UNIVERSITY OF JOHANNESBURG

Statistics 2B (Practical)

Compiled by Mr. V van Appel, Department of Statistics - University of Johannesburg
Practical video done by Mr. T Hansragh, Department of Statistics - University of Johannesburg

Practical 2: Sampling and sampling distributions

https://youtu.be/M29ehuHBrwY

The objective of statistical analysis is to gain knowledge about certain characteristics (properties) of a population that are of interest.

2.1 Methods of obtaining information

Standard methods to obtain information about the characteristics of a population include taking a census, simulation, designed experiments and sampling.

2.1.1 Census

When the population is small, the best way to study it, is to consider all the units in the population. This process of collecting information on the entire population is called a census. Usually it is not possible to do a census due to factors such as: monetary and time constraints and it may be difficult to find all the units in a population.

2.1.2 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. Example: The effects of an earthquake on buildings are often simulated.

2.1.3 Designed Experiment

If the researcher can control certain variables of interest in a study, a designed experiment may be used to obtain data. The objective is to gain information about the in influence that the variables have on response of the given experiment.

2.2 Sampling

Sampling is the most frequently used form of collecting information about a population. Methods of obtaining a representative sample from the population include: random sampling, simple random sampling, stratified sampling, systematic sampling, and cluster sampling.

Assume that the population size is N and that the sample size is n with $n < N$.

2.2.1 Random sampling

This is the process of selecting n elements from a population in such a way that each of the n units has the same probability $1/N$, of being selected. Thus the random variables $X_1, X_2, \ldots, X_n$ from a random sample of size n from a population with pdf $f(x)$, if $X_1, X_2, \ldots, X_n$ are mutually independent random variables such that the marginal pdf of each X_i is $f(x)$ (i.e. the X_i are i.i.d).

The objective of random sampling is to obtain a representative sample from the population, which can be used to make generalizations about the population. Typically random numbers which indicate the units to be included in the sample, are generated. When the population is finite, it is possible to list all the possible combinations of samples of size n using the R-command expand.grid().

List all the combinations of size $n = 3$ from a population consisting of $N = 4$ items. Consider the following R-codes:

```
> #Number of possible combinations
> 4^3
```

```
[1] 64
```

```
> expand.grid(1:4,1:4,1:4)
```

```
   Var1 Var2 Var3
1     1    1    1
2     2    1    1
3     3    1    1
4     4    1    1
5     1    2    1
6     2    2    1
7     3    2    1
8     4    2    1
9     1    3    1
10    2    3    1
11    3    3    1
12    4    3    1
13    1    4    1
14    2    4    1
15    3    4    1
16    4    4    1
17    1    1    2
```

```
18    2    1    2
19    3    1    2
20    4    1    2
21    1    2    2
22    2    2    2
23    3    2    2
24    4    2    2
25    1    3    2
26    2    3    2
27    3    3    2
28    4    3    2
29    1    4    2
30    2    4    2
31    3    4    2
32    4    4    2
33    1    1    3
34    2    1    3
35    3    1    3
36    4    1    3
37    1    2    3
38    2    2    3
39    3    2    3
40    4    2    3
41    1    3    3
42    2    3    3
43    3    3    3
44    4    3    3
45    1    4    3
46    2    4    3
47    3    4    3
48    4    4    3
49    1    1    4
50    2    1    4
51    3    1    4
52    4    1    4
53    1    2    4
54    2    2    4
55    3    2    4
56    4    2    4
57    1    3    4
58    2    3    4
59    3    3    4
60    4    3    4
61    1    4    4
62    2    4    4
63    3    4    4
64    4    4    4
```

Example 2.1 *Suppose a population consists of the $N = 3$ units: 2, 5, 8. Enumerate all possible combinations of size $n = 2$.*

Try doing this in R. The solution will be made available after the Practical class.

2.2.2 Simple random sampling

This is the most elementary form of sampling, where each particular sample of size n has the same probability of occurring. In finite population each of the $\binom{N}{n}$ samples of size n is taken without replacement and has the same probability of occurring.

