Available online at www.sciencedirect.com
SCIENCE DIRECT
Tourism Management 27 (2006) 42–50
TOURISM MANAGEMENT
www.elsevier.com/locate/tourman

Robust estimators and bootstrap confidence intervals applied to tourism spending

Alfonso Palmer Pol[a,*], Mauricio Beltrán Pascual[b], Pedro Cortiñas Vázquez[c]

[a]Facultad de Psicología, Universidad de las Islas Baleares, Ctra. Valldemossa Km. 7,5, 07122 Palma de Mallorca, Spain
[b]Junta de Castilla y León. Consejería de Hacienda, Dirección General de Estadística, Calle José Cantalapiedra s/n, 47012 Valladolid, Spain
[c]Facultad de Ciencias Económicas y Empresariales, Departamento de Economía Aplicada y Estadística, Universidad Nacional de Educación a Distancia, Paseo Senda del Rey, 11, 28040 Madrid, Spain

Received 19 January 2004; accepted 18 June 2004

Abstract

Statistical research carried out over the past few years evaluating important aspects of tourism has shown that supply and demand for tourism products have risen. This study focuses on various methods of evaluating a fundamental variable of tourist expenditure: average daily expenses per tourist. When analysing this variable, extreme values that invalidate the average location parameter are not uncommon. The presence of skewed values and the asymmetry of distribution justify using alternative methods for parameter estimation. Using data collected from the tourist expenditure survey taken in the Balearic Islands in 2001, this study presents results obtained from different robust location estimators, placing special emphasis on Huber and one-step's *M*-estimators, accompanied by calculating confidence intervals. Additionally, results were obtained by using a resampling method called the bootstrap estimation.

Keywords: Robust estimators; Tourism spending; Bootstrap confidence intervals

1. Introduction

For the Spanish economy in the year 2001, tourism and other directly associated activities accounted for approximately 12.5% of the GNP. In labour terms, they employ 13.5% of the country's total workforce and 21.1% of those working in service industries. Of all the world's countries, Spain is the second most important tourist destination, with a total of 49.5 million inbound tourists in the year 2001.

Within the framework of this situation, one Spanish region stands out particularly from the rest, and it is this region that has been used in our study, given its importance in the field of tourism.

Covering a surface area of 4,968.36 km^2 with a recorded population of 878,627 inhabitants in the year 2001, the Autonomous Community of the Balearic Islands is considered one of the world's exceptional tourist destinations, with visitor numbers placing the region on a par with countries that are international leaders in tourism. In 2001 there were 9.7 million non-resident tourists, of which 8.4 million were foreigners (international tourism), giving a world market share of 1.2%, as can be seen in Table 1.

At the same time, given the actual size of the Balearic Islands, the importance that inbound tourism represents for the region's economy can easily be understood. In the year 2001, revenue from tourism amounted to 5096.2 million €, whilst in 1997 the tourism-added value (TVA) accounted for 21.2% of the GDP of the Balearic Islands'

*Corresponding author. Tel.: +34 971 17 34 32; fax: +34 971 17 31 90.
E-mail addresses: alfonso.palmer@uib.es (A. Palmer Pol), mauriciobeltran@wanadoo.es (M.B. Pascual), pcortinas@cee.uned.es (P.C. Vázquez).

0261-5177/$ - see front matter
doi:10.1016/j.tourman.2004.06.016

Table 1
Market shares by destination

Leading world destinations[a]	International tourist arrivals	
	2001	Market share
World	693	
1. France	76.5	11.0
2. Spain	49.5	7.1
3. United States	45.5	6.6
4. Italy	39.0	5.6
5. China	33.2	4.8
6. United Kingdom	23.4	3.4
7. Mexico	19.8	2.9
8. Canada	19.7	2.8
9. Austria	18.2	2.6
10. Germany	17.9	2.6
11. Hungary	15.3	2.2
12. Poland	15.0	2.2
13. Hong Kong (China)	13.7	2.0
14. Greece	12.8	1.8
15. Portugal	12.1	1.7
Balearic Islands[b]	8.4	1.2

[a] *Source*: TO data for 2001 (in millions).
[b] *Source*: IBAE data for 2001.

economy, whilst the value added of tourism industries (VATI) was estimated as representing 31.2%.[1]

These figures highlight how necessary it is to explore estimation methods that can help determine economic realities, as well as contributing towards the formulation of realistic policies for this key sector of the economy.

