



# Bayesian Statistics: Scientific Data Gathering

Paul Hewson

**Overview:** This webfile is designed as a revision aid to key concepts around scientific data gathering. It is intended to supplement a formal encounter with a text book or a set of lectures. In particular, these notes are strongly based on Bolstad (2007). These notes are meant to be slightly interactive, mysterious green dots, squares and boxes appear which you can click on to answer questions and check solutions.

## 1. Design

According to Bolstad (2007) “Scientists gather data purposefully, in order to find answers to particular questions”

- Whilst collecting purposefully, as described previously *randomisation* is the key statistical contribution
  - Variability in data due to chance can be averaged out by increasing the sample size (we will formalise this concept next week when we consider the laws of large numbers - but look back at the computer work estimating  $\pi$  by Monte Carlo methods)
  - Variability due to other causes cannot be averaged out by increasing sample size

- We have two broad categories of statistical theory underpinning purposeful data collection:
  - Experimental design (and analysis)
  - Sample survey theory (and design)
- We shall also make reference throughout the course to more opportunistic methods of data collection

[Back](#)

Inference always depends on the assumed probability model being correct.

- When we don't use a randomised design to collect data, we increase the risk that any observed patterns we see are due to confounding variables
- Using a random survey or experiment means that the observations follow a probability model we followed in collecting the data (so we can be confident it is correct)
- When using non-randomly collected data, we have to assume that our model is correct, but we can never know this.



Back

◀ Doc

Doc ▶

## 2. Inference

We will consider three key terms:

- Population (summarised by parameters)
- Sample (summarised by statistics)
- Statistical inference



Back



## 2.1. Population

### Summarised by parameters

This is the entire group of (items, objects, subjects, people) about whom we are interested.

Thus a “population” could be:

- All humans in the world
- Every adult resident in Moscow in 2009
- All dairy cattle in England
- Every lightbulb manufactured by a particular factor in China
- Every child (under 14 years old) with Malaria

Note that for every object in this population, we are interested in knowing about one or more attributes. These could be features such as:

- Weight and height
- Income
- Daily milk yield
- Time until failure
- Presence of a particular immune reaction

Ultimately, we believe that the “attributes” vary amongst members of a population, but they vary in a pattern we can describe mathematically. Rather than carry around 6 billion data items for weight and height, we would prefer to use the mathematical description - the formula and one or two parameters to define it.

As noted before, it may be infeasible, expensive or impossible to collect data on an entire population. We therefore can not know the parameters.

[Back](#)[◀ Doc](#)[Doc ▶](#)

## 2.2. Sample

### Summarised by statistics

The sample is simply a subset of a population.

- We “draw” a sample from the population and collect information from the objects in that sample.
- Given total knowledge of the sample, we can calculate *summary statistics* that describe the sample - usually these include a measure of central tendency (mean, median or mode) and a measure of variability (variance, range, interquartile range and so on).

As we have all the data, we can know the statistics.



Back

◀ Doc

Doc ▶

## 2.3. Inference

This is the direction of thinking in statistics that most differs from mathematics. We seek to make inference.

- Based on the sample statistics, we hope to make a statement about possible values of the population.
- If the sample really is representative of the population, we can make good inference.
- If the sample is not representative, because of some sampling bias, our inference will be limited. The aim of random sampling is to avoid this bias

As mentioned before, we shall consider two general approaches to statistical inference on this course, classical (Frequentist) and Bayesian.

[Back](#)[Doc](#)[Doc](#)

- Frequentist inference: The population parameters are regarded as fixed but unknown constants
- Bayesian inference: The population parameters are regarded as random variables

Although the interpretation differs in detail, in either case, we end up making statements about a range of plausible values for the population parameters, based on a given sample.

[Back](#)[◀ Doc](#)[Doc ▶](#)

### 3. Sampling



Figure 1: Simple Random Sampling: individuals drawn at random from the entire population

## Quiz

1. Consider the following data on the effectiveness of Insecticide<sup>1</sup>  
 $X = 10, 7, 20, 14, 14, 12, 10, 23, 17, 20, 14, 12$   
Regarding this as your “population”, estimate the arithmetic mean  
(a) 14.0      (b) 14.5      (c) 15.0      (d) 15.5
2. Using the following Uniform random numbers, take a sample of size 6 and estimate the sample mean:  
 $U = 0.86 \ 0.09 \ 0.66 \ 0.25 \ 0.02 \ 0.30 \ 0.68 \ 0.89 \ 0.02 \ 0.62 \ 0.60 \ 0.30$   
(match the random numbers  $U$  with actual numbers  $X$ , and select the values of  $X$  that correspond to the 6 smallest numbers of  $U$  (these will be  $U \leq 0.3$ ). For example, the second entry is  $U$  is 0.09 ( $0.09 \leq 0.3$ ), and so we draw the corresponding value 7 (the second entry in  $X$ ) into our sample.  
(a) 12.2      (b) 12.8      (c) 13.4      (d) 14