In finite populations no distinction is made between sampling with or without replacement. Thus the probability of selecting a given unit is the same whether sampling is done with or without replacement.

Most sampling is done without replacement due to its case and increased efficiency in terms of variability when compared to sampling with replacement.

Assume that all the values in the population are numbered sequentially from 1 to N. To list all the possible combinations of size n when sampling without replacement from a finite population of size N (that is all $\binom{N}{n}$ combinations), load the PASWR package and use the function combn()

Example 2.2 *Given a population of size $N = 5$, list all the possible samples of size $n = 3$, that may be obtained through simple random sampling.*

```
> #List of possible samples (First install the packages and then load them)
> #install.packages("PASWR")
> library(PASWR,e1071)
> combn(5,3)
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    1    1    1    1    1    2    2    2     3
[2,]    2    2    2    3    3    4    3    3    4     4
[3,]    3    4    5    4    5    5    4    5    5     5
```

```
> #The combinations, listed vertically in the output, are: (1,2,3), ...,(3,4,5)
```

Example 2.3 *A teacher wants to randomly select a sample of 5 students from 180 students to present their work to the class.*

```
> # Assume the students in the class are numbered from
> # 1 to 180 according to the class list
> sample(1:180,5,replace=F)
```

```
[1] 174  44  23  77 132
```

(The other methods of sampling will not be discussed.)

2.3 Parameters

Once a sample is taken the main objective is to obtain the maximum and most precise information possible about the population's parameters from the sample.

A parameter $\theta = t(F)$, is a function of the probability distribution F. (Although F has been used to denote the *cdf*, it now refers to any description of a random variable's probabilities). Recall that parameters characteristics probability distributions and are inherent in all probability models. Therefore it is impossible to calculate a probability without prior knowledge of the distribution's parameters. We will follow the classical approach and consider parameters to be constants.

Example 2.4 *Suppose F is the exponential (λ)-distribution and $t(F) = E_F(X) = \theta$. Then $t(\cdot)$ is the expected value of X and $\theta = 1/\lambda$.*

2.4 Estimators

Population parameters are generally unknown and therefore must be estimated using sample data $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$. An estimator (or statistic) is a function of the sample, $T = t(\mathbf{X}) = \hat{\theta}$, while an estimate (a number) $t(x)$ is the observed value of the estimator that is obtained for an observed sample. Note that an estimator is a random variable, as it is a function of random variables and its value changes from sample to sample.

Example 2.5 *Consider the parameter, the population mean μ. It can be estimated by the sample mean*

$$T = t(\mathbf{X}) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.1}$$

which is a statistic or estimator constructed from the random sample $\{X_1, X_2, \ldots, X_n\}$. Until a sample is taken, the value of the statistic (the estimator) is unknown. Suppose a random sample has taken that contains the following values: $\mathbf{X} = \{3, 5, 6, 1, 2, 7\}$. Now the estimate of (value of the statistic) $T = t(\mathbf{X})$ is 4. We also write $\hat{\mu} = 4$.

2.5 Empirical probability distribution function (ecdf)

Let $x = (x_1, x_2, \ldots, x_n)$ denote a sample of size n from a distribution F. The *epdf* $\hat{f}$, is defined as the discrete distribution that assigns probability $1/n$ to each value x_i. The empirical probability distribution function (ecdf), is defined as:

$$\hat{F_n}(t) = \sum_{i=1}^{n} \mathbb{I}(x_i \leq t)/n, \qquad \textit{where} \quad \mathbb{I}(x_i \leq t) = \left\{ \begin{array}{ll} 1 & x_i \leq t \\ 0 & x_i > t \end{array} \right. \tag{2.2}$$

The R-command ecdf() can be used.

Example 2.6 *Simulate rolling a die 100 times and compute the epdf. Graph the ecdf.*

Try doing this in R. The solution will be made available after the Practical class.

The epdf $\hat{f}$ can be used to calculate the plug-in estimator $\hat{\theta} = t(\hat{f})$ of a parameter $\theta = t(f)$. For example, the plug-in estimator of the expected value is the sample mean.