Sampling, Sample Size Calculation and Sampling Distributions

1. Introduction

The objective of statistical analysis is to gain knowledge about certain characteristics (properties) of a population that are of interest. Standard methods to obtain information about the characteristics of a population include taking a census, simulation, designed experiments and sampling.

1.1 Census

A census is a survey conducted on the full set of objects belonging to a given population or universe. The United Nations defines a population census as "*the total process of collecting, compiling, evaluating, analysing and publishing or otherwise disseminating demographic, economic and social data pertaining, at a specified time, to all persons in a country or a well-defined part of the country*". The census plays an essential role in public administration and is typically conducted every five or ten years.

1.2 Simulation

Simulation studies generate numbers according to a researcher-specified model. For a simulation study to be successful, the chosen simulation model must follow the real-life process closely.

1.3 Experimental design

An experimental design is a detailed plan for collecting data under controlled conditions for manipulating input factors and using the data to identify causal relationships and measure the changes in the response. Through careful planning, the design of experiments allows your data collection efforts to have a reasonable chance of detecting effects and testing hypotheses that answer your research questions.

1.4 Sampling

Sampling is the most common method for collecting information about a population. For a population if size N, a sample of size $n < N$ is selected according to a specific sampling methodology. For inferential purposes, samples must be selected randomly and must represent the population under study to produce valid and reliable estimates.

2. Sampling Methods

There are two basic sampling approaches: non-probability sampling and probability sampling.

2.1 Non-probability (non-random) sampling

Non-probability sampling includes all sampling methods where the sample units are not selected at random. Non-probability samples are useful in providing initial insights into random variables. However, samples are unlikely to truly represent the target population under investigation. This will introduce bias into the statistical findings. As it is not possible to measure sampling error in non-probability samples, such samples cannot be used for inferential statistics, only for descriptive statistics.

1 *Convenience sampling*

The researcher chooses sampling units that are readily available and convenient to find, for example intercepting people as they exit the Pick-n-Pay at a certain shopping centre.

2 *Judgment sampling*

The researcher uses his/her judgment to select the sampling units, for example selecting labour union leaders to respond to a study into working conditions.

3 *Quota sampling*

The researcher set quotas from specific subgroups of a population. When the quota for a particular group is reached, no more units are selected from that group. For example, include a quota of 20 males and 20 females in a sample.

4 *Snowball sampling*

The researcher selects sampling units based on referrals from other samplings units. This method is used when sampling units are difficult to locate.

2.2 Probability (random) sampling

Probability sampling include all sampling methods where the sample units are selected at random from the target population, i.e., all N units of the population has a chance of being sampled. Random sampling reduces selection bias and are likely to produce unbiased estimates. Since it is possible to measure sampling error in probability samples, such samples can be used for inferential statistics.

1 *Simple random sampling*

Each unit of the target population of size N has an equal chance of being selected, namely $1/N$. This sampling method is appropriate if the population is homogeneous with respect to the random variable under study. For example, selecting a subset of numbers for the Lotto.

2 *Systematic random sampling*

Used when a sampling frame/list exists. Sampling begins by randomly selecting the first sampling unit. Subsequent units are then selected at uniform intervals of size k. The steps involved in selecting a systematic sample are as follows:

1. Determine $k = \dfrac{N}{n}$
2. Randomly select a starting point = $1, 2, \ldots, k$
3. Thereafter select every k^{th} unit on the list

3 *Stratified random sampling*

Used when the population is heterogeneous with respect to the random variable under study. The population is divided into segments or strata. Within each stratum the population is homogeneous. A random sample is selected from each stratum, proportional to the size of each stratum.

4 *Cluster random sampling*

Used when the target population can be naturally divided into clusters, where each cluster is similar in profile to every other cluster. As sample of clusters is then randomly selected for sampling. All units or a random sample of units are then selected from the subset of clusters.

2.3 Some comparisons

Simple vs. *stratified*

Let's say we need as sample of 200 households across Gauteng. It is important that we include household from all socio-economic classes (SEC). If we randomly select household from the total *N* households, it is possible that we may exclude, under-represent or over-represent one of the SEC subgroups. In this case it is important to use population information about the SEC distribution in Gauteng to identify strata reflecting the SEC and then randomly select household from each stratum according to the relative distribution.

Quota vs. *stratified*

Quota and stratified sampling approaches appear the same or at least very similar. The difference is that the sampling units are selected through either probability (stratified) or non-probability (quota) sampling methodology. With reference to the example above (200 households in Gauteng reflecting SEC): In quota samples we randomly select a sample of households and then "fill" each quota by observing a unit that fits the profile. Once a quota for an SEC subgroup is reached, any subsequent household that fits that quota is excluded. In stratified sampling we select our sample points in such a way that the strata are naturally reflected. All households selected are then included in that sample.

Stratified vs. *cluster*

In stratified sampling we select a random sample from each stratum. In cluster sampling we first select a random sample of clusters, and then select all or a random sample within each selected cluster.

