**A5Wa Resampling with Replacement (Bootstrapping)**

Because the CLT only applies to sums and means, when statisticians wanted to put a confidence interval on a median, they were out of luck. That is, until the bootstrap was invented.

Resampling procedures are based on the assumption that the underlying population distribution is the same as a given sample. The approach is to create a large number of samples from this pseudo-population using the techniques described in sampling and then draw some conclusions from some statistic (mean, median, etc.) of the sample.

Suppose you had sampled from a distribution you knew was skewed (like incomes or house prices or something like that) and so you are trying to estimate the median of that population. You sample the data and you take the median of the sample as an estimate of the median of the population. You'd like to put a confidence interval around that estimate, but how?

Method steps:

* Take your data as the "estimate" of the population.
* Draw thousands of resamples of equal size from the data.
* Compute the median of each.
* The distributions of medians are an estimate of the sampling distribution of the median on your population.
* Use the appropriate interval from your distribution. For example, for a 90 percent confidence interval, use the 5th and 95 percentiles from your distribution.

This is the real reason that statisticians are excited by the bootstrap. They can now give confidence intervals on all kinds of things that they couldn't before, such as:

* Medians.
* Modes.
* Correlations.

Resampling with replacement is called bootstrapping which will be described below for a simple one sample example/

|  |
| --- |
| **Stop random numbers changing by setting Excel Calculation Options**  If you wish to stop the Excel spreadsheet random number re-generating new values when opening the file and/or modified key cells, then you can stop the spreadsheet re-calculating as follows – this is useful if you want to screenshot solutions for your reports and wish to stop the values changing:   1. Open the Excel workbook you want to keep the random numbers from changing. 2. Click Formula tab > Calculation Options > check Manual in the drop-down list.     Figure 1  With this method, all the formulas in this workbook will not automatically calculate any more.  Re-member, to switch automatic re-calculating when you have completed the task. You can always re-calculate by pressing the **F9 key** on your computer keyboard. |

|  |
| --- |
| **Array formulas**  Array formulas are often referred to as CSE (Ctrl + Shift + Enter) formulas because instead of just pressing Enter, you press Ctrl + Shift + Enter to complete the formula. Consider Example 1, we have an array formula to calculate the frequencies in Cells AA10:AA19. To do this we highlight AA10:AA19 and then enter the formula:  =FREQUENCY (X4: X2003, Z10:Z19)  Now press (Ctrl + Shift + Enter) and the frequencies will be input within this range. You will see the frequency formula looks like {=FREQUENCY (X4:X2003, Z10:Z19). |

**Example 1 - One sample case for the median**

A company claims that they offer a therapy to reduce memory loss for senile patients. To test this claim they take a sample of 20 patients and test each patient’s percentage of memory loss, with the results given in Table 1. Calculate a 95% confidence interval around the median for the memory loss program.

Suppose that we would like to calculate a confidence interval for the median. Since there are no standard statistical tests for such confidence intervals, we approach the problem via **bootstrapping.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Subject | Memory | Subject | Memory | Subject | Memory | Subject | Memory |
| 1 | 3 | 6 | 6 | 11 | 10 | 16 | 12 |
| 2 | 4 | 7 | 7 | 12 | 11 | 17 | 12 |
| 3 | 4 | 8 | 7 | 13 | 11 | 18 | 13 |
| 4 | 6 | 9 | 9 | 14 | 11 | 19 | 13 |
| 5 | 6 | 10 | 9 | 15 | 11 | 20 | 15 |

Table 1

Excel solution

|  |  |
| --- | --- |
| Figure 2 represents the data and calculation of the mean and median values.    Figure 2 | The sample has a mean of 9 and a median of 9.5.  If the date was interval/scale then we could use the mean as the measure of average and then use either the normal or Student’s t distributions to construct a suitable confidence interval.  Given the data is ordinal then we should not use the mean as the measure of average but the median. We have no equivalent statistical method to calculate the confidence interval for the median.  Fortunately, we can employ a bootstrapping method to randomly sample with replacement from the data set and use these samples to calculate a point estimate for the median and construct a confidence interval.  We treat the sample as the population and draw 2,000 samples of size 20 (the same size as the original sample) with replacement. Referring to Figure 2, range D4:W4 represents the first sample, D5:W5 the second, etc. Each element in each sample is selected using the following function:  =INDEX ($B$4: $B$23, RANDBETWEEN (1, 20)) |

Steps:

1. Enter this Excel formula into Cell D4 and copy across to Cell W4.
2. Now copy formula D4:W4 down to D2003:W2003 (this gives 2000 samples)
3. Calculate the Median in Cell X4 = Median (D4:W4)
4. Copy formula down from X4 to X2004 (this creates 2000 median values)

Figure 3 contains the first 27 sampled values for this data set out of 2000 possible sample value. Each sample is in a row, starting from D4:W4, D5:W5, ……., D2003:W2003. The median is calculated using the Excel median () function with the first row given via =median (D4:W4). This formula is then copied down from X4 to X2003.