The main purpose of the Survey on Tourism Spending and Stays in the Balearic Islands is to determine certain socio-economic characteristics of visiting tourists, compile quantitative information on tourism spending in the Autonomous Community of the Balearic Islands and obtain qualitative data on tourist attitudes.

The methodology used is a personal interview, based on a carefully structured questionnaire. The reporting unit is a tourist spending at least one night in the Balearic Islands. This is a yearly survey, reflecting seasonal activity (the high, mid and low seasons) and each individual island.

The survey is conducted at airports and maritime ports in the Autonomous Community of the Balearic Islands with international and domestic passenger traffic, i.e. Palma de Mallorca, Minorca and Ibiza Airports and the maritime ports of Palma de Mallorca, Minorca, Ibiza and San Antonio. The questionnaires are completed by tourists at the end of their stay in the Balearic Islands.

In the year 2001, a sample of 10,178 individual surveys was collected. The distribution of the interviewees was based on the volume of international and domestic passenger traffic at each of the survey points.

Since this paper focuses on a description and comparison of different estimators for the calculation of the average daily spending variable, to avoid complicating the description, no weighting factor will be used, unlike the original survey analysis, which does so in order to take into account both the nationality of the tourists and the island they visit.

One variable that is clearly of political and social interest is average spending in a tourist destination per person per day. This refers exclusively to expenditure in the destination itself and not to any payment made in the tourist's country of origin. When the data used in this study were being processed, the asymmetrical behaviour of this variable was observed, with numerous atypical values (444 outlying values) and a very pronounced degree of asymmetry (see Fig. 1) with a skewness value of 6.286 (SE = 0.024), and a kurtosis value of 93.09 (SE = 0.049), meaning that the distribution of the average spending variable per person per day was right-skewed and leptokurtic.

This behaviour, observed in Fig. 1 (Stem and Leaf Plot), has a decisive influence on the way the estimation of the total average spending variable is made, and failing to take this into account can lead to incorrect estimations and mistaken conclusions that do not reflect reality.

As Table 2 shows, the mean value of the sample group's average tourism expenditure per person per day is 9719.32 pesetas,[2] with a standard deviation of 8797.39, which indicates a high variation in the variable's different values (with a variance of 77393881). Under such conditions of skewness, the sample mean will be a bad measure of central location to use. In other words, it is not a valid measure for the intended calculation, since it is clearly affected by its outlying (atypical) values in what is obviously a skewed distribution, meaning that the arithmetical mean is a non-resistant measure. This effect is attributable to the circumstances of the variable under observation, which has a minimum fixed limit (a minimum spending level of 0 pesetas) but no pre-established upper limit. In our case, there is a small group of tourists who spend no money at all during their stay (2.56%), whilst the highest value that can be observed in our data set is an expenditure figure of 216,000 pesetas. However, the central 50% of the daily spending values range between 4741.26 and 12,317.86 pesetas.

[1] Montserrat, A., Beltrán, M., Parra, F., & Cortiñas, P. (2001). *Valor Añadido de las Industrias Turísticas y Valor Añadido Turístico en Baleares (The value added of tourism industries and tourism value added in the Balearic Islands.)*. Congress on Tourism Satellite Accounts, Vancouver.

[2] The expenditure figures are shown in pesetas: the official currency when the survey was conducted. The exchange rate for euros is: 1 euro = 166.386 pesetas.

```
 Frequency    Stem & Leaf

   355,00        0 .  00000000000006789&
   361,00        1 .  01233445566677788899
   599,00        2 .  000111222334455556677788899
   688,00        3 .  0000111122233334455556667778888999
   712,00        4 .  0000111122233344455556666777888999
   722,00        5 .  0000001112223333444455566677778888999
   795,00        6 .  00000111222222333344444555666667778888999
   729,00        7 .  0000111122233344445555556667778888999
   708,00        8 .  00001112222333334445556667777888999
   642,00        9 .  00000111222333344455556667788899
   616,00       10 .  000000112233444555666777888999
   544,00       11 .  00111122222333444556667778899
   477,00       12 .  000112233445555566778899
   371,00       13 .  00112233345667789
   263,00       14 .  0012233456789
   253,00       15 .  00123456789
   208,00       16 .  01234566789
   169,00       17 .  01234579&
   156,00       18 .  012345677&
    96,00       19 .  0236&&
    92,00       20 .  0&&&&
    80,00       21 .  28&&&
    57,00       22 .  15&&
    41,00       23 .  5&
   444,00 Extremes    (>=23707)

 Stem width:   1000,00
 Each leaf:      20 case(s)

 & denotes fractional leaves.
```

Fig. 1. Stem and leaf plot.