3. Sampling in R

There are many R libraries that can implement different sampling methodologies. We will not be using those libraries and will only use the base functions built into R to apply the different methodologies. For illustration purposes, the iris dataset in R is used to demonstrate how to select samples according to the four different probability sampling methodologies. The dataset gives measurements of sepal and petal widths and lengths for three different species of iris, with a total sample size of 150. Note: this dataset does not represent a population or sampling frame.

If you want to be able to reproduce your sample later, use the *set.seed*() function to set the seed value before calling the function that generates the sample. Remember, you must always set the seed value again before running the code to get the exact same random subset. For this illustration, set the seed using the value 1234567.

3.1 Simple random sampling

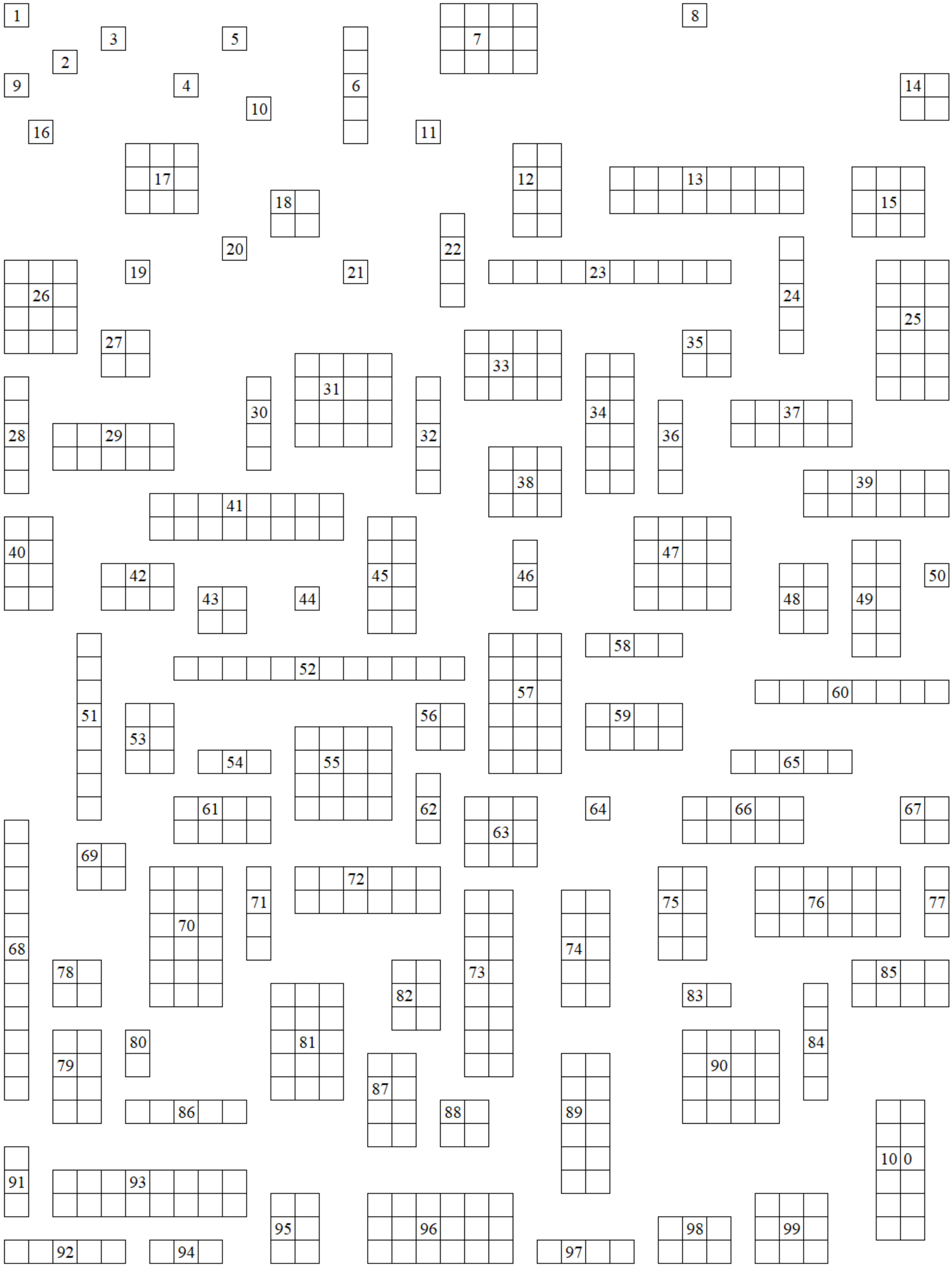
Select a simple random sample of size $n = 6$, without replacement, from the iris dataset:

```
set.seed(1234567)
srs=iris[sort(sample(nrow(iris), 6, replace = FALSE)),]
srs
     Sepal.Length  Sepal.Width  Petal.Length  Petal.Width  Species
34   5.5           4.2          1.4           0.2          setosa
59   6.6           2.9          4.6           1.3          versicolor
64   6.1           2.9          4.7           1.4          versicolor
78   6.7           3.0          5.0           1.7          versicolor
124  6.3           2.7          4.9           1.8          virginica
128  6.1           3.0          4.9           1.8          virginica
```

Exercise 1

Consider the figure on the next page, consisting of 100 rectangles (*Source: Elementary Survey Sampling, R.L. Scheaffer, W. Mendenhall III, R. Lyman Ott, Duxbury, 2006*). The goal is to choose a sample of 5 rectangles to estimate the average area of the 100 rectangles. The sizes of the 100 rectangles are given in the .csv file provided.