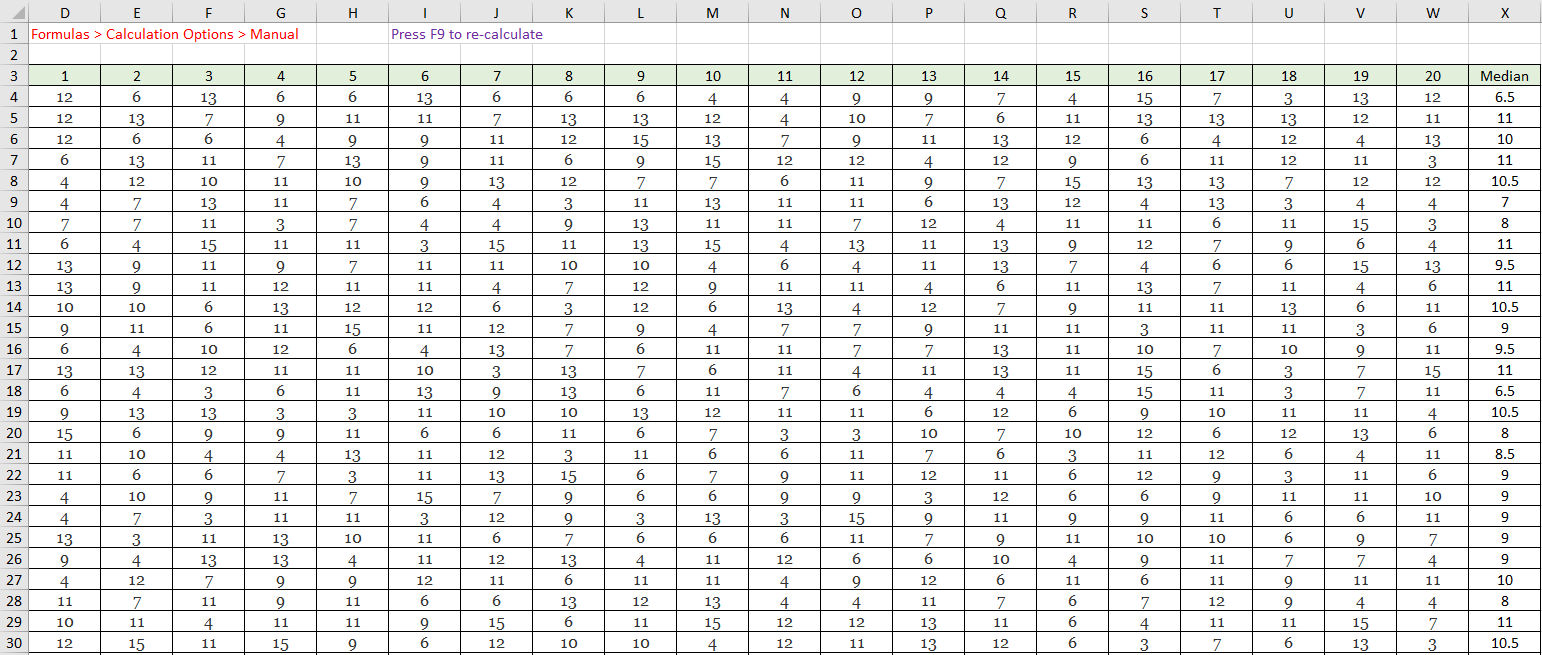


Figure 3

Next, we plot the distribution of the medians (i.e. range X4: X2003) in a histogram using Excel’s Histogram data analysis tool (or Excel’s charting capability), augmented with percentage and cumulative % columns. The results are shown in Figures 4 - 7.