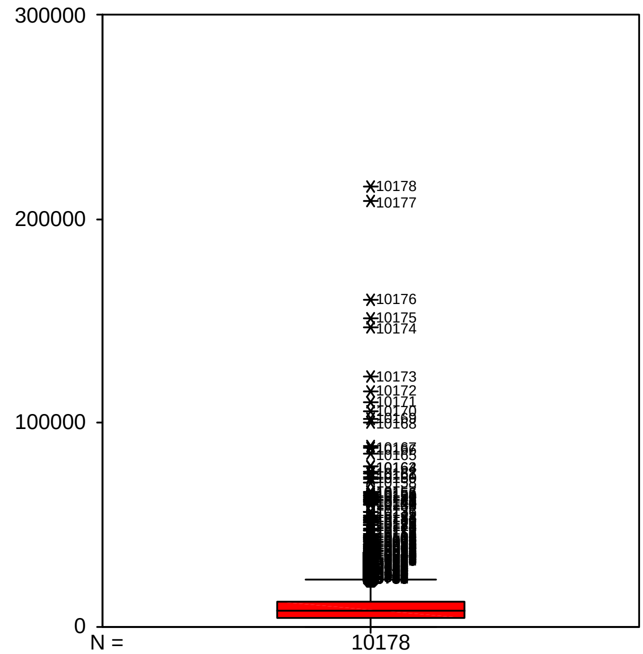

Fig. 2. Box plot of spending in the Balearic Islands per person per day.

Table 2
Mean, standard deviation, standard error, confidence interval

Mean	Standard deviation	Standard error of mean	95% confidence interval of mean
9719.32	8797.39	87.2	9548.39–890.25

The box plot, Fig. 2, shows the distribution of the observed values, detecting, as commented above, the presence of outlying values with a strong influence on the distribution of the average spending variable per person per day.

Due to the behaviour of the variable under analysis, alternative estimators that could avoid the effects of these outliers were deemed necessary. In literature, numerous estimators that fulfil these requirements are contemplated (Andrews et al., 1972; Tukey, 1977; Wilcox, 1997). Among them are the following methods for the calculation of measures of central location:

(a) The first consists of using the median to estimate the central value. The main problem with the median is the fact that its calculation is based on the number of sample observations, with a resulting value that is midway between them. It is considered the simplest robust estimator.
(b) The second is the use of the trimmed mean. This is the mean of the observations which remain after eliminating a certain percentage (α) of the observations at each extreme of the distribution, and its efficiency as an estimator is generally good. Another method uses the winsorised mean, whereby instead of eliminating the α percentage of observations at each extreme they are replaced by the value of the most extreme observation remaining in the distribution.
(c) The third group comprises what are known as robust means. Of these, the *M*-estimators based on maximum-likelihood estimations are used in this study. These involve the calculation of the values' weighted mean in such a way that the weightings decrease as the values move away from the centre of the data set. These calculation methods are relatively valid for distributions with a very high variance, and statistically they are more efficient than the median method, since they include the real values of the observed samples in their calculations together with any influences derived from them. Robust estimators were introduced by Huber (1964) and developed by Hampel (1968), who introduced the concept of the influence function.

At the same time, each of the above methods provides a specific point estimate for the distribution, and it is crucial to provide the said estimator's confidence interval (CI) in order to discover the true measure of central location for the population.

The bootstrap method can be used for this purpose. This is a technique developed by Efron (1979, 1987), and is now a commonly used resampling method. The technique (Efron & Tibshirani, 1993) consists of a simulation whereby a large number of samples *B* of size *n* are removed from the same sample of size *n*, with the replacement of the removed elements. In each of these simulated samples, the value of the statistic of interest is calculated.

With this technique, conclusions on the population from which the data have been obtained can be reached by repeatedly sampling the said data. Each time a bootstrap sample is extracted, a different estimate results.

