Statistics for Computing

MA4413 Lecture 2A

Kevin O'Brien

Kevin.obrien@ul.ie

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

Autumn Semester 2012

Today's Class

- Sampling without replacement.
- Factorials
- Permutations
- Combinations
- Sample mean
- Median
- Measures of dispersion

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 $0.1 \times 0.1 \times 0.1 \times 0.1 = 0.0001$

Factorials Numbers

A factorial is a positive whole number, based on a number n , and which is written as "n!". The factorial $n!$ is defined as follows:

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1$$

Remark $n! = n \times (n-1)!$

Example:

- $3! = 3 \times 2 \times 1 = 6$
- $4! = 4 \times 3! = 4 \times 3 \times 2 \times 1 = 24$

Remark $0! = 1$ not 0.

Permutations and Combinations

Often we are concerned with computing the number of ways of selecting and arranging groups of items.

- A ***combination*** describes the selection of items from a larger group of items.
- A ***permutation*** is a combination that is arranged in a particular way.

- Suppose we have items A,B,C and D to choose two items from.
- AB is one possible selection, BD is another. AB and BD are both combinations.
- More importantly, AB is one combination, for which there are two distinct permutations: AB and BA.

Combinations

Combinations: The number of ways of selecting k objects from n unique objects is:

$$^{n}C_{k} = \frac{n!}{k! \times (n-k)!}$$

In some texts, the notation for finding the number of possible combination is written

$$^{n}C_{k} = \binom{n}{k}$$

Example of Combinations

How many ways are there of selecting two items from possible 5?

$$^{5}C_{2}\left(\text{ also }\binom{5}{2}\right)=\frac{5!}{2!\times 3!}=\frac{5\times 4\times 3!}{2\times 1\times 3!}=10$$

Discuss how combinations can be used to compute the number of rugby matches for each group in the Rugby World Cup.

The Permutation Formula

The number of different permutations of r items from n unique items is written as ${}^{n}P_{k}$

$${}^{n}P_{k} = \frac{n!}{(n-k)!}$$

Permutations

Example: How many ways are there of arranging 3 different jobs, between 5 workers, where each worker can only do one job?

$$^{5}P_3 = \frac{5!}{(5-3)!} = \frac{5!}{2!} = 60$$

Example of Combinations

A committee of 4 must be chosen from 3 females and 4 males.

- In how many ways can the committee be chosen.
- In how many cans 2 males and 2 females be chosen.
- Compute the probability of a committee of 2 males and 2 females are chosen.
- Compute the probability of at least two females.

Example of Combinations

Part 1
We need to choose 4 people from 7:
This can be done in

$$ {}^{7}C_4 = \frac{7!}{4! \times 3!} = \frac{7 \times 6 \times 5 \times 4!}{4! \times 3!} = 35 \text{ ways.} $$

Part 2
With 4 men to choose from, 2 men can be selected in

$$ {}^{4}C_2 = \frac{4!}{2! \times 2!} = \frac{4 \times 3 \times 2!}{2! \times 2!} = 6 \text{ ways.} $$

Similarly 2 women can be selected from 3 in

$$ {}^{3}C_2 = \frac{3!}{2! \times 1!} = \frac{3 \times 2!}{2! \times 1!} = 3 \text{ ways.} $$

Using R

When implementing combination calculations in R, we use the `choose()` function.

```
> choose(5,0)
[1] 1
> choose(5,1)
[1] 5
> choose(5,2)
[1] 10
> choose(5,3)
[1] 10
> choose(5,4)
[1] 5
> choose(5,5)
[1] 1
```

Example of Combinations

Part 2

Thus a committee of 2 men and 2 women can be selected in $6 \times 3 = 18$ ways.

Part 3

The probability of two men and two women on a committee is

$$\frac{\text{Number of ways of selecting 2 men and 2 women}}{\text{Number of ways of selecting 4 from 7}} = \frac{18}{35}$$

Example of Combinations

Part 4

- The probability of at least two females is the probability of 2 females or 3 females being selected.
- We can use the addition rule, noting that these are two mutually exclusive events.
- From before we know that probability of 2 females being selected is 18/35.

Example of Combinations

Part 4

- We have to compute the number of ways of selecting 1 male from 4 (4 ways) and the number of ways of selecting three females from 2 (only 1 way)
- The probability of selecting three females is therefore $\frac{4\times 1}{35} = 4/35$
- So using the addition rule

$$Pr(\text{ at least 2 females }) = Pr(\text{ 2 females }) + Pr(\text{ 3 females })$$

$$Pr(\text{ at least 2 females }) = 18/35 + 4/35 = 22/35$$

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

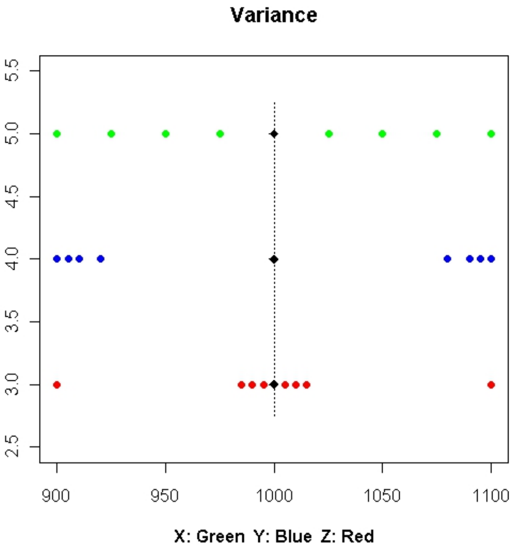
- $X = \{900, 925, 950, 975, 1025, 1050, 1075, 1100\}$
- $Y = \{900, 905, 910, 920, 1080, 1090, 1095, 1100\}$
- $Z = \{900, 985, 990, 995, 1005, 1010, 1015, 1100\}$

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
```