HOW MANY TAXIS?

JOHN GOEBEL & DAN TEAGUE

You are standing in the rain trying to hail a cab in a large city. While waiting, seven cabs pass by that already have a passenger. The numbers on the cabs are 405, 73, 280, 179, 440, 301, 218.

Suppose you want to estimate the number of taxis in the city while you are waiting. Assuming that the taxis are numbered consecutively from 1 to *N* and all are still in service, how can you use the observed numbers to estimate *N*, the total number of taxis in the city?

How many taxis do you think there are? How can you test your method for estimating *N*?

Possible Solutions:

Student solutions will vary according to their mathematical background. Students in a statistics course will more likely use techniques from that course, but our experience has been that the more creative solutions often come from those students who don't have a strong statistical background. Students who know a particular technique often use that technique without thinking further. If you don't know the technique, you have to think more deeply about the problem and often come up with a "better" solution as a result. The following solutions have been developed by our students.

Twice the Mean: The average of a list of consecutive integers is in the middle of the list. If we find the average of the known cab numbers, we can double it to get to the end of the list. From a statistical point of view, the mean of a sample estimates the mean of the population. The average of the known taxi numbers is 270.7. So our estimate of *N* is $2(270.7) \approx 541$.

Twice the Median less 1: The median of the sample of taxis is 280. In a uniform distribution, for example, (1, 2, 3, 4, 5), the largest value is twice the median minus 1. Using this method, we estimate the number of taxis as $2(280) - 1 = 559$. Sometimes students will subtract 1 from the solution based on means above, but, for some reason with our students, the subtraction is more likely to happen with the median.

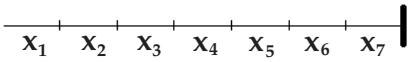
Median Plus IQR: Students who have studied some data analysis know about the Inter-Quartile Range (IQR) associated with boxplots. This is the distance between the location of the smallest 25% of the data and the largest 25% of the data. This interval contains the middle 50% of the data. Students argue that if you find the median and add the IQR, that will take you to the upper extreme. The median for the sample of cabs is 280. The first quartile is the median of the smaller half of the data. The first quartile is 179. The third quartile is the median of the larger half of the data; in this case 405. The IQR is $405 - 179 = 226$. So the median plus the IQR gives an estimate of $280 + 226 = 506$ taxis in the city.

1.5 IQR Solution: A uniform distribution would not have any outliers. Students studying statistics will know that an outlier is conventionally defined as 1.5 times the Inter-Quartile Range beyond the 3rd Quartile. The third quartile is 405 and the IQR is 226. In this sample, an outlier would be beyond $405 + 1.5(226) = 744$. We estimate the number of cabs at 744.

Symmetric Range Solution: The largest in the sample should be approximately the same distance from the maximum value as is the smallest in the sample is from the minimum value. The minimum is 1 and the smallest in the sample is 73, so the largest in the sample 440 should be approximately 72 away from the maximum. This method gives an estimate of 512 taxis.

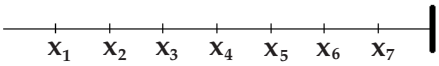
Mean Gap Size Solution: A random sample from a list of consecutive integers should be spread evenly along the number line. We know the first taxi number is 1, so the gaps between successive known cab numbers are 71, 107, 39, 62, 21, 104, and 35. We want to use these gap sizes to predict the final gap, from 440 to the end of the list. The mean gap size is 62.7. If we add the mean gap size to the maximum known value, we have an estimate of the largest value in the list. Our prediction is $440 + 63 = 503$ taxis.

Percentile Solution: The 7 numbers in the sample each represent $\frac{1}{7}$ or 14.3% of N.



If we associate each number with the middle of its 14.3%, we find that 440 is the 93rd percentile, or 93% of the maximum. So, $440 = 0.93x$. Then we expect to have $\frac{440}{0.93} = 473$ cabs.

$\frac{n+1}{n}$ **Max Solution:** The 7 numbers divide the number line from 1 to N into 8 regions.



