A (which we write as P(A)) is

$$P(A) = \frac{m}{n}$$

The probability of the event A may be interpreted as the proportion of times that event A will occur if we repeat the random experiment an infinite number of times.

Rules:

1. $0 \leq P(A) \leq 1$: the probability of any event lies between 0 and 1 inclusive.
2. $P(S) = 1$: the probability of the sample space is always equal to 1.
3. $P(A^c) = 1 - P(A)$: how to compute the probability of the complement.

Conditional Probability

Suppose B is an event in a sample space S with $P(B) > 0$. The probability that an event A occurs once Bhas occurred or, specifically, the conditional probability of A given B (written $P(A|B)$), is defined as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- This can be expressed as a multiplication theorem

$$P(A \cap B) = P(A|B) \times P(B)$$

- The symbol $|$ is a vertical line and does not imply division.
- Also $P(A|B)$ is not the same as $P(B|A)$.

Remark: The Prosecutor's Fallacy , with reference to the O.J. Simpson trial.

Independent Events

Events A and B in a probability space S are said to be independent if the occurrence of one of them does not influence the occurrence of the other.

More specifically, B is independent of A if $P(B)$ is the same as $P(B|A)$. Now substituting $P(B)$ for $P(B|A)$ in the multiplication theorem from the previous slide yields.

$$P(A \cap B) = P(A) \times P(B)$$

We formally use the above equation as our definition of independence.

Mutually Exclusive Events

- Two events are mutually exclusive (or disjoint) if it is impossible for them to occur together.
- Formally, two events A and B are mutually exclusive if and only if $A \cap B = \varnothing$

Consider our die example

- Event A = 'observe an odd number' = $\{1,3,5\}$
- Event B = 'observe an even number' = $\{2,4,6\}$
- $A \cap B = \varnothing$ (i.e. the empty set), so A and B are mutually exclusive.

Addition Rule

The addition rule is a result used to determine the probability that event A or event B occurs or both occur. The result is often written as follows, using set notation:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- $P(A)$ = probability that event A occurs.
- $P(B)$ = probability that event B occurs.
- $P(A \cup B)$ = probability that either event A or event B occurs, or both occur.
- $P(A \cap B)$ = probability that event A and event B both occur.

Remark: $P(A \cap B)$ is subtracted to prevent the relevant outcomes being counted twice.

Addition Rule (Continued)

For mutually exclusive events, that is events which cannot occur together: $P(A \cap B) = 0$. The addition rule therefore reduces to

$$P(A \cup B) = P(A) + P(B)$$

Addition Rule: Worked Example

Suppose we wish to find the probability of drawing either a Queen or a Heart in a single draw from a pack of 52 playing cards. We define the events Q = 'draw a queen' and H = 'draw a heart'.

- $P(Q)$ probability that a random selected card is a Queen
- $P(H)$ probability that a randomly selected card is a Heart.
- $P(Q \cap H)$ probability that a randomly selected card is the Queen of Hearts.
- $P(Q \cup H)$ probability that a randomly selected card is a Queen or a Heart.

Solution

- Since there are 4 Queens in the pack and 13 Hearts, so the $P(Q) = 4/52$ and $P(H) = 13/52$ respectively.
- The probability of selecting the Queen of Hearts is $P(Q \cap H) = 1/52$.
- We use the addition rule to find $P(Q \cup H)$:

$$P(Q \cup H) = (4/52) + (13/52) - (1/52) = 16/52$$

- So, the probability of drawing either a queen or a heart is $16/52 (= 4/13)$.

More on probability

For this lecture and the next.

1. Contingency Tables
2. Conditional Probability: Worked Examples
3. Joint Probability Tables
4. The Multiplication Rule
5. Law of Total Probability
6. Bayes' Theorem
7. Exam standard Probability Question
8. Random Variables

Contingency Tables

Suppose there are 100 students in a first year college intake.

- 44 are male and are studying computer science,
- 18 are male and studying engineering
- 16 are female and studying computer science,
- 22 are female and studying engineering.

We assign the names M, F, C and E to the events that a student, randomly selected from this group, is male, female, studying computer science, and studying engineering respectively.

Contingency Tables

The most effective way to handle this data is to draw up a table. We call this a ***contingency table***. A contingency table is a table in which all possible events (or outcomes) for one variable are listed as row headings, all possible events for a second variable are listed as column headings, and the value entered in each cell of the table is the frequency of each joint occurrence.

	C	E	Total
M	44	18	62
F	16	22	38
Total	60	40	100

Contingency Tables

It is now easy to deduce the probabilities of the respective events, by looking at the totals for each row and column.

- P(C) = 60/100 = 0.60
- P(E) = 40/100 = 0.40
- P(M) = 62/100 = 0.62
- P(F) = 38/100 = 0.38

Remark:

The information we were originally given can also be expressed as:

- $P(C \cap M) = 44/100 = 0.44$
- $P(C \cap F) = 16/100 = 0.16$
- $P(E \cap M) = 18/100 = 0.18$
- $P(E \cap F) = 22/100 = 0.22$

Joint Probability Tables

A ***joint probability table*** is similar to a contingency table, but for that the value entered in each cell of the table is the probability of each joint occurrence. Often, the probabilities in such a table are based on observed frequencies of occurrence for the various joint events.

	C	E	Total
M	0.44	0.18	0.62
F	0.16	0.22	0.38
Total	0.60	0.40	1.00

Marginal Probabilities

- In the context of joint probability tables, a ***marginal probability*** is so named because it is a marginal total of a row or a column.
- Whereas the probability values in the cells of the table are probabilities of joint occurrence, the marginal probabilities are the simple (i.e. unconditional) probabilities of particular events.
- From the first year intake example, the marginal probabilities are $P(C)$, $P(E)$, $P(M)$ and $P(F)$ respectively.

Conditional Probabilities : Example 1

Recall the definition of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Using this formula, compute the following:

1. $P(C|M)$: Probability that a student is a computer science student, given that he is male.
2. $P(E|M)$: Probability that a student studies engineering, given that he is male.
3. $P(F|E)$: Probability that a student is female, given that she studies engineering.
4. $P(E|F)$: Probability that a student studies engineering, given that she is female.

Refer back to the contingency table to appraise your results.

Conditional Probabilities : Example 1

Part 1) Probability that a student is a computer science student, given that he is male.

$$P(C|M) = \frac{P(C \cap M)}{P(M)} = \frac{0.44}{0.62} = 0.71$$

Part 2) Probability that a student studies engineering, given that he is male.

$$P(E|M) = \frac{P(E \cap M)}{P(M)} = \frac{0.18}{0.62} = 0.29$$

Conditional Probabilities : Example 1

Part 3) Probability that a student is female, given that she studies engineering.

$$P(F|E) = \frac{P(F \cap E)}{P(E)} = \frac{0.22}{0.40} = 0.55$$

Part 4) Probability that a student studies engineering, given that she is female.

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{0.22}{0.38} = 0.58$$

Remark: $P(E \cap F)$ is the same as $P(F \cap E)$.