Metro Ciencia

Editorial: Hospital Metropolitano
ISSN (impreso) 1390-2989 - **ISSN (electrónico)** 2737-6303
Edición: Vol. 29 N° 3 (2021) Octubre - Diciembre
DOI: https://doi.org/10.47464/MetroCiencia/vol29/4/2021/65-72
URL: https://revistametrociencia.com.ec/index.php/revista/article/view/302
Pág: 65-72

Introduction to Estimation

Introducción a la estimación

Maria Carolina Velasco[1], Isaac Zhao[2]

Latin American Center for Clinical Research. Quito - Ecuador[1]
Worcester Polytechnic Institute. Worcester, MA - United States[2]

Recibido: 24/09/2021 Aceptado: 01/10/2021 Publicado: 30/11/2021

ABSTRACT

The present paper is an introduction to standard point and interval estimation methods. It covers statistical principles and sampling processes that are building blocks for arguably more advanced statistical analyses and inference methods. By providing examples and sample code in R, the paper sets an important practical basis for the theory of inferential statistics.

Keywords: Point estimation, interval estimation, central limit theorem, law of large numbers, sampling distribution, bootstrap.

RESUMEN

Este artículo es una introducción a métodos de estimación puntual y por intervalos. Cubre principios estadísticos y procesos de muestreo que son importantes para aplicar análisis y métodos de inferencia más avanzados. A través de ejemplos y código de muestra en R, el artículo establece una base práctica para la teoría de la estadística inferencial.

Palabras claves: Estimación puntual, estimación por intervalos, teorema del límite central, ley de los grandes números, distribución muestral, bootstrap.

IDs Orcid
Maria Carolina Velasco: https://orcid.org/0000-0002-8482-9865
Isaac Zhao: https://orcid.org/0000-0002-4352-4969

Correspondencia: Maria Carolina Velasco
e-mail: ma.carolina.velasco@gmail.com

1. INTRODUCTION

Estimation is the process of making inferences about the population based on the information from a sample. The present paper provides readers with important background on statistical principles, sampling processes, and point and interval estimators. These topics are building blocks for hypothesis testing, regression models, and many other statistical methods. Hence, although a bit more theoretical, the content discussed herein is important for arguably more advanced statistical analyses.

Throughout the paper we draw from Shahbaba's introductory book to biostatistics[4] and Bruce et al.'s practical guide for data scientists[2]. Additionally, we analyze the dataset called "Survival from Malignant Melanoma" (referred to as melanoma moving forward) to exemplify methods and tools. The dataset, publicly available in the "boot" package in R., contains demographic and tumor characteristics of patients with malignant melanoma in Denmark[1].

2. SAMPLING

Research questions are answered by analyzing data of a representative sample taken from the population of interest. With caution, the conclusions reached at the sample level via inference methods can be applied at the population level.

To simplify statistical analyses, it is common to assume that all members of a sample are taken independently from each other such that the selection of one participant does not affect the selection of another. Likewise, we assume that the members of a sample have the same probability distribution. That is, if we plotted all the variables together, they would resemble a specific distribution[4]. Both characteristics are part of a property called i.i.d. that describes independent and identically distributed random variables.

2.1 Point Estimation and the Law of Large Numbers

Researchers use sample quantities, or sample statistics, to estimate unknown population parameters. We can represent unknown population quantities either with a single value via point estimation or with a range of possible values via interval estimation.

Some of the most common point estimators of interest are the sample mean $\hat{x}$, sample proportion $\hat{p}$, and sample variance s^2. Point estimators are random variables, meaning that, if we take different samples from the population, we may get different estimates each time[4]. However, the Law of Large Numbers (LLN) states that the sample mean of i.i.d. random variables becomes closer to the true population mean as the sample size increases. The LLN also applies to the sample proportion since it is the mean of binary random variables. Justified by this law, the sample mean or sample proportion are estimates of the population mean[4].