1. Import the *Rectangle.csv* file into R.
2. Judgment sample:
   a) Without studying the figure too carefully, quickly choose five rectangles that you think represent the population of rectangles in the figure (this is your judgment sample)
   b) Extract the sample of five rectangles you selected from the full dataset
   c) Calculate the average area of the five rectangles in your sample
3. Simple random sample:
   a) Use simple random sampling to randomly select five rectangles from the full dataset
   b) Calculate the average area of the five rectangles in your sample
4. Enter your two sample averages from the judgment and simple random samples in the survey on Moodle. I will post the .csv file with averages calculated by all students on Moodle for your to import.
5. For the full dataset (population), the variable of means from the judgment samples, and the variable of means from the simple random samples, create a plot of each variable and describe the shape, centre and variability of the variable (visually and numerically).
6. Discuss how the two distributions of sample means are similar and how they differ.
7. Which method of producing the sample means do you think is better if the goal is to use the sample mean to estimate the population mean?


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
10 0

3.2 Systematic random sampling

For a systematic random sample of size $n = 6$ from the iris dataset, the systematic increment is $k = \frac{N}{n} = \frac{150}{6} = 25$.

Randomly select a starting point = 1, 2, …, 25, thereafter select every 25th row of the dataset:

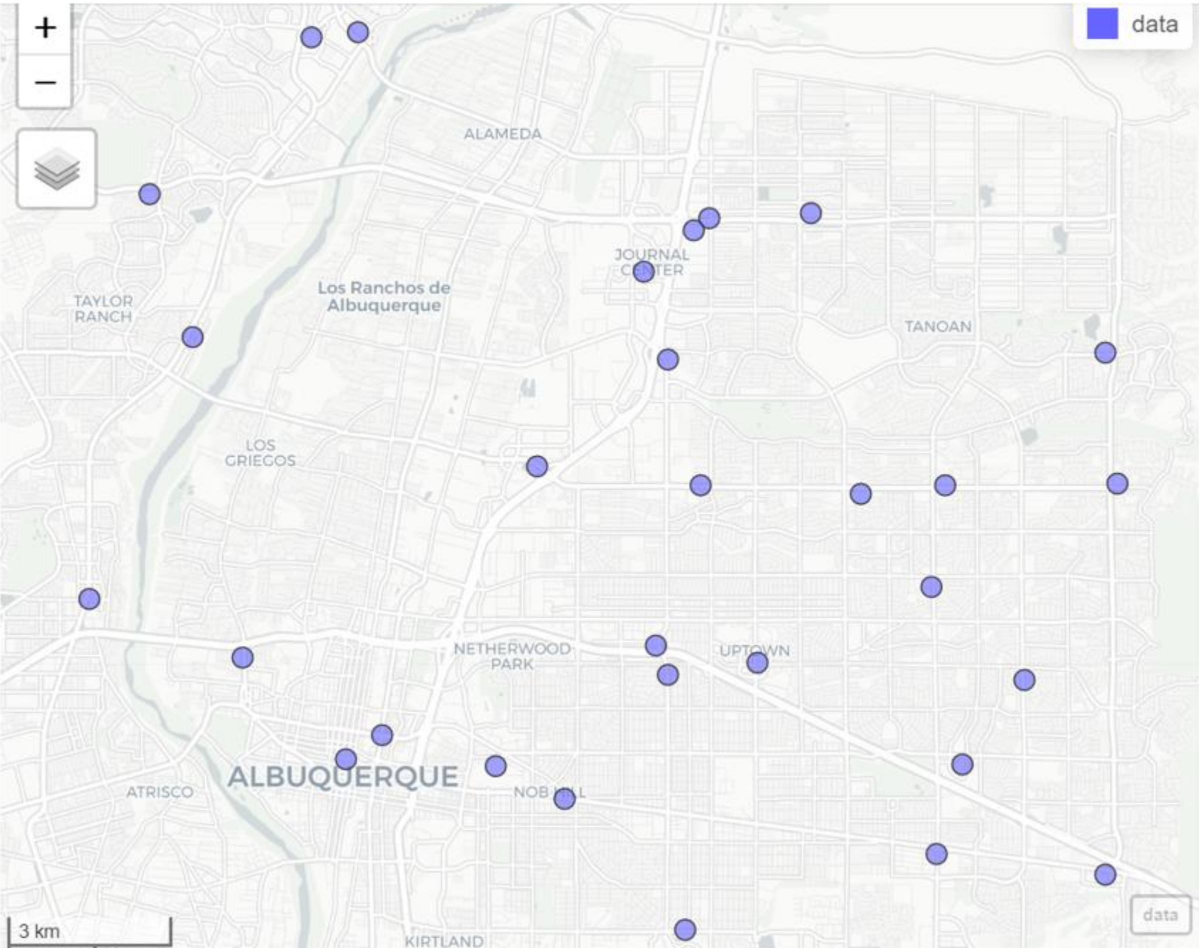
```
set.seed(1234567)
sys_list=sample(25,1)
for(i in 2:6){
  sys_list[i]=sys_list[i-1]+25
}
sys_list
[1]     17      42      67      92      117     142

sys=iris[sys_list,]
sys
        Sepal.Length  Sepal.Width  Petal.Length  Petal.Width  Species
17      5.4           3.9          1.3           0.4          setosa
42      4.5           2.3          1.3           0.3          setosa
67      5.6           3.0          4.5           1.5          versicolor
92      6.1           3.0          4.6           1.4          versicolor
117     6.5           3.0          5.5           1.8          virginica
142     6.9           3.1          5.1           2.3          virginica
```