---

<sup>1</sup>taken from Beall, G., (1942) “The Transformation of data from entomological field experiments”, *Biometrika*, **29**: 243-262



Back

◀ Doc

Doc ▶

### 3.1. Stratified Random Sampling

There are various reasons for using stratified random sampling:

- Administrative convenience: if we are estimating heights of school children it might be convenient to visit each class in turn.
- It may be more efficient to sample each strata separately for technical reasons, such as the variance within strata being smaller than the overall variance



Back

◀ Doc

Doc ▶

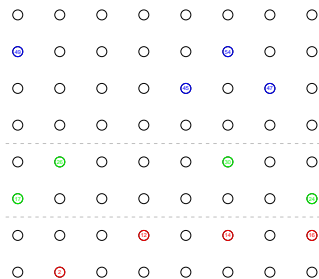


Figure 2: Samples taken at random from within strata

Essentially, we operate separate, independent, random sampling frames within each strata. If this is the case, we need to use different methods

to estimate the summary statistics. If we have  $K$  strata, each of size  $n_k$  where the strata are indexed  $k = 1, \dots, K$ , we calculate the mean for each individual strata  $k$  as before:

$$\bar{x}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_i$$

But we now need to combine the means for each of these strata, taking into account the different numbers in each strata. We use the following weighted sum:

$$\bar{x}_{str} = \sum_{k=1}^K W_k \bar{x}_k$$

where the weights,  $W_k$ , are given by:

$$W_k = \frac{n_k}{N}$$

with  $N$  denoting the total population size.



**Quiz** Consider further data on the effectiveness of Insecticide taken from groups A, B and Other ( $k = 1, 2, 3$  respectively):

$$X_1 (17, 7, 20, 14); n_1 = 12$$

$$X_2 (13, 17, 11, 16); n_2 = 12$$

$$X_3 (1, 5, 10, 3); n_3 = 48$$

1. What is the sample mean for each of the three strata ( $\bar{x}_1, \bar{x}_2, \bar{x}_3$ )  
(a) 14.25, 14.75, 4.00                      (b) 14.25, 14.50, 4.25  
(c) 14.50, 14.25, 4.75                      (d) 14.50, 14.50, 4.50
2. And what is the overall stratified mean  
(a) 7.66                      (b) 7.96                      (c) 11.17                      (d) 33.5



Back



Doc



Doc

## 3.2. Cluster sampling

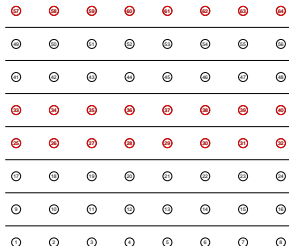


Figure 3: Clusters are randomly selected from a population

### 3.3. Final Points

- Non-sampling errors: for example there are a range of other issues such as response bias (the people who respond to a survey are different to those who do not respond)
- Randomised response methods: for very sensitive questions we can ask a choice of two questions - the sensitive question and an alternative (for which we know the population values already). Respondents toss a coin and answer either of the two questions. When analysing the results we assume 50% have answered the “known” question, and can infer the answer to the sensitive question

## 4. Observational Studies and Designed Experiments

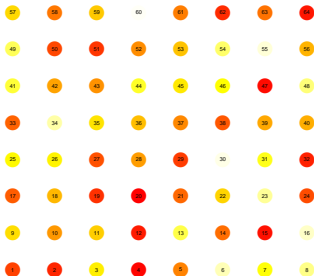


Figure 4: A fictional study group. Shading/coloring denotes potential confounding variable

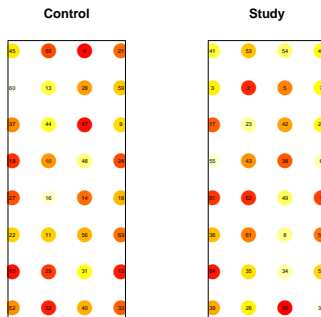


Figure 5: Study group randomly allocated to control/treatment groups.



Back



The previous two figures attempt to illustrate the random allocation of a study group of 64 individuals to two study groups. The levels of the potential confounding variable are also randomly spread between the two groups.

- Experimental study: one in which we control the allocation of individuals to “treatment” groups (and may also control the level of treatment in those groups)
- Observational study: one in which we observe what happens in a group of individuals who “choose” to take a treatment, and compare that with individuals who do not “choose” to take a treatment

What are the implications in terms of inference of using an observational study rather than a randomly controlled experiment?