2.2 The Sampling Distribution and the Central Limit Theorem

Sample estimates may take on different values from one sample to another, which is why they have a probability distribution that summarizes the likelihood of observing all the possible values. The sampling distribution is the probability distribution of a sample statistic over many samples[2,4]. For instance, in the melanoma dataset, the mean age of the study participants is 52 years. Nevertheless, if the study had taken another sample of patients with melanoma, the mean age of the new participants would vary. This is called sampling variability. Oftentimes, we only have access to one sample and, hence, to only one sample statistic. To obtain the sampling distribution, we resort to taking smaller samples from the original sample. A widespread method to do so is called bootstrap, covered in Section 2.3.

Two fundamental statistical principles pertain to the sampling distribution: the Law of Large Numbers and the Central Limit Theorem. The LLN states the larger the sample, the closer the sample mean will be to the true population mean. Therefore, the larger the sample, the narrower the variability of the sampling distribution[2].

The Central Limit Theorem (CLT) indicates that, for i.i.d random variables, the sampling distribution of the sample mean approximates a normal distribution as the sample size increases. Accordingly, the sampling distribution of the mean age of patients with melanoma would follow a normal distribution with a big enough sample size. The CLT simplifies statistical problems because it is valid even if the underlying distribution of the source population is not normal. The normal-approximation formulas that are derived from the CLT are commonly used in statistical inference methods like confidence intervals and hypothesis testing[2].

2.3 The Bootstrap and Standard Error

Bootstrapping is the process of sampling with replacement from the original sample to estimate the sampling distribution of a statistic. Sampling with replacement means that, after we take an observation from the sample, we replace it such that the probability of choosing an observation remains unchanged from draw to draw[2]. Let's suppose we want to find the sampling distribution of the mean age of melanoma patients. As explained by Bruce et al.[2], the steps to perform the bootstrap are:

1. Draw a random value from the original sample, record it, and then replace it.

2. Repeat step 1 n times.
3. Compute the test statistic (e.g. sample mean) of the n resampled values.
4. Repeat steps 1 and 2 K times.
5. Use the K results to obtain insights on the sample statistic and its sampling distribution

The bootstrap is a powerful tool for evaluating the variability of a sample statistic. The standard deviation of the sampling distribution is called the standard error (SE):

$$SE = \frac{s}{\sqrt{n}},$$

Where s is the standard deviation of the sample statistic and n is the sample size. The larger the sample size, the smaller the standard error. This is why the bootstrap can be used to assess how the sample size affects the sampling variability[2].

The function boot from the boot package in R[1] implements the bootstrap and computes the standard error at once[2]. The output indicates that the mean age of melanoma patients in the original sample is 52.5 years. By generating bootstrap replicates of the mean age, the algorithm estimates that the sample statistic has a bias of 0.0068 and a standard error of 1.17 years. We use the function set.seed in the code to avoid having slightly different results between consecutive runs of the algorithm.

```
# reate a function to obtain the mean for a given sample specified by
the index idx
stat_fun <- function(x, idx) mean(x[idx])
# Initiate boot package
library(boot)
# Set reproducible random samples
set.seed(1)
# Generate R=1000 bootstrap replicates of the mean age of melanoma
patients, using 1000 replicates
boot_obj <- boot(data=melanoma$age, R=1000, statistic=stat_fun)
# Returns the observed sample mean in the original data, its bias and
its standard error.
boot_obj

ORDINARY NONPARAMETRIC BOOTSTRAP
Call:
boot(data = melanoma$age, statistic = stat_fun, R = 1000)
Bootstrap Statistics :
    original       bias    std. error
t1* 52.46341  0.006760976    1.171995
```

There are noteworthy applications for the bootstrap. A benefit of the method is that it does not rely on the CLT or any other distribution assumptions. Therefore, it is commonly used in analyses that do not assume a mathematical approximation to the sampling distribution. The bootstrap is also widely used in predictive studies to assess the stability and improve the predictive power of a model. For instance, in a process called bagging, the predictions of multiple bootstrap samples are aggregated to outperform the predictions of a single model[2].

