Statistics for Computing

Lecture 1B

Kevin O'Brien

Kevin.obrien@ul.ie

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

Autumn Semester 2013

(Brought over from last lecture)

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. Sampling (Samples and Populations)
9. Sampling with and without Replacement

(Later in Class : A look at Descriptive Statistic)

Multiplication Rule

The multiplication rule is a result used to determine the probability that two events, A and B, both occur. The multiplication rule follows from the definition of conditional probability.

The result is often written as follows, using set notation:

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

Recall that for independent events, that is events which have no influence on one another, the rule simplifies to:

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

Multiplication Rule

From the first year intake example, check that

$$P(E|F) \times P(F) = P(F|E) \times P(E)$$

- $P(E|F) \times P(F) = 0.58 \times 0.38 = 0.22$
- $P(F|E) \times P(E) = 0.55 \times 0.40 = 0.22$

Law of Total Probability

The law of total probability is a fundamental rule relating marginal probabilities to conditional probabilities. The result is often written as follows, using set notation:

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

where $P(A \cap B^c)$ is probability that event A occurs and B does not.

Using the multiplication rule, this can be expressed as

$$P(A) = P(A|B) \times P(B) + P(A|B^c) \times P(B^c)$$

Law of Total Probability

From the first year intake example , check that

$$P(E) = P(E \cap M) + P(E \cap F)$$

with $P(E) = 0.40$, $P(E \cap M) = 0.18$ and $P(E \cap F) = 0.22$

$$0.40 = 0.18 + 0.22$$

Remark: M and F are complement events.

Bayes' Theorem

Bayes' Theorem is a result that allows new information to be used to update the conditional probability of an event.

Recall the definition of conditional probability:

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

Using the multiplication rule, gives Bayes' Theorem in its simplest form:

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

Probability: Worked Example

An electronics assembly subcontractor receives resistors from two suppliers: Deltatech provides 70% of the subcontractors's resistors while another company, Echelon, supplies the remainder.
1% of the resistors provided by Deltatech fail the quality control test, while 2% of the resistors from Echelon also fail the quality control test.

1. What is the probability that a resistor will fail the quality control test?
2. What is the probability that a resistor that fails the quality control test was supplied by Echelon?

Probability: Worked Example

Firstly, let's assign names to each event.

- D : a randomly chosen resistor comes from Deltatech.
- E : a randomly chosen resistor comes from Echelon.
- F : a randomly chosen resistor fails the quality control test.
- P : a randomly chosen resistor passes the quality control test.

We are given (or can deduce) the following probabilities:

- $P(D) = 0.70$,
- $P(E) = 0.30$.

Probability: Worked Example

We are given two more important pieces of information:

- The probability that a randomly chosen resistor fails the quality control test, given that it comes from Deltatech: $P(F|D) = 0.01$.
- The probability that a randomly chosen resistor fails the quality control test, given that it comes from Echelon: $P(F|E) = 0.02$.

Probability: Worked Example

The first question asks us to compute the probability that a randomly chosen resistor fails the quality control test. i.e. $P(F)$.

All resistors come from either Deltatech or Echelon. So, using the ***law of total probability***, we can express $P(F)$ as follows:

$$P(F) = P(F \cap D) + P(F \cap E)$$

Probability: Worked Example

Using the ***multiplication rule*** i.e. $P(A \cap B) = P(A|B) \times P(B)$, we can re-express the formula as follows

$$P(F) = P(F|D) \times P(D) + P(F|E) \times P(E)$$

We have all the necessary probabilities to solve this.

$$P(F) = 0.01 \times 0.70 + 0.02 \times 0.30 = 0.007 + 0.006 = 0.013$$

Probability: Worked Example

- The second question asks us to compute probability that a resistor that fails the quality control test was supplied by Echelon.
- In other words; of the resistors that did fail the quality test only, what is the probability that a randomly selected resistor was supplied by Echelon?
- We can express this mathematically as $P(E|F)$.
- We can use ***Bayes' theorem*** to compute the answer.

Probability: Worked Example

Recall Bayes' theorem

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

$$P(E|F) = \frac{P(F|E) \times P(E)}{P(F)} = \frac{0.02 \times 0.30}{0.013} = 0.46$$

Sampling

The major use of statistics is to use information from a ***sample*** to infer something about a ***population***.

- A ***population*** is a collection of data whose properties are analyzed. The population is the complete collection to be studied, it contains all subjects of interest.
- A ***sample*** is a part of the population of interest, a sub-collection selected from a population.
- A ***parameter*** is a numerical measurement that describes a characteristic of a population, while a ***sample statistic*** is a numerical measurement that describes a characteristic of a sample.
- In general, we will use a statistic to infer something about a parameter.