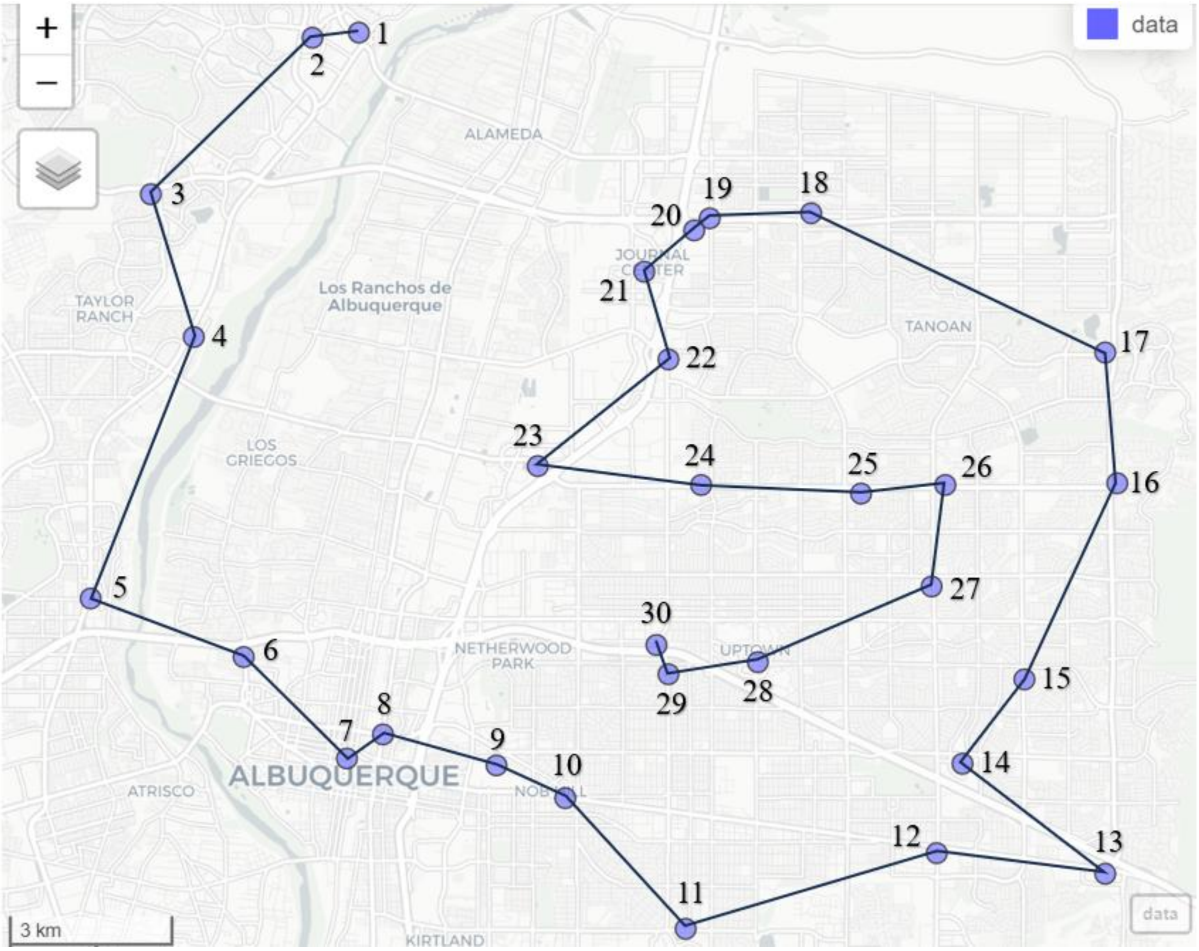
Exercise 2

A researcher wants to select a random sample of Starbucks Coffee shops in Albuquerque, New Mexico, USA. The georeferenced locations of the 30 shops in Albuquerque (hereafter abbreviated as ABQ) are shown on the first map below.

1. Open the ABQ.R script file, which includes some code you will need to create the initial map.
2. Import the ABQ.csv datafile with latitude (x-axis) and longitude (y-axis) coordinates of the 30 locations, together with path numbers, as displayed in the second map below.
3. Import the ABQcoord.csv datafile with the latitude (x-axis) and longitude (y-axis) coordinates, together with the place names of 12 areas on the map.