For example, if $x=(x_1,\ldots,x_n)$ represents the vector of the data and $T_n(x)$ represents the statistic to be calculated, this method consists of extracting a sufficiently large number of random samples with replacements. The number of bootstrap samples to be used varies according to the type of statistics to be estimated. Each one of the samples is called a bootstrap sample,

$$x^* = (x_1^*, x_2^*, \ldots, x_n^*).$$

If $x^{*1}, x^{*2},\ldots, x^{*B}$ represents B bootstrap samples, the values of the statistic under analysis T_n in each of them, $T_n(x^{*1}), T_n(x^{*2}),\ldots, T_n(x^{*B})$, are called bootstrap replications.

If $T_n(\cdot)$ represents the estimated parameter, e.g. the mean, then:

$$T_n(\cdot) = \frac{1}{B}\sum_{b=1}^{B} T_n(x^{*b}).$$

Similarly, the bootstrap estimator of the sample error can be calculated by using the following expression:

$$\hat{\text{se}}_{\text{boot}} = \sqrt{\frac{1}{B}\sum\nolimits_{b=1}^{B}(T_n(x^{*b}) - T_n(\cdot))^2}.$$

Thus, with the bootstrap method, the confidence interval of the parameter estimated from the distribution generated by the bootstrap replications of the said parameter can be calculated.

1.1. Robust estimations of the mean using Huber's M-estimator

As commented above, in this type of study the sample mean has a highly variable behaviour. Consequently, the classic estimator that is normally used is not appropriate, and so calculations must be based on robust estimators. *M*-estimators use an iterative calculation process, whereby an estimate is obtained with each iteration by weighting the observations according to their distance from the core of the data set. Huber's estimator is an *M*-estimator possessing the characteristics of robustness and efficiency. It has a lineal Ψ function in the centre and is constant at the extremes, being the most efficient estimator of those which limit their sensitivity to large errors. Likewise, it reaches the maximum possible breakage point.

The relative distance between each observation and the centre of the distribution is expressed thus:

$$u_i = \frac{y_i - T}{s},$$

where s is the normalised median absolute deviation from the median (MADN), which is a robust measurement of dispersion, defined as the median of the absolute deviations from each observation to the median and divided by 0.6745, and T, a measure of central location (in the first iteration, the median).

The weighting function, which is dependent on a single constant k, is expressed thus:

$$\omega(u_i) = \begin{cases} 1, & |u_i| \leqslant k, \\ \frac{k}{u_i}\,\text{sgn}(u_i), & |u_i| > k. \end{cases}$$

For weighting constant k the value 1.28 has been used. This corresponds to the value of percentile 90 in a normal standard distribution. As the weighting function shows, Huber's *M*-estimator only weights those observations situated at a relative distance from the centre of the distribution with a value higher than constant k. For observations closer to the centre, the original value is retained.

2. Results

In order to ensure a clear description of the average spending variable per person per day in the Balearic Islands, we will now outline the results obtained for the different robust techniques mentioned above.

2.1. Estimations of the median

For the original data set, the classic median (i.e. quantile 50 of the data distribution) gave a figure of

$$M = 8177.66.$$

This value has been criticised, as it uses very little information contained in the data set, since only the central value intervenes in its calculation, in the event of an odd number of observations, or the two central values, in the case of an even number of observations. Table 2 shows the value of the Harrell–Davis estimator of the median (Harrell & Davis, 1982). Unlike the classic median value, this estimator uses all the observations, weighting them according to a beta distribution.

Thus, the Harrell–Davis estimator is obtained from the expression:

$$M = \sum_{i=1}^{n} \omega_i X_{(i)},$$

where the weightings of the order statistics are dependent on the cumulative probability of two possible values for variable Y, which follows a beta distribution with parameters $a=b=0.5(n+1)-1$:

$$\omega_i = P\left(\frac{i-1}{n} \leqslant Y \leqslant \frac{i}{n}\right).$$

All measures of central location must be accompanied by a coefficient of variation. A standard error estimator (EE) for the median is the Maritz–Jarrett error estimator (Maritz & Jarrett, 1978), which also uses beta-weighted observations. It is expressed thus:

$$\mathrm{EE}(M) = \sqrt{C_2 - C_1^2},$$

where

$$C_k = \sum_{i=1}^{n} \omega_i X_{(i)}^k.$$

Finally, when a location is being described, an interval estimation of the said coefficient is needed, providing a certain degree of confidence that the value of the corresponding parameter will actually fall within the limits obtained. Thus, in Table 3 the 95% CI of the Harrell–Davis median is provided, using the Maritz–Jarrett standard error.

The similarity can be observed between the value of the median achieved via classic methods (= 8177.66) and that obtained by the Harrell–Davis technique (= 8166.02).