Data > Data Analysis > Histogram

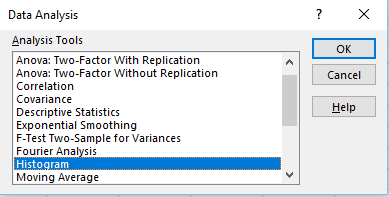


Figure 4

Click OK

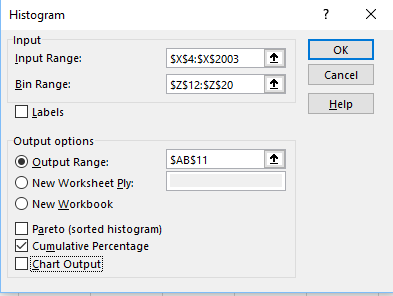


Figure 5

Click OK

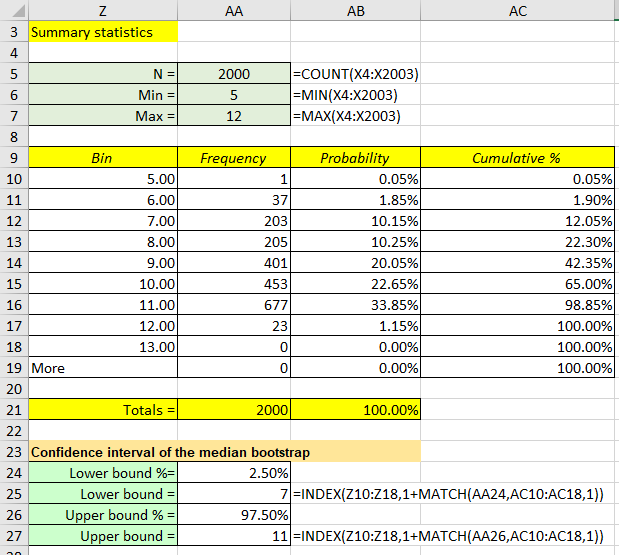


Figure 6

Create Excel histogram (bar chart)

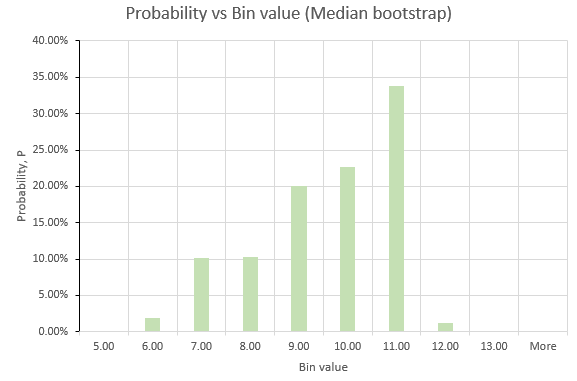


Figure 7

From Figure 6:

* From the Cumulative % column we observe that the 2.5th percentile value is when the Bin value is 7. This implies that the 2.5th percentile value is 7.
* From the Cumulative % column we observe that the 97.5th percentile value is when the Bin value is 11. This implies that the 97.5th percentile value is 11.

Thus, we have a sample median of 9.5 with a 95% confidence interval (7, 11) that contains the median.

**Example 2 - One sample case for the mean, median**

Suppose we have 30 sampled data values collected from a questionnaire as illustrated in Figure 8.

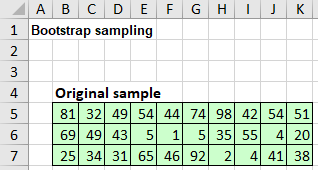


Figure 8

We can now create 200 samples of size 10

In Cell B10, enter the formula:

=INDEX(sample,ROWS(sample)\*RAND()+1,COLUMNS(sample)\*RAND()+1)

1. Copy the formula across from B10:K10.
2. Copy the formula down from B10:K10 to B209:K209

This gives us 200 samples with each sample of size 10.

Figure 9, illustrates the first 7 samples:

* B10:K109
* B11:K11
* B16:K16.

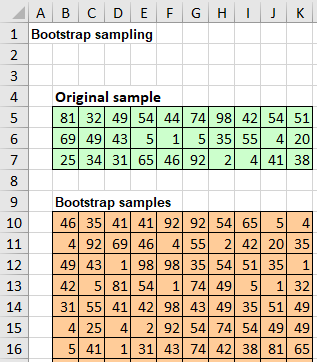


Figure 9

Calculate the descriptive statistics mean and median for each sample as illustrated in Figure 10.

Cell M11, mean = AVERAGE (B10: K10), copy formula M10: M209

Cell N11, mean = MEDIAN (B10: K10), copy formula N10: N209

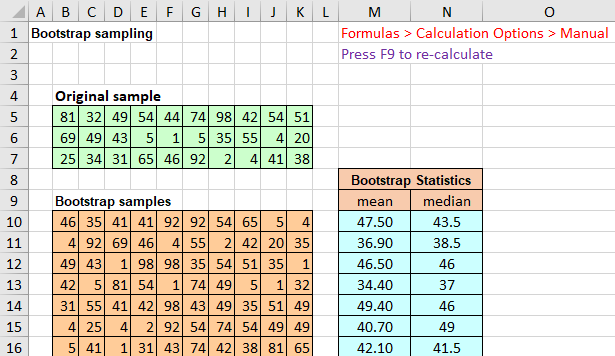


Figure 10

Create a frequency table – including: frequency, probability, cumulative frequency.