Sampling without replacement

- Sampling is said to be “without replacement” when a unit is selected at random from the population and it is not returned to the main lot.
- The first unit is selected out of a population of size N and the second unit is selected out of the remaining population of $N-1$ units and so on.
- For example, if you draw one card out of a deck of 52, there are only 51 cards left to draw from if you are selecting a second card.

Sampling without replacement

A lot of 100 semiconductor chips contains 20 that are defective. Two chips are selected at random, without replacement from the lot.

- What is the probability that the first one is defective?
  (Answer : 20/100 , i.e 0.20)
- What is the probability that the second one is defective given that the first one was defective?
  (Answer: 19/99)
- What is the probability that the second one is defective given that the first one was not defective?
  (Answer: 20/99)

Sampling With Replacement

Sampling is called "with replacement" when a unit selected at random from the population is returned to the population and then a second element is selected at random. Whenever a unit is selected, the population contains all the same units.

- What is the probability of guessing a PIN number for an ATM card at the first attempt.
- Importantly a digit can be used twice, or more, in PIN codes.
- For example 1337 is a valid pin number, where 3 appears twice.
- We have a one-in-ten chance of picking the first digit correctly, a one-in-ten chance of the guessing the second, and so on.
- All of these events are independent, so the probability of guess the correct PIN is $0.1 \times 0.1 \times 0.1 \times 0.1 = 0.0001$

Descriptive Statistics

We will digress from Probabilility for a while, and look at **Descriptive Statistics**.

- Sample Mean
- Sample Median
- Measures of dispersion

Descriptive Statistics

- Measures of Centrality
  - Mean
  - Median
- Measures of Dispersion
  - Range
  - Variance
  - Standard Deviation

Measures of Centrality

- Measures of centrality give one representative number for the location of the centre of the distribution of data.
- The most common measures are the ***mean*** and the ***median*** .
- We must make a distinction between a sample mean and a population mean: The sample mean is simply the average of all the items in a sample.
- The population mean (often represented by the Greek letter μ) is simply the average of all the items in a population.
- Because a population is usually very large, the population mean is usually an unknown constant.
- We will return to the matter of population means in due course. For now, we will look at sample means.

Sample Mean

- The sample mean is an estimator available for estimating the population mean . It is a measure of location, commonly called the average, often denoted $\bar{x}$, where x is the data set.
- Its value depends equally on all of the data which may include outliers. It may not appear representative of the central region for skewed data sets.
- It is especially useful as being representative of the whole sample for use in subsequent calculations.
- The sample mean of a data set is defined as :

$$\bar{x} = \frac{\sum x_i}{n}$$

- $\sum x_i$ is the summation of al the elements of x, and n is the sample size.

Computing the sample mean

Suppose we roll a die 8 times and get the following scores:
$x = \{5,2,1,6,3,5,3,1\}$

What is the sample mean of the scores $\bar{x}$?

$$\bar{x} = \frac{5+2+1+6+3+5+3+1}{8} = \frac{26}{8} = 3.25$$

Using R to compute mean (and median)

When implementing this in R, we would use the following code

```
> # create the "vector" x with the required values
> x=c(5, 2, 1, 6, 3, 5, 3, 1)
>
> mean(x)
[1] 3.25
>
> # See next slides first.
> sort(x)
[1] 1 1 2 3 3 5 5 6
> median(x)
[1] 3
```

Median

- The other commonly used measure of centrality is the median.
- The median is the value halfway through the ordered data set, below and above which there lies an equal number of data values.
- For an odd sized data set, the median is the middle element of the **ordered** data set.
- For an even sized data set, the median is the average of the middle pair of elements of an **ordered** data set.
- It is generally a good descriptive measure of the location which works well for ***skewed data***, or data with ***outliers***.
- For later, the median is the 0.5 quantile, and the second quartile Q_2.