Another way to obtain both the standard error of the median and its 95% CI is to use bootstrap resampling techniques, in this case $B=100$ bootstrap samples, whose results are shown in Table 4.

In this case, it can be seen that the bootstrap standard error of the median is slightly higher than the value of the Maritz–Jarrett standard error. This implies that the bootstrap CI is wider than the corresponding CI obtained with Maritz–Jarrett procedures.

Table 3
HD estimation of the median, MJ standard error and its 95% CI

Harrell–Davis estimator of the median	Maritz-Jarrett estimate of the standard error of the median	95% confidence interval of the median
8166.02	69.98	8038.53–8312.85

Table 4
Bootstrap estimator of the standard error

Harrell–Davis estimator of the median	Bootstrap estimator of the standard error of the median	95% bootstrap confidence interval
8166.02	72.99	8019.27–8312.76

2.2. The winsorised mean

One solution to the problem of widely differing values is to winsorise the data. This entails replacing a percentage of outlying values with the value immediately prior to them. The winsorised mean is obtained by winsorising a certain proportion of data (α) from each end of the distribution:

$$W(\alpha) = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

where the winsorised values of the variable take the following pattern:

$$X_i = \begin{cases} X_{\gamma+1}, & i = 1, \ldots, \gamma, \\ X_i, & i = \gamma+1, \ldots, n-\gamma, \\ X_{n-\gamma}, & i = n-\gamma-1, \ldots, n, \end{cases}$$

where γ is the whole part of α_n.

The standard error of a winsorised mean is obtained by calculating the square root of the variance in the winsorised values:

$$\mathrm{EE}[W(\alpha)] = \sqrt{\frac{\sum_{i=1}^{n}[X_i - W(\alpha)]^2}{n-1}}$$

Table 5 shows the values of the winsorised mean for different percentages of winsorised data, together with the values of their associated standard errors.

The table shows how the value of the winsorised mean drops as greater percentages of data are winsorised, due to the distribution's right-skewed shape. This means that each time the data are winsorised, the variable's high values are winsorised, with the replacement of those values that tended to increase the mean.

One problem with winsorised means is the high degree of variability they entail (see the standard errors in Table 5) due to the considerable accumulation of data at the winsorised extremes.

2.3. The trimmed mean

Instead of winsorising the data, a more drastic solution consists of eliminating the same percentage of observations α from each end of the distribution and obtaining the mean value of the remaining ones. This is

Table 5
Winsorised means

% Winsorised	Winsorised mean	Standard error
0.05	9145.42	5645.33
0.10	8867.79	4836.25
0.15	8663.49	4220.53
0.20	8492.20	3600.69
0.25	8401.66	3074.66
0.30	8324.75	2516.26

Table 6
Trimmed means

% Trimmed	Trimmed mean	Effective sample size	Standard error	95% confidence interval
0.05	8826.87	9162	62.18	8704.99–8948.74
0.10	8572.66	8144	59.92	8455.19–8690.12
0.15	8422.15	7126	59.76	8304.99–8539.30
0.20	8332.24	6108	59.48	8215.63–8448.85
0.25	8273.81	5090	60.95	8154.32–8393.31
0.30	8232.73	4072	62.35	8110.49–8354.98

the trimmed mean, achieved using the following expression:

$$T(\alpha) = \frac{X_{\gamma+1} + \cdots + X_{n-\gamma-1}}{n(1-2\alpha)}.$$

The standard error of a trimmed mean is obtained using the standard error of the corresponding winsorised mean, as follows:

$$\mathrm{EE}[T(\alpha)] = \frac{\mathrm{EE}[W(\alpha)]}{(1-2\alpha)\sqrt{n}}.$$

Table 6 shows the trimmed mean values of different percentages of trimmed data. Each time the data are trimmed, the new effective sample size, the standard error of each trimmed mean and the 95% CI are all shown.

As greater percentages of data are trimmed, the value of the trimmed mean can be seen to fall, thanks to the distribution's right-skewed shape. This means that, with each trim, what are eliminated are the variable's high values (i.e. the ones that inflated the value of the mean).

Table 6 shows that trimming the data eliminates the variability problem we came across when they were winsorised. As a result, the standard errors of the trimmed means are small.

Using the bootstrap procedure with $B = 599$ bootstrap samples, the CI of a trimmed mean can be obtained. Table 7 shows the 95% bootstrap CI for the trimmed means of different percentages of trimmed data.