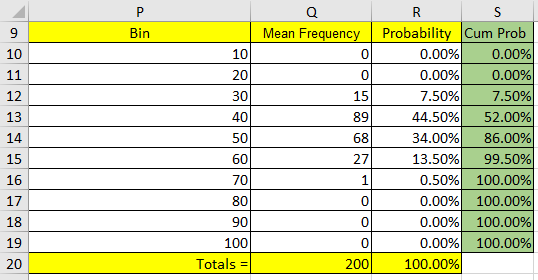


Figure 11

* The frequency is an array calculation for cells Q10:Q19:

{=FREQUENCY (M10: M209, P10: P19).

* Total frequency in Cell Q20: =SUM (Q10: Q19).
* The probability in cells R10: =Q10/$Q$20. Copy formula down from R10:R19.
* Total probability in cell R20: =SUM (R10:R19).
* Calculate cumulative probability in cells S10:S19. In cell S10: = R10. In cell S11: =S10+R11. Copy formula down from S11:S19.

Figure 12, illustrates the bar chart for the probability against the mean bootstrap values (you could repeat this for the median bootstrap values).

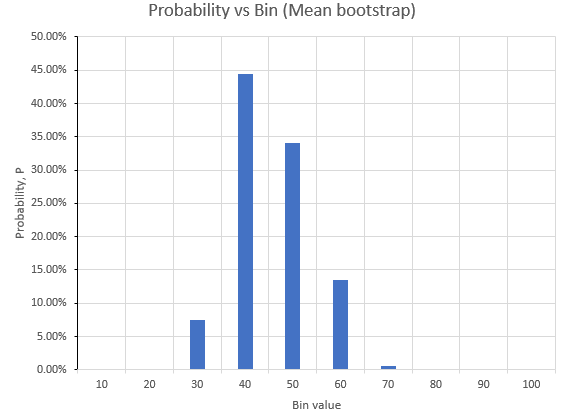


Figure 12

Figure 13, illustrates the point estimates and confidence intervals for the mean and median bootstrap values.

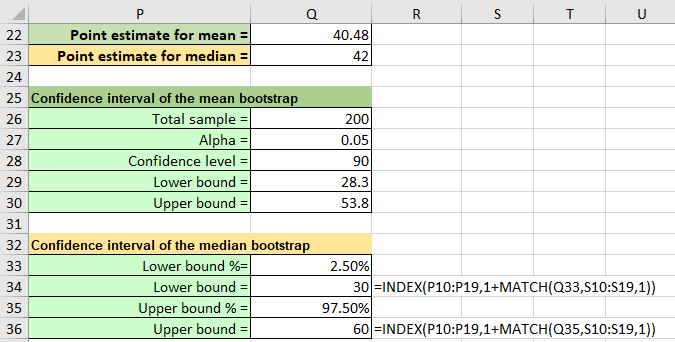


Figure 13

Point estimates

* Point estimate of the mean in cell Q22: = AVERAGE (M10:M209).
* Point estimate of the median in cell Q23: = MEDIAN (N10:N209)

Confidence interval of the mean bootstrap

* Total sample in cell Q26: =COUNT (M10:M209).
* Alpha in cell Q27: = 0.05
* Confidence level in cell Q28: =100\*(1-2\*Q27)
* Lower bound in cell Q29: =SMALL (M10: M209, Q27\*Q26)
* Upper bound in cell Q30: =SMALL (M10: M209, Q26\*(1-Q27))

Confidence interval of the median bootstrap

* Lower bound % in cell Q33: 2.5%
* Lower bound in cell Q34: =INDEX (P10:P19,1+MATCH (Q33, S10:S19,1))
* Upper bound % in cell Q35: 97.5%
* Upper bound in cell Q36: =INDEX (P10:P19,1+MATCH (Q35, S10:S19,1))

For the population mean:

* Point estimate for the mean = 40.48
* Confidence interval for the mean = 28.3, 53.8

For the population median:

* Point estimate for the median = 42
* Confidence interval for the median = 30, 60

**Overall conclusions**

In summary, using a bootstrapping method you can use a sample taken from a population and estimate point estimates for the population parameters (mean, median, standard deviation, semi-interquartile range, etc), distribution shapes for these statistics, and confidence intervals for the population parameters.

In contrast to Monte Carlo simulation which you need to know the distribution of the population, in Bootstrap sampling you also create your own data (based upon observation) without knowing the distribution of the population. All you need is the sample data.