It is important to note that, although the bootstrap allows us to have an infinite number of samples, it does not compensate for small sample sizes because the method itself does not generate new data[2].

3. PROBABILITY DISTRIBUTIONS

Probability distributions are mathematical functions that help us model a range of phenomena by estimating the probability of events and the variability of occurrence. A key challenge in statistical problems is the identification of a distribution that can be properly applied to a variable based on its characteristics[4]. There are several distributions that have been well-researched and analyzed; this section focuses on the normal and Student's t-distributions.

3.1 The Normal Distribution

The normal distribution is a bell-shaped curve that is symmetric around the mean, implying that the mean, median, and mode are close to each other and coin-

cide at the peak of the curve. The normal distribution is specified with two parameters: the mean μ, representing the maximum point of the curve, and the variance σ^2, representing the spread of the curve around the mean. A normally distributed random variable *X* is denoted as $X \sim N(\mu, \sigma^2)$.

3.1.1 Sampling Distribution of the Sample Mean

According to the CLT, regardless of the underlying distribution of the random variables $X_1, X_2, \ldots, X_n$, the sampling distribution of the sample mean is normal with the parameters:

$$\underline{X} \sim N(\mu, \sigma^2/n)$$

The mean age of patients in the melanoma study is 52 years and the standard deviation is 17 years. With a sample size of *n* = 205, the sample mean follows the distribution:

$$\underline{X} \sim N(52, 17^2/205)$$
$$\underline{X} \sim N(52, 1.4)$$

The standard error of the sample mean is equal to $s/\sqrt{n} = 17/\sqrt{205} = 1.18$ years and it reflects the extent of the variability of the sample mean as an estimator for the population mean[4].

The right panel of Figure 1 shows the density of $\underline{X}$. The sampling distribution of the mean age is centered on the population mean (vertical line). However, compared to the (unknown) theoretical distribution of age (left panel), the sampling distribution has a much smaller variance. Note the different scales on the x-axis[4].

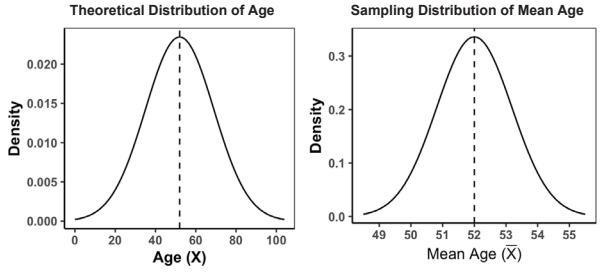

Figura 1. Left panel: The (unknown) theoretical distribution of age, $X \sim N(52, 289)$. Right panel: Density curve for the sampling distribution $\underline{X} \sim N(52, 1.4)$.

3.1.2 Sampling Distribution of the Sample Proportion

Based on the CLT, the sample proportion is normally distributed:

$$\hat{p} \sim N(p, \frac{p(1-p)}{n}),$$

As long as $np \geq 10$ and $n(1-p) \geq 10$, where *n* is the sample size and *p* is the sample proportion.

Given that 126 out of the 205 participants of the melanoma study were female, the sample proportion follows the distribution $\hat{p} \sim N(0.61, 0.001)$.

3.2 The Standard Normal Distribution

The standard normal distribution is a normal distribution with a mean of zero and a standard deviation (and variance) of one, *N*(0,1). The random variable of a standard normal distribution is called a z-score and it represents the number of standard deviations a value (or score) is from the mean. We can transform or standardize any random variable of a normal distribution with a z-score. The z-score for the *i*th observation of a sample is computed as follows:

$$z_i = \frac{x_i - \underline{X}}{s},$$

where x_i is a normal random variable, $\underline{X}$ is the sample mean, and *s* is the sample standard deviation.