The 25% trimmed mean is called the *midmean* as it represents the mean of the central 50% of the observations. In our case, this *midmean* value is 8273.81.

The question users ask is: What percentage of trimmed data should be used?

For a general guideline, we can look at Figs. 3a–d, which provide the distributions of daily spending after trimming 0%, 5%, 10%, and 15% of the data respectively.

When the data are trimmed by 5% (Fig. 3b), it can be observed that the shape of the distribution does not change in comparison with the original (Fig. 3a), but it can also be seen that there are still outlying values on the right of the distribution. Consequently, the trimmed mean will still be influenced by these values, which must be removed.

Table 7
Bootstrap confidence interval for the trimmed mean

% Trimmed	95% bootstrap confidence interval for the trimmed mean
0.05	8705.49–8951.83
0.10	8456.92–8690.71
0.15	8299.93–8538.11
0.20	8207.13–8448.67
0.25	8145.32–8391.63
0.30	8104.40–8352.48

When the data are trimmed by 10% (Fig. 3c), no outliers can now be seen. As a result, the trimmed mean will be resistant. Furthermore, the distribution of the data still has a shape similar to the original one, although it is trimmed on the right-hand side.

After trimming 15% of the data (Fig. 3d), it can be seen that the resulting distribution has a very homogenous shape, meaning that too many values have been removed from the distribution, causing it to lose its original form.

From these graphs and their interpretations, we can get an idea of the percentage of data that must be trimmed if the estimator is to be resistant and not lose relevant information. In our case, a 10% trim is the best choice.

2.4. Huber's M-estimator

The results for Huber's *M*-estimator will now be described. In the calculations, the normalised MAD has been used as a robust measurement of dispersion: MADN = 5511.225.

Table 8 shows the value of the *M*-estimator. Likewise, it also provides two estimates of the standard error of the *M*-estimator. The first value was obtained by using the influence function of the *M*-estimator, whilst the second was obtained by using the bootstrap procedure with $B = 100$ bootstrap samples. Finally, the 95% bootstrap confidence interval is provided for the value of the *M*-estimator.

```
Stem &  Leaf

   0 .  000000000000006789&
   1 .  0123344556667788899
   2 .  00011122233445555667778889
   3 .  000011112223333445555666777888899
   4 .  000011112223334445555666777888999
   5 .  000000111222333344445556667777888899
   6 .  0000011122222233334444455566666777888899
   7 .  0000111122233344445555566677788889
   8 .  00001112222333334445556667777888999
   9 .  00000111222333344455556667788899
  10 .  0000001122334445556667778889
  11 .  00111222233344455666777889
  12 .  0001122334455556677889
  13 .  00112233345667789
  14 .  0012233456789
  15 .  00123456789
  16 .  01234566789
  17 .  01234579&
  18 .  012345677&
  19 .  0236&&
  20 .  0&&&&
  21 .  28&&&
  22 .  15&&
  23 .  5&
444,00 Extremes    (>=23707)
```

(a) Trimmed by 0%

```
Stem &  Leaf

   1 .  56667788899
   2 .  00011122233445555667778889
   3 .  000011112223333445555666777888899
   4 .  000011112223334445555666777888999
   5 .  000000111222333344445556667777888899
   6 .  0000011122222233334444455566666777888899
   7 .  0000111122233344445555566677788889
   8 .  00001112222333334445556667777888999
   9 .  00000111222333344455556667788899
  10 .  0000001122334445556667778889
  11 .  00111222233344455666777889
  12 .  0001122334455556677889
  13 .  00112233345667789
  14 .  0012233456789
  15 .  00123456789
  16 .  01234566789
  17 .  01234579&
  18 .  012345677&
  19 .  0236&&
  20 .  0&&&&
  21 .  28&&&
37,00 Extremes    (>=22013)
```

(b) Trimmed by 5%

```
Stem &  Leaf

   2 .  55556677788899
   3 .  000011112223333445555666777888899
   4 .  000011112223334445555666777888999
   5 .  000000111222333344445556667777888899
   6 .  0000011122222233334444455566666777888899
   7 .  0000111122233344445555566677788889
   8 .  00001112222333334445556667777888999
   9 .  00000111222333344455556667788899
  10 .  0000001122334445556667778889
  11 .  00111222233344455666777889
  12 .  0001122334455556677889
  13 .  00112233345667789
  14 .  0012233456789
  15 .  00123456789
  16 .  01234566789
  17 .  012345
```

