

S3

Will Bevington

Callum O'Brien

Alex Pace

October 8, 2015

Contents

1	Combining Random Variables	2
2	Sampling Frames	2
2.1	Random Sampling	2
2.1.1	Random Number Sample	2
2.1.2	Lottery Sample	2
2.1.3	Positives	3
2.1.4	Negatives	3
2.2	Systematic Sampling	3
2.2.1	Positives	3
2.2.2	Negatives	3
2.3	Stratified Sampling	3
2.3.1	Positives	3
2.3.2	Negatives	3
2.4	Quota Sampling	4
2.4.1	Positives	4
2.4.2	Negatives	4
3	Types of Data	4
3.1	Primary Data	4
3.1.1	Positives	4
3.1.2	Negatives	4
3.2	Secondary Data	4
3.2.1	Positives	4
3.2.2	Negatives	5
4	Estimating Population Parameters using a Sample	5

1 Combining Random Variables

Let X and Y be two independent random variables with means $E(X)$ and $E(Y)$ respectively, and variances $\text{Var}(X)$ and $\text{Var}(Y)$ respectively;

$$E(X \pm Y) = (EX) \pm E(Y)$$

$$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$$

$$E(aX \pm b) = aE(X) \pm b$$

$$\text{Var}(aX \pm b) = a^2\text{Var}(X) + b^2$$

The latter two formulae should be recalled from S1. We can combine these to acquire:

$$E(aX \pm bY) = aE(X) \pm bE(Y)$$

$$\text{Var}(aX \pm bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$$

Additionally, the combination of two independent normal distributions is also a normal distribution;

$$X \sim N(\mu_x, \sigma_x^2)$$

$$Y \sim N(\mu_y, \sigma_y^2)$$

$$(aX \pm bY) \sim N(a\mu_x \pm b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2)$$

2 Sampling Frames

Population The whole set of items that are of interest.

Census Observes or measures every member of a population.

Sample Survey A selection of observations taken from a subset of the population which is used to find out information about the population as a whole.

Random Sample A sample in which every possible sample of size n has an equal chance of being selected.

Sampling Frame A list identifying every single sampling unit that could be included in the sample

2.1 Random Sampling

2.1.1 Random Number Sample

Give each sampling unit in the sampling frame a number and use a random number generator or random number tables to select required number of sampling units.

2.1.2 Lottery Sample

Put the sampling units from the sampling frame into a "hat" and select randomly without replacing.

2.1.3 Positives

- Random and free from bias
- Easy to carry out

2.1.4 Negatives

- Not suitable for large sample sizes

2.2 Systematic Sampling

Pick at required intervals from an ordered list, e.g. I want a sample of 15 from 60: $\frac{60}{15} = 4$ therefore choose a starting point randomly from one of the first four sampling unit from the ordered list, then choose every fourth sampling unit after until you have selected 15.

2.2.1 Positives

- Suitable for large samples
- Is easy to carry out

2.2.2 Negatives

- Sample is not random unless the ordered list is random
- Can introduce bias

2.3 Stratified Sampling

A form of random sampling: The population is split into mutually exclusive groups (strata). Random samples are taken from each strata, the relative size of each corresponds to the same ratio as each strata's representation in the total population.

2.3.1 Positives

- Works well with large samples that can be split into mutually exclusive groups
- Reflects a populations structure

2.3.2 Negatives

- Takes longer than random sampling
- Within each strata the problems are the same as with any random sample.
- Ill defined strata can overlap (meaning they are no longer mutually exclusive)
- Can't provide accurate data when strata overlap

2.4 Quota Sampling

When no sampling frame is available, quota sampling may be used. The population is divided into groups (as with stratified sampling). Quotas for each group are created that correspond with the groups representation in the total population. The interviewer then selects sampling units until each quota is reached.

2.4.1 Positives

- Administering the test is easy
- Test is low cost
- Test is quick if the sample is small

2.4.2 Negatives

- Introduces interviewer bias
- Can't estimate sampling errors

3 Types of Data

3.1 Primary Data

When you collect data, or someone collects data on your behalf.

3.1.1 Positives

- You have control over the type and method of collection
- The exact data needed is collected
- The Accuracy is known

3.1.2 Negatives

- Expensive (money and time)

3.2 Secondary Data

Second hand data, collected by another person or organisation.

3.2.1 Positives

- Cheaper than gathering primary data (time and money)
- Large amounts of data are easily available on the internet
- Access to data over time (trends)

3.2.2 Negatives

- Bias is not always acknowledged
- Accuracy is not known
- Certain data can be in a form that is difficult to deal with

4 Estimating Population Parameters using a Sample

A statistic which is used to estimate a population parameter is called an estimator. A particular value is called an estimation. If X is a random variable then $E(X)$ would be an estimator of the mean. If a statistic T is an estimator for a population parameter θ and $E(T) = \theta$ then T is an unbiased estimator for θ . Otherwise, the bias of T is given by the expression

$$E(T) - \theta$$

Estimators for population parameters can be written using “hat notation,” wherein an estimator for a population parameter θ is denoted by $\hat{\theta}$.

$$\bar{X} = \frac{1}{n} \sum_i X_i \Rightarrow E(\bar{X}) = \mu_X \quad (1)$$

$$S^2 = \frac{1}{n-1} \left(\sum_i X_i^2 - n\bar{X}^2 \right) \Rightarrow E(S^2) = \sigma_X^2 \quad (2)$$

Proof of (1) assuming $E(X + Y) = E(X) + E(Y)$,

$$E(\bar{X}) = E\left(\frac{1}{n} \sum_i X_i\right) = \frac{1}{n} E\left(\sum_i X_i\right)$$

$$E(\bar{X}) = \frac{1}{n} \sum_i E(X_i) = \frac{1}{n} n\mu = \mu$$