Recall that the mean age of patients in the melanoma study is 52 years and the standard deviation is 17 years. Assuming that age has a relatively normal distribution, the z-score of a 69-year old patient is $z = \frac{69-52}{17} = 1$. As the standard normal distribution in Figure 2 shows, the age of this patient would be found at the right hand side exactly one standard deviation away from the mean.

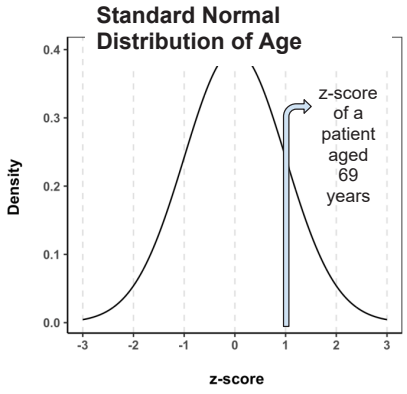

Figura 2. Standard normal distribution of age, assuming that it is normally distributed.

3.3 Student's t-distribution

Another known probability distribution is called the Student's t-distribution, which is frequently used in estimation problems when the sample size is small

and the population variance is unknown. The t-distribution is a bell-shaped curve that is "flatter" when compared to the normal distribution. To account for the uncertainty in the variation, the t-distribution gives a lower probability to the center and a higher probability to the tails[2].

The t-distribution is represented by a parameter called the degrees of freedom (*df*) that is a function of the sample size. As exemplified in Figure 3, the larger the sample size ($n > 30$ as a rule of thumb), the more the t-distribution approximates the standard normal distribution.

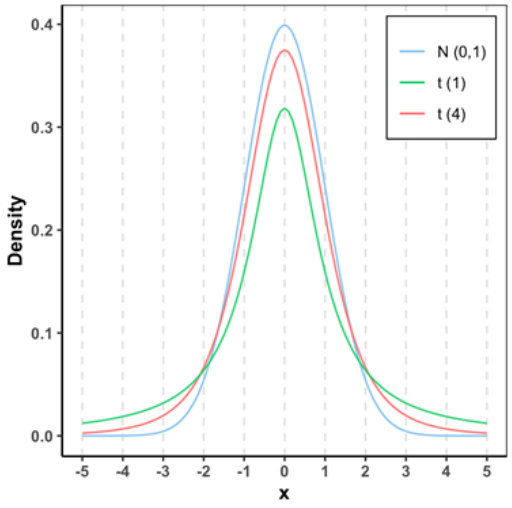

Figura 3. Overlapped distributions: standard normal distribution, t-distribution with df=1, and t-distribution with df=4.

4. CONFIDENCE INTERVALS

Point estimation does not consider the uncertainty of the sample statistic (i.e. the standard error). Confidence intervals overcome this limitation by providing a range of possible values that is likely to contain the unknown population parameter. A confidence interval is an expression of the point estimate and its standard error, at a specific confidence level. A common confidence level is 95%, indicating that the interval is 95% likely to contain the unknown population parameter.

Let's dive into what a 95% confidence level means by assuming that the true mean age of the patients with melanoma in Denmark is 50 years. If we were to take 40 different independent samples from this population and find the 95% confidence interval for each of the sample means, we would expect 38 (or 95%) of those intervals to contain the true population mean[4]. This example is presented in Figure 4. The confidence interval obtained in the study could be either one of the 38 intervals that contain the true population mean or one of the two intervals that do not.

We often have access to only one point estimate and one confidence interval in research studies. Hence, a confidence level of 95% means that we have a 95% confidence that the procedure that generated the interval contains the true population parameter[2,4].