4. Create the map:
   a) First plot the *xy*-coordinates of the 30 locations (Note, this will resemble the map below, but the coordinates are not adjusted for curvature of the earth so there are small differences in positioning)
   b) Now add the path labels using the *text*() function, where the labels are placed above each location (pos=3) and the text size is shrunk to half the original size (cex=0.5)
   c) Now add the place names of the 12 areas using the *text*() function
   d) To create a line graph of the path in the correct order from 1 to 30, create a new data frame from the ABQ data frame, sorted in the order of the path numbers 1 to 30, and add the lines from the ordered data frame
5. Systematic random sampling:
   a) Determine the value *k* for a sample of size $n = 5$
   b) Randomly select a starting point = 1, 2, …, *k*
   c) Select every k^{th} location on the path
   d) Plot the selected sample points
6. Simple random sampling:
   a) Select a simple random sample of size $n = 5$ without replacement from the 30 locations
7. Plot the sampled locations selected through the two different sampling methodologies on the initial map in different colours.
8. Comment on the resulting samples.
9. If you want, you can try a different path to the one I created. To do this, you must change the path numbers in the data frame to new path numbers according to your own path. You can then repeat the systematic sample selection again and plot the map.


+
−
data
ALAMEDA
JOURNAL CENTER
Los Ranchos de Albuquerque
TAYLOR RANCH
TANOAN
LOS GRIEGOS
NETHERWOOD PARK
UPTOWN
ALBUQUERQUE
ATRISCO
NOB HILL
3 km
KIRTLAND
data


+
−
data
1
2
ALAMEDA
3
19
18
20
JOURNAL CENTER
21
Los Ranchos de Albuquerque
TAYLOR RANCH
4
TANOAN
17
22
23
24
25
26
16
LOS GRIEGOS
5
27
30
NETHERWOOD PARK
UPTOWN
6
29
28
15
8
7
9
10
14
ALBUQUERQUE
ATRISCO
NOB HILL
12
13
11
3 km
KIRTLAND
data

Selecting a sample of n = 6, but the sample have to be relative to some proportion of a variable.

For example, Selecting a sample of 9 from the iris dataset in proportion to the Species variable, we got 3 Species whiich means that each species will have a proportion of 3.

3.3 Stratified random sampling

8

Select a stratified random sample of size *n* = 6, without replacement, from the iris dataset, proportional to the three different iris species:

```
#determine the sample size per species
strat_freq=6*prop.table(table(iris$Species))
strat_freq
setosa          versicolor      virginica
2               2               2
```

Group the population dataset in proportion to the variable.

```
#create data frame for each species
Species1=iris[iris$Species=="setosa",]
Species2=iris[iris$Species=="versicolor",]
Species3=iris[iris$Species=="virginica",]
```

```
#select individual sample sizes from each species subset and combine into a single data frame
set.seed(1234567)
strat=rbind(Species1[sort(sample(nrow(Species1),strat_freq["setosa"])),],
          Species2[sort(sample(nrow(Species2),strat_freq["versicolor"])),],
          Species3[sort(sample(nrow(Species3),strat_freq["virginica"])),])
strat
          Sepal.Length  Sepal.Width  Petal.Length  Petal.Width  Species
14        4.3           3.0          1.1           0.1          setosa
49        5.3           3.7          1.5           0.2          setosa
77        6.8           2.8          4.8           1.4          versicolor
87        6.7           3.1          4.7           1.5          versicolor
133       6.4           2.8          5.6           2.2          virginica
143       5.8           2.7          5.1           1.9          virginica
```

Exercise 3

The *Demographic.csv* data file contains the heights and gender of 1600 people in an artificial population.

1. Select a simple random sample of size *n* = 80 (5% of the population) and estimate the average height.
2. Select a stratified random sample of size *n* = 80 (5% of the population), proportional to the gender distribution in the population, and estimate the average height.
3. Compare the estimates from the two sampling methods with the actual average height of the population, and comment on the results.

3.4 Cluster random sampling

For illustration purposes, we assume that the three different iris species represent clusters, such that the measurements within each species are diverse but similar across species. Select a cluster random sample of size *n* = 6, without replacement, from the iris dataset, where all sampling units are selected from one of the three iris species (clusters):

```
#randomly select 1 of the 3 clusters
set.seed(1234567)
sample(c("setosa","versicolor","virginica"),1)
[1] "setosa"
```

```
#create a data frame for the cluster
Cluster1=iris[iris$Species=="setosa",]
```

```
#select a sample of size 6 from the cluster
clust=Cluster1[sort(sample(nrow(Cluster1),6)),]
clust
```

```
        Sepal.Length  Sepal.Width  Petal.Length  Petal.Width  Species
14      4.3           3.0          1.1           0.1          setosa
27      5.0           3.4          1.6           0.4          setosa
33      5.2           4.1          1.5           0.1          setosa
34      5.5           4.2          1.4           0.2          setosa
37      5.5           3.5          1.3           0.2          setosa
43      4.4           3.2          1.3           0.2          setosa
```

Exercise 4

A manufacturing plant produces aluminium cans that are sold to other factories to fill with canned goods. The cans are packaged in cases, and each case contains 12 cans. As part of the quality control, the quality control inspector selects a random sample of cans at the end of the production cycle to estimate the proportion of defective cans produced during the cycle. It is generally assumed that the plant produces defective cans independently from one another during a cycle. The *Cans.csv* data file contains the population data for 60 cases of cans produced during a single production cycle. Note, this level of information is typically not available and is given to you for illustration purposes.