(c) Trimmed by 10%

```
Stem &  Leaf

   3 .  22333333334444455555556666667777777888888899999
   4 .  0000000011111112222222333333344444455555555566666667777777888888899999
   5 .  0000000000001111122222233333333444444455555556666666777777788888888899999
   6 .  0000000001111112222222222222333333344444444445555556666666666677777888888888899999
   7 .  000000011111112222222333334444444455555555556666667777777888888888899999
   8 .  000000001111112222222223333333333444445555556666666777777778888888899999
   9 .  0000000001111112222223333333344444455555555666666777888888899999
  10 .  000000000001111122222333334444455555566666667777778888888899999
  11 .  00000111111222222223333334444445555666666777778888889999
  12 .  000000111112222233334445555555666777778888899
  13 .  00001111222233333344555666677778899
  14 .  00001122233333445566678889
  15 .  0000111&
```

(d) Trimmed by 15%

Fig. 3. Trimmed by: (a) 0%; (b) 5%; (c) 10%; and (d) 15%.

Table 8
Huber's *M*-estimator ($k = 1.28$)

Huber's *M*-estimator	Standard error	Bootstrap standard error	95% bootstrap confidence interval
8557.28	59.42	61.34	8446.64–8658.47

2.5. The one-step estimator

One problem with the trimmed mean is the fact that the same proportion of observations is trimmed from each end of the distribution of ordered values, even if the outliers are only on one side of the distribution.

One way of overcoming this drawback is to determine empirically the percentage of data that must be trimmed from each side of the distribution. This is precisely what a one-step estimator does.

The one-step estimator used follows the methodology of the Huber estimator, meaning that it eliminates those observations that lie outside the thresholds -1.28 and 1.28: the value of Huber's constant. It is this that then determines the number of observations to be eliminated from each end of the distribution.

This estimator is the result of the first iteration of the *M*-estimator. Consequently, Huber's one-step estimator (Staudte & Sheather, 1990) is expressed thus:

$$\mathrm{os}_{\mathrm{Huber}} = \frac{1.28 \times (\mathrm{MADN}) \times (i_2 - i_1) + \sum_{i=i_1+1}^{n-i_2} X_{(i)}}{n - i_1 - i_2},$$

where i_1 and i_2 represent the number of values to be eliminated, as obtained below:

i_1 represents the number of observations that complies with $(X_i - M)/\mathrm{MADN} < -1.28$, and i_2 the number that complies with $(X_i - M)/\mathrm{MADN} > 1.28$.

Table 9 shows the value of the one-step estimator for the data set under analysis, together with its 95% bootstrap confidence interval, calculated with $B = 399$ bootstrap samples.

Table 9
One-step estimator

Huber's one-step estimator	95% bootstrap confidence interval
8557.00	8446.84–8658.33

3. Conclusions

Clearly, if one is to obtain realistic information on a variable like average spending per person per day, the best-suited statistical tool must be chosen, since realistic initiatives can only be implemented on the basis of realistic results.

In a recent, as yet unpublished[3] bibliometric study, based on 12 relevant tourism journals covering the period 1998–2002 and a total of 1790 articles, not one single article was found that used robust or bootstrap methods. This indicates the existence of a large gap in tourism research, from the perspective of descriptions of variables with an asymmetric behaviour, as always occurs if, for example, a region's tourism spending is analysed.

As can be observed from the results achieved using different mean parameter estimation methods, shown in Table 10, there are serious discrepancies between the result of the arithmetical mean and those of different robust estimators, due to the influence of the data set's outlying values.

As seen previously, the median values obtained with the classic method or the Harrell–Davis method are practically the same, with a difference of 11.64 pesetas. In turn, they are both very different from the value of the arithmetical mean, showing a difference of 1550 pesetas in the estimation of average spending per person per day. However, in the case of the median, because the distribution is right-skewed, it gives a lower value than the mean, tending to underestimate the distribution's measure of central location.

It can also be seen that the one-step estimator, based on Huber's methodology, is practically identical to Huber's *M*-estimator. It is also virtually the same as the 10% trimmed mean (the most appropriate trim for this data set, as mentioned above). Compared to the arithmetic mean, all these values show a difference of about 1160 pesetas in the estimation of average daily spending per person per day.