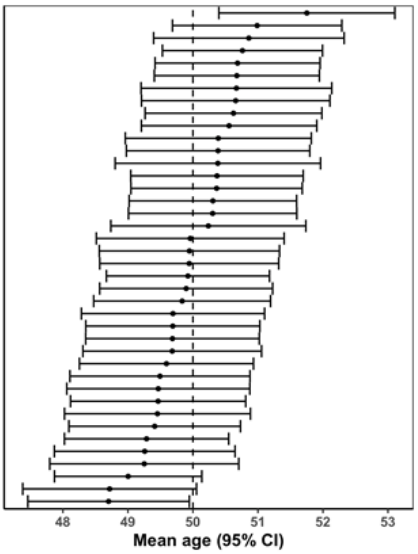

Figura 4. 95% confidence intervals obtained from 40 independent samples. 95% of all the intervals (38/40) include the true population mean of μ=50 (vertical dotted line).

4.1 Population Mean

Assuming that the population variance σ^2 is known, the 95% confidence interval for the population mean μ is:

$$[\underline{x} - 2\frac{\sigma}{\sqrt{n}}, \underline{x} + 2\frac{\sigma}{\sqrt{n}}]$$

The multiplier in the confidence interval formula is called the z-critical value and it depends on the confidence level. In the example above, the z-critical value is 2 because we are calculating an interval with a 95% confidence level. The empirical rule for normal distributions, shown in Figure 5, states that 95% of the observations fall roughly within two standard deviations from the mean. To compute the 99.7% confidence interval, we would instead multiply by 3. Likewise, to obtain a 68% confidence interval, we would use the multiplier 1.

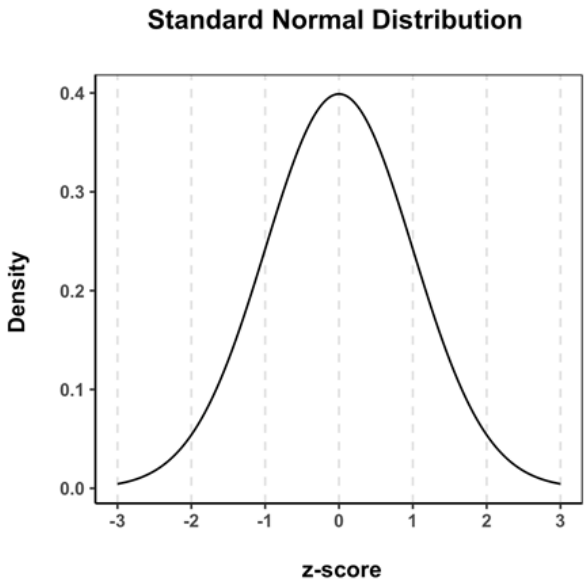

Figura 5. The 68–95–99.7 empirical rule for normal distributions.

The definition of the z-critical value is an important step in interval estimation because it allows us to obtain the area that contains the desired confidence level for the interval. For any confidence level c, the z-critical value is denoted as $z_{crit} = z_{\frac{1-c}{2}}$. In R, we can easily compute the value with the function qnorm. As seen with the empirical rule, thez_{crit} for a 95% confidence level is *1.96 ≈ 2*.

We can then proceed to estimate the confidence interval for the population mean when the population variance is known with the following formula:

$$[\underline{x} - z_{crit}\frac{\sigma}{\sqrt{n}}, \underline{x} + z_{crit}\frac{\sigma}{\sqrt{n}}].$$

```
# Obtain the z-critical value for a 95% confidence interval
z_crit = qnorm(0.975)
```

Accordingly, with a probability of 95%, the mean age of the population of melanoma patients in Denmark is within the interval:

$$[52.46 - 1.96\frac{17}{\sqrt{205}}, 52.46 + 1.96\frac{17}{\sqrt{205}}]$$
$$[50.18, 54.75]$$

The 95% confidence interval can be computed in R as described below:

```
# Obtain the mean age
mean_age = mean(melanoma$age, na.rm=T)
# Obtain the standard deviation of age
sd_age = sd(melanoma$age, na.rm=T)
# Obtain the sample
sample_size = nrow(melanoma)

# Obtain the z-critical value
z_crit = qnorm(0.975)

# Obtain the standard error
standard_error =sd_age/sqrt(sample_size)

# Obtain the lower bound of the CI
mean_age - (z_crit*standard_error)
50.18123
# Obtain the upper bound of the CI
mean_age + (z_crit*standard_error)
54.7456
```