1. Which sample methodology do you think is the best approach to use for this problem in terms of cost, time and representation: simple random sampling, systematic sampling stratified random sampling or cluster sampling? Justify your answer.
2. Use cluster sampling to randomly select 5% of the cans produced during the cycle:
   a) ~~Total number of cans = 60 × 12 = 720 → 5% of cans produced = 720 × 0.05 = 36~~
   b) Randomly select 9 cases, then randomly select 4 cans from each case
   c) Create a bar graph of the number of cans selected per case
   d) Estimate the proportion of defective cans using the sample information
3. Use simple random sampling to randomly select 1% of the cans produced during the cycle:
   a) Randomly select *n* = 140 cans
   b) Create a bar graph of the number of cans selected per case
   c) Estimate the proportion of defective cans using the sample information
4. Compare the two estimates from the two different sampling methodologies with the true proportion of defective cans produced during the production cycle.

4. Sample Size Calculations

When we design a study, a very important component is to determine the appropriate sample size to answer the research question. The sample size can be calculated using either the confidence interval method or the hypothesis testing method. The objective of the confidence interval method is to obtain narrow confidence intervals with high reliability/precision, while the objective of the hypothesis testing method is to ensure that the test has a certain power. For the purpose of this module we will only focus on sample size calculation for inference about a single mean and a single proportion for simple random samples without replacement (SRSWR).

4.1 Confidence interval method

For confidence interval estimation, we want an appropriate sample size that will ensure that the margin of error at a certain level of confidence is sufficiently small to be informative and meaningful. In other words, we do not want the interval estimate to be too wide so that it makes no sense from a practical perspective. This is a practical issue rather than a statistical issue, where the analyst must specify the desired margin of error and required confidence level. The formulae for sample size calculation include population parameters, which are typically unknown. Analysts often use estimates of these parameters from previous similar or comparable studies, or by conducting a pilot study. When incorporating the finite population correction for confidence interval estimation from an SRSWR the sample size is:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

- n_0 is the sample size for a SRSWR
- N is the population size

One-sample mean

In studies where we want to estimate the mean of a continuous random variable for a single population from an SRSWR, the formula to calculate the appropriate sample size is:

$$n_0 = \left(\frac{z_{1-\frac{\alpha}{2}} \sigma}{e} \right)^2$$

- $z_{1-\frac{\alpha}{2}}$ is the value from the standard normal distribution reflecting the required confidence level
- σ is the population standard deviation of the variable of interest
- e is the desired margin of error

One-sample proportion

In studies where we want to estimate the proportion of successes in a dichotomous random variable for a single population from an SRSWR, the formula to calculate the appropriate sample size is:

$$n_0 = p(1-p)\left(\frac{z_{1-\frac{\alpha}{2}}}{e} \right)^2$$

- $z_{1-\frac{\alpha}{2}}$ is the value from the standard normal distribution reflecting the required confidence level
- p is the population proportion of the variable of interest
- e is the desired margin of error

Using R to calculate sample size

To determine sample sizes in R for confidence intervals we use the formulae defined above.

Exercise ~~4~~5

A researcher wants to estimate the mean systolic blood pressure of children between the ages 3 and 5 with congenital heart disease. He wants the calculated (1 – α)% CI to be within 3 units of the true mean. Assume that the population size *N* is large relative to any appropriate sample size, i.e., there is no need to use the finite population correction. Write a loop in R to calculate the sample sizes for the 90%, 95% and 99% confidence intervals, using three different assumed population standard deviation values of 10, 15 and 20 and create the following contingency table to summarise all the sample sizes.

	90% CI	95% CI	99% CI
Sigma = 10	31	43	74
Sigma = 15	68	97	166
Sigma = 20	121	171	295

Exercise 6

Suppose we want to estimate the proportion of recipes in a new cookbook that do not include any animal products. Write a function in R to calculate the SRS sample size required to attain a desired margin of error at a required confidence level for any population size. The function must give the calculated sample size when we ignore the finite population correction and when we incorporate the finite population correction. Use the function to determine the two different sample sizes for an SRS from the *N* = 1251 recipes such that the margin of error is 0.03 with 95% confidence.

4.2 Hypothesis testing method

In hypothesis testing we account for two types of error, namely Type I and Type II. A Type I error occurs when we incorrectly reject the null hypothesis, and its probability is denoted by the significance level α. We control for this error by choosing a small value for α (1%, 5% or 10%). A Type II error occurs when we fail to reject a false null hypothesis, and its probability is denoted by β. The power of a test it defined as the tests ability to correctly reject a false null hypothesis, and its probability is denoted by 1 – β. A good hypothesis test has a low probability of a Type I error and a high power.