Thus, while all the robust estimators are relatively far removed from the value given by the arithmetic mean, with a difference of between 1445 (midmean) and 1162 (Huber) pesetas per day, when compared with the median they show a difference of between 107 and 406 pesetas.

[3]Palmer, A., Sesé A., & Montaño J.J. (2004). *Tourism and statistics: bibliometric study 1998–2002*. (currently being revised).

Table 10
Central value estimation results, using different methods

	Estimation of the central value
Sample mean	9719.32
Harrell–Davis median	8166.02
10% trimmed mean	8572.66
Midmean	8273.81
Huber's *M*-estimator	8557.28
One-step estimator	8557.00

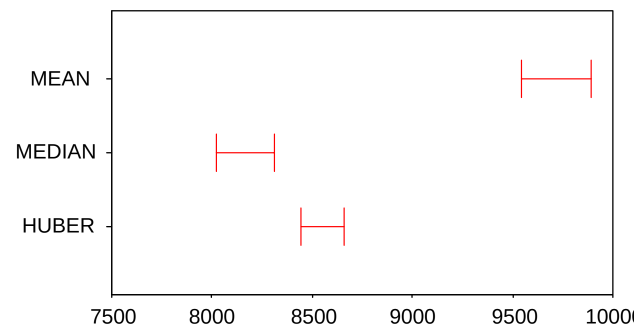

Fig. 4. Confidence interval of the mean, median and Huber estimator.

In summary, the arithmetic mean gives an excessively high figure for tourist spending of 9719.32 pesetas, given the positive skewness of the variable's distribution. In a situation such as this, another more realistic estimation of spending must be found that is better adjusted to the data set under analysis. This can be obtained by using different robust estimators. In this case, the figure obtained ranges somewhere between 8166.02 pesetas (the median value) and 8572.66 pesetas (the value of the 10% trimmed mean, with a Huber's *M*-estimator of 8557). This is between 12% and 16% less than the spending figure given by the arithmetical mean.

Next to the value given by each estimator, it is also important to provide the parameter's confidence interval. This can be obtained by classic methods or by the bootstrap resampling technique.

Nowadays there are statistical programmes (S-Plus, R) capable of using robust estimations and bootstrap methods, as well as ready-made macros with all these capacities. In addition, these platforms allow users to create the statistical applications they require. Consequently, these robust techniques can be used by any researcher working in the field of tourism, and it is hoped that the results of the study presented here (obtained using the *R* statistical system) will help to increase people's awareness of their importance.

The difference between the arithmetic mean (normally used as a distribution's measure of central location), the median (used typically for asymmetric distributions) and Huber's *M*-estimator (as an example of a robust measurement) can clearly be seen in Fig. 4.

The confidence interval for the mean (9548.39–9890.25) provides a set of very high values for average spending per person per day in the Balearic Islands during the year 2001, whilst that of the median (8019.27–8312.76) falls below the said values. Huber's confidence value (8446.64–8658.47) is situated some way between them, overcoming the problems the two former systems present and giving values that are more in line with the reality.

As a numerical example, if one uses the arithmetic mean as a measure of daily spending then for example, the 10,178 tourists interviewed spent almost 99 million pesetas per day, or 832 million pesetas if the median is used. On the other hand, if one uses a robust measurement of spending, like Huber's *M*-estimator, the more realistic conclusion is that 87 million pesetas were spent per day. As illustrated, there is a difference of plus 12 million pesetas daily expenditure, signifying a difference of 12.1% between the more realistic spending figure and the arithmetic mean, and a difference of minus 3.8 million pesetas between the Huber and the median. In assessing how relevant it is to use a carefully selected, appropriate measure for estimating tourism spending, this difference speaks for itself.

The final conclusion is that if there is asymmetry, a good estimator method is Huber's *M*-estimator, which should be accompanied with its corresponding bootstrap confidence interval.

References

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972). *Robust estimates of location: survey and advances*. Princeton: Princeton University Press.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*, 171–185.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. London: Chapman & Hall.

Hampel, F.R. (1968). *Contributions to the theory of robust estimation*. Ph.D. dissertation, University of California, Berkeley.

Harrell, F. E., & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika*, *69*, 635–640.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, *35*, 73–101.

Maritz, J. S., & Jarrett, R. G. (1978). A note on estimating the variance of the simple median. *Journal of the American Statistical Association*, *73*, 194–196.

Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. New York: Wiley.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. New York: Academic Press.