## 4.1. Blocked designs

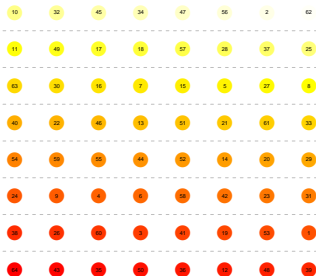


Figure 6: Study group “sorted” into 8 blocks with respect to potential confounding variable



Back

◀ Doc

Doc ▶

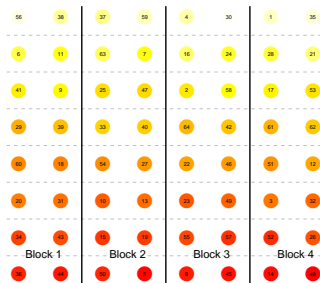


Figure 7: Within each block, subjects are randomly allocation to one of four treatment groups

The nomenclature here comes from agricultural field trials. The con-

founding variable may be parts of the field (north facing, south facing, hilly, well drained etc.). Treatments will be applied to crops, based on random allocation within each block.

[Back](#)[◀ Doc](#)[Doc ▶](#)

## 5. Summary

Whilst we haven't covered survey methods or experimental design in any kind of detail, we should now be very clear about the following terms:

- Population and Parameter
- Sample and Statistic
- Random sampling, and variants thereof such as stratified random sampling and random cluster sampling
- The difference between observational and experimental studies, as well as variants on random allocation (such as blocking)

There's a quick quiz to see whether we've covered this material:

You need to click the **Begin Quiz** before you start, and the **End Quiz** box when you finish (you will then be told the correct answers).

1. Any characteristic of a population distribution may properly be referred to as a
  - (a) standard deviation.
  - (b) raw score.
  - (c) standard score.
  - (d) standard error.
  - (e) parameter
2. A factor that is varied by an experimenter in order to assess its effect is known as a:
  - (a) dependent variable
  - (b) independent variable
  - (c) control variable
  - (d) none of the above

3. Characteristics of a population are called . . . . ., while those of a sample are termed . . . . .
- (a) statistics; measures                      (b) statistics; parameters  
(c) parameters; statistics                    (d) statistics; variables
4. A population is:
- (a) a number or measurement collected as a result of observation
- (b) a subset of a population
- (c) a characteristic of a population which is measurable
- (d) a complete set of individuals, objects, or measurements having some common observable characteristics
- (e) none of these



5. Consider the average reading achievement test score of the currently enrolled students in Mallory Towers Primary School. The set of test scores for Miss Brody's class comprise
- (a) an element.
  - (b) a sample.
  - (c) a statistic.
  - (d) a population.
6. Now consider the average score for *all* students in Mallory Towers School. This is a
- (a) sample.
  - (b) statistic.
  - (c) parameter.
  - (d) variable.



Back

◀ Doc

Doc ▶

7. An experiment is conducted to determine if the use of certain specified amounts of a drug will increase the IQ scores differentially for high and low anxious students in the fifth grade.

In this experiment, IQ serves as:

- (a) a primary independent variable
- (b) a moderator variable
- (c) a dependent variable
- (d) a control variable
- (e) an intervening variable.



Back

◀ Doc

Doc ▶

8. An experiment is conducted to determine if the use of certain specified amounts of a drug will increase the IQ scores differentially for high and low anxious students in the fifth grade.

In this experiment, the drug serves as:

- (a) a primary independent variable
- (b) a moderator variable
- (c) a dependent variable
- (d) a control variable
- (e) an intervening variable.



Back



9. Male students are assigned randomly to either a rote learning (memorization) treatment or to a discovery learning treatment. At the end of the experiment, students are tested for their ability to answer questions on an achievement test. The results indicate that fast learners in the discovery treatment do better than the slow learners in this treatment, but there is no difference in performance between the two types of learners in the rote treatment.

In this experiment, the achievement test serves as:

- (a) a primary independent variable
- (b) a moderator variable
- (c) a dependent variable
- (d) a control variable
- (e) an intervening variable

Points:

## Solutions to Quizzes

**Solution to Quiz:** The mean is given by

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n x_i \\ = & \frac{1}{12} 10 + 7 + 20 + 14 + 14 + 12 + 10 + 23 + 17 + 20 + 14 + 12 \\ = & \frac{174}{12} \end{aligned}$$



**Solution to Quiz:**

We use the Uniform random numbers (0.3 and below) to choose six values of the data at random:

10	7	20	14	14	12	10	23	17	20	14	13
0.86	0.09	0.66	0.25	0.02	0.30	0.68	0.89	0.02	0.62	0.60	0.30

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n x_i \\ &= \frac{1}{16} 7 + 14 + 14 + 12 + 17 + 13 \\ &= \frac{77}{6} \end{aligned}$$



**Solution to Quiz:** We have firstly that:

$$W_1 = \frac{12}{12+12+48} = 0.166,$$

$$W_2 = \frac{12}{72} = 0.166$$

$$\text{and } W_3 = \frac{48}{72} = 0.667$$

So

$$x_{str} = 0.166 \times 14.5 + 0.166 \times 14.25 + 0.667 \times 4.75 = 7.96$$



**Solution to Quiz:** This is a sample - a summary of these (such as a mean) would be a statistic. What kind of sampling scheme are we using here? ■

**Solution to Quiz:** Do note that we have defined (previous question) the school as being our entire population. ■