As part of the study design we must determine the appropriate sample size that will ensure that the test has high power, where the analyst must specify the level of significance (which relates to confidence), the desired power, the variability of the variable of interest and the effect size. Note that the finite population correction is not used in sample size calculations for hypothesis testing from an SRSWR.

The effect size is a practical issue rather than a statistical issue and refers to the difference between the null and alternative hypotheses values divided by the population standard deviation (typically estimated from previous similar or comparable studies, or a pilot study). Effect size, referred to as practical significance, is not the same as statistical significance. Significance shows that an effect exists in a study, while practical significance shows that the effect is large enough to be meaningful in the real world.

Statistical significance can be misleading as an increase in sample size makes it more likely to find a significant effect, no matter how small the effect truly is in reality. Knowing the expected effect size means we can calculate the minimum sample size needed for enough statistical power to detect an effect of that size. The larger the effect size, the more powerful the study and the easier it is to detect an effect (meaningful difference).

Guidelines for the size of the effect for hypothesis test for single means and single proportions:

- 0.2 = small
- 0.5 = medium
- 0.8 = large

One-sample mean

For hypothesis testing for the mean of a continuous random variable for a single population from an SRSWR, the formula to calculate the appropriate sample size is:

$$n = \left(\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{d} \right)^2$$

- $z_{1-\frac{\alpha}{2}}$ is the value from the standard normal distribution reflecting the desired confidence level
- $z_{1-\beta}$ is the value from the standard normal distribution reflecting the desired power
- d is the effect size (Cohen's D), where:
  - $d = \frac{\mu_1 - \mu_0}{\sigma}$, i.e., the meaningful difference in the population mean divided by the population standard deviation

One-sample proportion

For hypothesis testing for the proportion of successes in a dichotomous random variable for a single population from an SRSWR, the formula to calculate the appropriate sample size is:

$$n = \left(\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{h} \right)^2$$

- $z_{1-\frac{\alpha}{2}}$ is the value from the standard normal distribution reflecting the desired confidence level
- $z_{1-\beta}$ is the value from the standard normal distribution reflecting the desired power
- h is the effect size (Cohen's *h*), where:
  - $h = 2\operatorname{asin}\sqrt{p_1} - 2\operatorname{asin}\sqrt{p_0}$
  - Note, this is one of many different effect size calculations for proportions, and is the formula used in R

Using R to calculate sample size

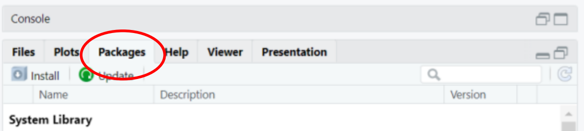
To determine sample sizes in R for hypothesis testing, we can either use the formulae defined above or we can make use of an R package called "pwr". This package/library is not part of the base R package and must be installed from CRAN.

1) Install a package using the following code:

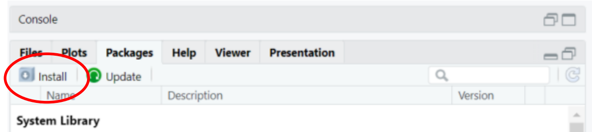
```
install.packages('pwr')
library(pwr)
```

2) Directly install a package in RStudio as follows:

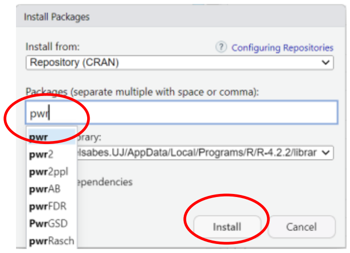
- Click on Packages



- Click on Install



- Type the name of the package you want to install, e.g., pwr, and click on Install

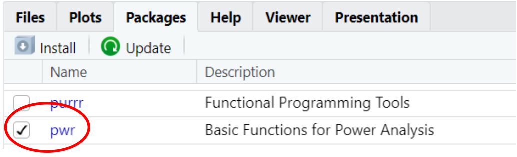

Install Packages

Install from: Repository (CRAN)

Packages (separate multiple with space or comma): pwr

...lsabes.UJ/AppData/Local/Programs/R/R-4.2.2/librar

- Scroll down on the list of packages and tick pwr



	Name	Description
□	purrr	Functional Programming Tools
☑	pwr	Basic Functions for Power Analysis

- You may get a warning message that the package was built under a different version of R, but this is not an issue

We will use the *pwr.t.test()* and *pwr.p.test()* functions for calculations for mean and proportions, respectively. The functions are structured is such a way that we can find the sample size, level of significance, power or effect size for given values of all other arguments. The *ES.h()* function calculates the effect size using the arcsin formula. The arguments of these functions, including default values and lists of possibilities, are as follows:

```
pwr.t.test(n = NULL, d = NULL, sig.level = 0.05, power = NULL,  type = c("two.sample", "one.sample",
        "paired"), alternative = c("two.sided", "less", "greater"))
pwr.p.test(h = NULL, n = NULL, sig.level = 0.05, power = NULL,alternative = c("two.sided","less","greater"))
ES.h(p1, p2)    #p1 = proportion under alternative hypothesis, p2 = proportion under null hypothesis
```