The largest number, 440, is $\frac{7}{8}$ of the distance from the beginning to the end of the number line, so $440 = \frac{7}{8}x$. The largest number should be $\frac{8}{7} \cdot 440 = 503$.

We have 8 different methods for estimating the number of cabs, yielding values that vary from a low of 489 to a high of 706. Which method works the best? (See **Table 1.**)

Assessing the Solutions

In practice, of course, we have no way to know which of the estimates is the best, since we don't know the true value of N. Without knowing the actual number of taxis, how can we decide which of the methods is most appropriate? One way to analyze problems like this is through simulation. The modeling assumption for the problem is that the 7 numbers we have observed are a random selection from the integers 1 to N. We can fix a value of N, say $N = 500$, and repeatedly select 7 numbers at random from 1 to N, compute all of the different measures, compare the estimates to 500, and pool the results of several hundred trials. We will do a few by hand to illustrate the procedure and then give a calculator program that will automate the process. What is the typical error for each method? Will some consistently overestimate while others underestimate?

Using my calculator, I selected 7 integers at random from 1 to 500. In the first sample, I got 210, 311, 71, 191, 440, 418, and 417. In the second I got 458, 5, 43, 124, 145, 69, and 462. Let's see how the different measures estimate the largest value, which we know to be 500. (See **Table 2.**)

Notice that all but one of the methods overestimated the true value of N. Is this generally going to be the case? (See **Table 3.**)

It seems pretty clear, even after only two trials, that there are problems with the first two methods. It is possible, even likely, that the estimate based on twice the mean or twice the median is actually less than one of the numbers in the sample. In the second trial, we have estimates of 373 and 247 for N and we have known values of 458 and 462! If you happen to get a few small values in your sample, you can produce an estimate that is contradicted by the sample itself. The Outlier Method overestimated the true value of N by a large amount both times. Will it do so consistently?

Based on our two samples, it appears that the average gap size, the percentile techniques, and the $\frac{n+1}{n}$ method produced the best estimates. The symmetric range also gave reasonable estimates, though more varied than the last three in the two examples. Will these results be consistent over a large number of samples?

We will use a calculator program to compute, for each of 200 sets of 7 random integers, estimates using all 8 methods. The estimates are compared to 500, the true value of N, and the *difference* between the estimate and 500 is stored in a list. (See **Table 4.**)

(The program does not check to see that 7 unique integers were selected. The probability of repeated selections is small and should not alter the results significantly.)

Results of 200 Simulations

We are interested in the mean of the 200 errors for each of the 8 methods to measure the accuracy of the estimate. We will use the standard deviations of the estimates to measure the precision of the estimates. What we want is an estimator that has a mean of zero (is unbiased) and a small standard deviation (gives precise estimates). We will also look at the histograms of the 200 estimates to see how the estimates are distributed around zero. The results of one run of the program are given in **Table 5.**

A second run of the program gives the results in **Table 6.**

And a third run of the program illustrates the consistency of the mean and standard deviations for these estimates. (See **Table 7.**)

Notice that the Outlier Method overestimated N by a large amount each time, while the Percentile method underestimated each time. Even though there are problems with the mean and median methods, they are, on average, fairly accurate. The standard deviation of the median method appears to be about 1.5 times that of the mean method for these three runs of the program. On average, the median + IQR method gave excellent results, but with much more variability than in the Average Gap Size and $\frac{n+1}{n}$ methods. (It is intriguing that the standard deviation of the Average Gap Size and $\frac{n+1}{n}$ methods are exactly the same.)

```
ClrAllLists
SetUpEditor A,B,MEAN,MED,MIQR,          Set up lists to use
OUT,SYM,GAP,PER,N
For(X,1,200)                            Repeats 200 times
randInt(1,500,7)→A                      Select 7 random integers
SortA(LA)                               Sort the integers
1-Var Stats LA                          Find statistical values
2mean(LA) – 500→MEAN(X)                 Twice the Mean solution
(2median(LA) – 1) – 500→MED(X)          Twice the Median solution
Med+(Q1 – Q1) – 500→MIQR(X)             Median + IQR solution
(Q1+1.5(Q1 – Q1)) – 500→OUT(X)          Outlier solution
(max(LA)+min(LA) – 1) – 500→SYM(X)      Symmetric Range solution
ΔList(LA)→B                             Find gap sizes
augment(LB,{min(LA) – 1})→B             Add first gap (from 1)
(max(LA)+mean(LB)) – 500→GAP(X)         Gap Size solution
max(LA)/0.93 – 500→PER(X)               Percentile solution
(8/7)max(LA) – 500→N(X)                 (n+1)/n solution
End
```