The population variance is rarely known in research studies because we often need to estimate the population variance along with the population mean. In such cases, we estimate σ with the standard error $\frac{s}{\sqrt{n}}$ and we use the t-distribution instead of the normal distribution. Therefore, the confidence interval for the population mean when the population variance is unknown is:

$$[\underline{x} - t_{crit}\frac{s}{\sqrt{n}}, \underline{x} + t_{crit}\frac{s}{\sqrt{n}}],$$

Where t_{crit} is obtained from a t-distribution with n-1 degrees of freedom.

We can use R to easily calculate the 95% CI for mean age in the melanoma dataset assuming σ^2 is unknown. The code below returns two numbers, the lower and upper bounds of the confidence interval. We are 95% confident that the true population mean age is between 50.17 and 54.76.

```
# Use the t.test function, specify the 95% confidence level, and use
$conf.int to extract the confidence interval.
t.test(melanoma$age,conf.level=0.95)$conf.int
50.16761 54.75922
```

Note that the confidence interval based on the t-distribution is wider than the one based on the normal distribution because the former accounts for the unknown variance. However, if the sample size increases, the t-distribution approaches the standard normal distribution, as seen in Figure 3.

4.2 Population Proportion

Section 2.2 specified that, under the CLT, $\hat{p}$ is normally distributed with $\mu_p = p$ and $\sigma^2_p = \frac{p(1-p)}{n}$. Therefore, assuming the population variance σ^2 is known, the confidence interval for the population proportion p is obtained as follows:

$$[p - z_{crit}\sqrt{\frac{p(1-p)}{n}}, p + z_{crit}\sqrt{\frac{p(1-p)}{n}}]$$

The 95% confidence interval for the proportion of females in the melanoma dataset can be estimated in R with the code below. We are 95% confident that the true population proportion of females is between 0.54 and 0.68.

```
# Obtain the number of females in the sample
n_female = sum(melanoma$sex == 0)
# Obtain the sample size
sample_size = length(!is.na(melanoma$sex))
# Use the prop.test function, specify the 95% confidence level, and
use $conf.int to extract the confidence interval.
prop.test(x = n_female, n = sample_size, conf.level = 0.95)$conf.int
0.5440159 0.6808823
```

5. CONCLUSION

The present paper deals with point and interval estimation methods. It provides an introduction to the Law of Large Numbers and the Central Limit Theorem. These principles simplify problems in statistics by making normal approximations and justifying the estimation of population parameters. By discussing basic probability distributions and sampling methods like the bootstrap, the paper sets an important basis for the theory of inferential statistics.

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding Sources

The study did not require funding sources.

Author Contributions

MCV, IZ: Content outline, investigation of paper references, and edits.

MCV: Write-up and analysis interpretation.

IZ: R analysis, figures, and code.

All the authors read and approved the final version of the manuscript.

REFERENCIAS BIBLIOGRÁFICAS

1. Andersen PK, Borgan O, Gill RD, Keiding N. Survival from malignant melanoma. R Package boot: Bootstrap R (S-Plus) Functions. 1993. [cited 2021 Aug 17]. Available from: https://stat.ethz.ch/R-manual/R-patched/library/boot/html/melanoma.html
2. Bruce, P, Bruce A, Gedeck P. Practical Statistics for Data Scientists. O'Reilly Media, Inc., 2020, p. 57-78.
3. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria; 2021. [cited 2021 Aug 17]. Available from: https://www.R-project.org/.
4. Shahbaba B. Biostatistics with R: An introduction to statistics through biological data. New York: Springer; 2012. p. 17-79.

Velasco MC, Zhao I. Introduction to Estimation. Metro Ciencia [Internet]. 29 de noviembre de 2021; 29(4):65-72. **https://doi.org/10.47464/MetroCiencia/vol29/4/2021/65-72**