Exercise 7

A researcher hypothesizes that fasting blood glucose for people without diabetes increases if they drink at least two cups of coffee per day. Previous studies have shown that the mean fasting blood glucose level in people free of diabetes is 95 mg/dL with a standard deviation of 9.8 mg/dL. If the mean fasting blood glucose level in people free of diabetes who drink at least two cups of coffee per day is 100 mg/dL, this would be viewed as significant from a practical perspective. How large a sample must be selected to ensure that the power of the test is 80% to detect this difference at a 5% level of significance? Look at how the format of the alternative hypothesis relates to the effect size value calculation.

Exercise 8

A medical device manufacturer produces implantable stents. Historical records showed that approximately 10% of the stents are defective. The quality control manager believes that this is no longer the case. A pilot sample showed that 13% of stents are defective. Calculate the sample size that will ensure that the appropriate hypothesis test at a 1% level of significance has 85% power to detect a 0.03 difference. Then calculate the sample size for the same criteria for small, medium and large effect sizes.

5. Sampling Distributions

5.1 Parametric

In statistical inference, sample statistics are used to estimate population parameters. To assess the accuracy of an estimate of a parameter, the probability distribution of the statistic of interest is used to place probabilistic bounds on the sampling error. The probability distribution associated with all the possible values that a statistic can assume is called the sampling distribution of the statistic. In practice, sampling is generally done without replacement. If the population is finite, or if the population is relatively small, then the finite population correction must be incorporated to derive the sampling distribution of a statistics. We will only do this in the theory lectures.

5.2 Non-parametric bootstrapping

An alternative method to estimate or approximate the sampling distribution of a statistic is through re-sampling methods. Re-sampling methods are based on taking samples from the sampled data and then using these re-samples to calculate statistics. This method can actually give more accurate answers than using the single original sample to calculate an estimate of a parameter and derive the sampling distribution. Re-sampling methods require fewer assumptions than the traditional methods and sometimes give more accurate answers. The most re-sampling common is bootstrapping.

Suppose that a random sample of size *n* is taken from an unknown probability distribution, and the values of a variable X are observed and used to estimate a parameter. The traditional approach of estimating is to make some assumptions about the population and to derive the sampling distribution of the statistic based on these assumptions. Often the bootstrap is used to empirically verify mathematical results and it might even be preferable to extensive mathematical calculations.

In the bootstrap method the original sample takes the place that the population holds in the traditional approach. Subsequently, a random sample of size *n* is drawn from the sample with replacement, referred to as the bootstrap sample. This procedure is repeated a large number of times and the statistic of interest (say the mean) is calculated for each of the bootstrap samples, yielding a variable of means, which is then used to estimate the mean and standard error of the sampling distribution of the statistic of interest. The variation in the value of the estimator between bootstrap samples will be a measure of the variation to be expected in the estimator, had it been possible to take several samples from the population. The larger the sample size of the original sample, the more it will re-sample the population it was drawn from and the more accurate this measure of precision for the estimator will be.

The general procedure for estimating the expected value and standard error of an estimator is:

1) Generate B independent bootstrap samples, each consisting of *n* values selected with replacement from the sample.
2) Compute the statistic of interest for each bootstrap sample.
3) Calculate the mean and standard deviation of the values calculated in (2) as the estimate of the expected value and standard error of the sampling distribution of the statistic.

For most statistics, the bootstrap distribution approximates the shape, spread and bias of the actual sampling distribution, but can differ in the location of the centres.

Exercise 9

Consider the following sample of heights (measured in centimetres) of *n* = 22 Grade 6 learners:

141.0	156.5	162.0	159.0	157.0	143.5	154.0	158.0	140.0	142.0	150.0
148.5	138.5	161.0	153.0	145.0	147.0	158.5	160.5	167.5	155.0	137.0

1. Create a vector of heights for this data in R.
2. Determine the minimum, maximum, median and mean values of height.
3. Plot the histogram of height values and comment on the shape of the distribution in terms of modality and symmetry.
4. Plot the box-and-whisker plot of the height values and comment on the shape of the distribution in terms of symmetry.
5. Draw 2 bootstrap samples from the sample of observations, each of size *n* = 22:
   a) Show which observations from the original sample are included in each of your bootstrap samples and comment on what you see
   b) For each bootstrap sample, calculate the minimum, maximum, median and mean values of height and compare these with the statistics calculated in (2)
6. Draw 100 bootstrap samples from the sample of observations, each of size *n* = 22:
   a) Create a matrix with the median and mean values of height for each bootstrap sample
   b) Plot the histogram and box-and-whisker plot of the median height, and of the mean height
   c) Estimate the mean and standard deviation of the distribution of sample medians and the distribution of sample means
   d) Which is the least variable (more precise) estimate: median or mean?