TABLE 4.

Method	Twice Mean	Twice Median	Median + IQR	Outlier Method	Symmetric Range	Average Gap Size	Percentile Solution	$\frac{n+1}{n}$ Solution
Estimate	541	559	706	544	512	503	489	503

TABLE 1.

ESTIMATES FOR THE DIFFERENT METHODS

Method	Twice Mean	Twice Median	Median + IQR	Outlier Method	Symmetric Range	Average Gap Size	Percentile Solution	$\frac{n+1}{n}$ Solution
Estimate	588	621	538	645	510	503	473	503
Error	88	121	38	145	10	3	–27	3

TABLE 2.

ESTIMATES AND ERRORS FOR THE FIRST SAMPLE 210, 311, 71, 191, 440, 418, AND 417.

Method	Twice Mean	Twice Median	Median + IQR	Outlier Method	Symmetric Range	Average Gap Size	Percentile Solution	$\frac{n+1}{n}$ Solution
Estimate	373	247	539	873	466	528	497	528
Error	–127	–253	39	373	–34	28	–3	28

TABLE 3.

ESTIMATES AND ERRORS FOR THE SECOND SAMPLE 458, 5, 43, 124, 145, 69, AND 462.

Method	Twice Mean	Twice Median	Median + IQR	Outlier Method	Symmetric Range	Average Gap Size	Percentile Solution	$\frac{n+1}{n}$ Solution
Mean	1.78	–7.78	–1.98	255.0	3.06	2.04	–27.5	2.18
St Dev	117.3	176.4	127.8	191.6	83.4	61.6	57.9	61.6

TABLE 5.

Method	Twice Mean	Twice Median	Median + IQR	Outlier Method	Symmetric Range	Average Gap Size	Percentile Solution	$\frac{n+1}{n}$ Solution
Mean	–9.11	–9.61	–1.78	250.5	–6.76	–3.49	–32.7	–3.35
St Dev	100.9	155.4	116.0	180.4	76.1	60.9	57.3	60.9

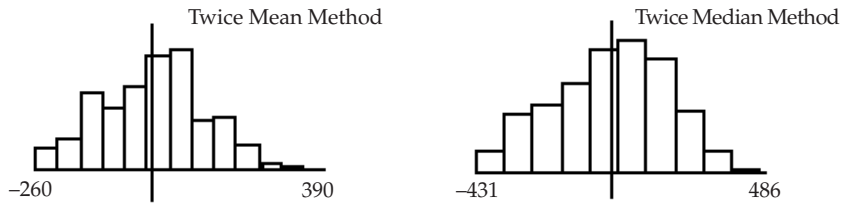
TABLE 6.

Method	Twice Mean	Twice Median	Median + IQR	Outlier Method	Symmetric Range	Average Gap Size	Percentile Solution	$\frac{n+1}{n}$ Solution
Mean	12.89	19.14	–0.615	233.3	11.11	3.77	–25.9	3.91
St Dev	108.8	164.0	118.7	180.5	75.3	54.8	51.5	54.8

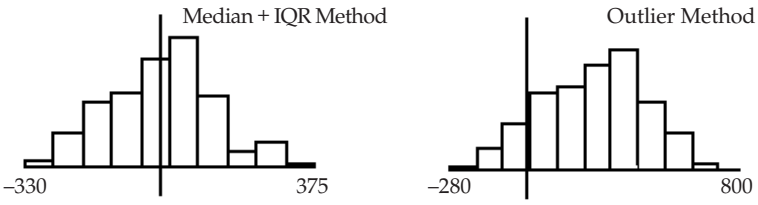
TABLE 7.