Computing the median

Example:

With an odd number of data values, for example nine, we have:

- Data : $\{96, 48, 27, 72, 39, 70, 7, 68, 99\}$
- Ordered Data : $\{7, 27, 39, 48, 68, 70, 72, 96, 99\}$
- Median : 68, leaving four values below and four values above

With an even number of data values, for example 8, we have:

- Data : $\{96, 48, 27, 72, 39, 70, 7, 68\}$
- Ordered Data : $\{7, 27, 39, 48, 68, 70, 72, 96\}$
- Median : Halfway between the two 'middle' data points - in this case halfway between 48 and 68, and so the median is 58

Using R to compute mean (and median)

When implementing this in R, we would use the following code

```
> x1=c(96, 48, 27, 72, 39, 70, 7, 68, 99 )
> sort(x1)
[1]  7 27 39 48 68 70 72 96 99
> median(x1)
[1] 68
>
> x2=c(96, 48 ,27 ,72, 39, 70, 7, 68)
> sort(x2)
[1]  7 27 39 48 68 70 72 96
> median(x2)
[1] 58
```

Dispersion

- The data values in a sample are not all the same. This variation between values is called ***dispersion***.
- When the dispersion is large, the values are widely scattered; when it is small they are tightly clustered.
- There are several measures of dispersion, the most common being the variance and standard deviation. These measures indicate to what degree the individual observations of a data set are dispersed or 'spread out' around their mean.
- In engineering and science, high precision is associated with low dispersion.

Range

- The range of a sample (or a data set) is a measure of the spread or the dispersion of the observations.
- It is the difference between the largest and the smallest observed value of some quantitative characteristic and is very easy to calculate.
- A great deal of information is ignored when computing the range since only the largest and the smallest data values are considered; the remaining data are ignored.
- The range value of a data set is greatly influenced by the presence of just one unusually large or small value in the sample (outlier).

Example

The range of $\{65, 73, 89, 56, 73, 52, 47\}$ is $89 - 47 = 42$.

Introducing Variance

Consider the three data sets X, Y and Z

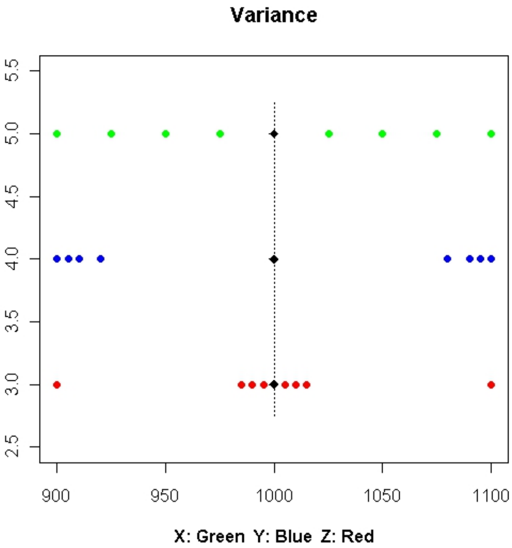
- $X = \{900, 925, 950, 975, 1025, 1050, 1075, 1100\}$
- $Y = \{900, 905, 910, 920, 1080, 1090, 1095, 1100\}$
- $Z = \{900, 985, 990, 995, 1005, 1010, 1015, 1100\}$

For each of the data sets, the following statements can be verified

- The mean of each data set is 1000
- There are 8 elements in each data set
- The minima and maxima are 900 and 1100 for each set
- The range is 200.

From the plot on the next slide, notice how different the three data sets are in terms of dispersion around the mean value.

Introducing Variance



Variance

- The (population) variance of a random variable is a non-negative number which gives an idea of how widely spread the values are likely to be; the larger the variance, the more scattered the observations on average.
- Stating the variance gives an impression of how closely concentrated round the expected value the distribution is; it is a measure of the 'spread' of a distribution about its average value.
- We distinguish between population variance (denoted σ^2) and sample variance (denoted s^2). For now, we will look only at sample variance.

Sample Variance

- Sample variance is a measure of the spread of or dispersion within a set of sample data.
- The sample variance is the sum of the squared deviations from their mean divided by one less than the number of observations in the data set.
- For example, for n observations $x_1, x_2, x_3, \ldots, x_n$ with sample mean $\bar{x}$, the sample variance is given by

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

Sample Standard Deviation

- Standard deviation is the square root of variance
- Standard deviation is commonly used in preference to variance because it is denominated in the same units as the mean.
- For example, if dealing with time units, we could have a variance of something like 25 *square minutes* , whereas the equivalent standard deviation is 5 minutes.
- Population standard deviation is denoted σ.
- Sample standard deviation is denoted s.

Using R

Using R to compute standard deviation and variance for these data sets.

```
> X=c(900,925,950,975,1025,1050,1075,1100)
> Y=c(900,905,910,920,1080,1090,1095,1100)
> Z=c(900,985,990,995,1005,1010,1015,1100)
>
> sd(X);sd(Y);sd(Z)
[1] 73.19251
[1] 97.87018
[1] 54.37962
>
>var(X);var(Y);var(Z)
[1] 5357.143
[1] 9578.571
[1] 2957.143
```