Graphical Analysis

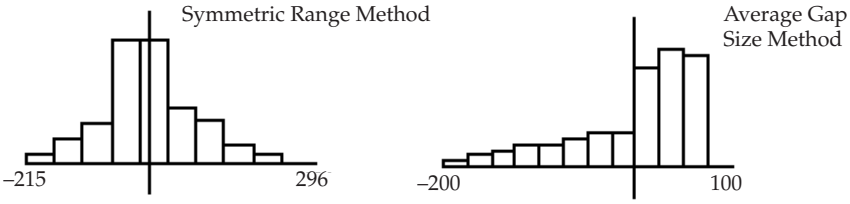
The histograms of the first 200 sample estimates are shown below:



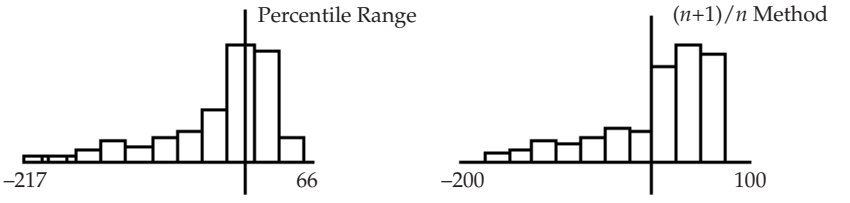
Both the mean and median methods have fairly symmetric distributions. They appear to have 0 near the center of their distribution, suggesting that they are unbiased. They overestimate as often as they underestimate the size of the population. Notice that the mean has a much smaller spread and so is a more consistent estimator than is the median.



The median + IQR method also appears to be unbiased, but the Outlier method is clearly biased, consistently overestimating the size of *N*.



The Symmetric Range also appears to be unbiased with a range that is smaller (more consistent estimates) than any of the previous measures. The Gap Procedure has an unusual distribution. Even though the distribution is asymmetric, it is unbiased because the average error is close to zero. Also, though the distribution is strongly skewed left, it has a much smaller standard deviation than the other measures. Of the 200 estimates, 133 were overestimates. The procedure is more likely to overestimate the value of *N*, but the overestimates are considerable smaller than the underestimates.



From the histograms, students can see that the Average Gap, Percentile, and $(n+1)/n$ methods are variations of the same procedure. In the Percentile Method, we multiply the maximum value (Max) by 1.08, while in the $(n+1)/n$ Method, we multiply the Max by 1.14. The average gap size is just (Max – 1)/*n*, so our estimate using the Gap Procedure is Max + (Max – 1)/*n*. This has the same distribution as the $(n+1)/n$ method with a slightly smaller mean.

Conclusion

Based on our simulations, we conclude that the Median + IQR solution has the smallest average error, but too large a variation. Better results come from either the Average Gap Size method and the $\frac{n+1}{n}$ method. They combine a small average error with the smallest standard deviation. In statistical terms, we say that by our simulation analysis, they are the minimal variance unbiased estimators of *N*.

An important point to make with this project is the need for simulations to assess the quality of the method. For any given sample, one method may do better than another, but it is the long-run capability of the method that must be assessed. On any given run, even our best methods may be off a lot; but used consistently, they will, on average, be accurate. This is the best we can hope for.

Finally, the problem has a more important role than just estimating the number of cabs that won't pick you up in the rain in a large city. According to Noether, the Allies were able to estimate the size of the German tank fleet by analyzing the serial numbers of the tanks observed in battle using a process little different from those described above. ❑

References

Flaspohler, David C. and Ann L. Dinkheller, "German Tanks: A Problem in Estimation," *The Mathematics Teacher*, Vol. 92, Num. 8, November, 1999.

Noether, Gottfried E., *Introduction to Statistics, The Nonparametric Way*, Springer-Verlag, New York, 1991.