

## How Many Taxis?

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 we 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. The following solutions are typical.

**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. 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:** The Interquartile Range (IQR) 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 you 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.

**Twice IQR Solution:** The interquartile range contains 50% of the data. Since the distribution is uniform, the length of the IQR should be the same as the length of the smallest half of the data and the largest half of the data. Consequently, twice the IQR should be an estimate of the range of the data. So  $2(226) = 452$  is our estimate for the number of taxis in the city.

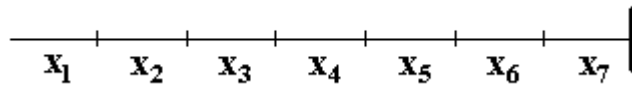
**1.5 IQR Solution:** A uniform distribution would not have any outliers. In our course, an outlier is defined as 1.5 times the IQR beyond the 3rd Quartile. The third quartile is 405 and the IQR is 226. In this sample, an outlier would be beyond  $405 + 1.5(226) = 744$ . We estimate the number of cabs at 744.

**Symmetric Range Solution:** In a uniform distribution, 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

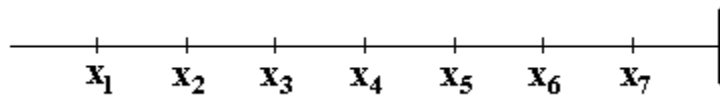
440 to the end of the list. The mean gap size is 62.7. If we add the mean gap size to the maximum known value, we have an estimate of the largest value in the list. Our prediction is  $440 + 63 = 503$  taxis.

**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 9 different methods for estimating the number of cabs, yielding values that vary from a low of 452 to a high of 706. Which method works the best? How would we know?

Method	Twice Mean	Twice Median	Median + IQR	Twice IQR	Outlier Method	Symmetric Range	Average Gap Size	Percentile Solution	$(n+1)/n$ Solution
Estimate	541	559	706	452	544	512	503	489	503

**Table 1:** Estimates for the different methods

### 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 over-estimate while others under-estimate?

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.

Method	Twice Mean	Twice Median	Median + IQR	Twice IQR	Outlier Method	Symmetric Range	Average Gap Size	Percentile Solution	$(n+1)/n$ Solution
Estimate	588	621	538	454	645	510	503	473	503
Error	88	121	38	-46	145	10	3	-27	3

**Table 2:** Estimates and Errors for the first sample 210, 311, 71, 191, 440, 418, and 417.

Notice that all but two of the methods over-estimated the true value of  $N$ . Is this generally going to be the case?

Method	Twice Mean	Twice Median	Median + IQR	Twice IQR	Outlier Method	Symmetric Range	Average Gap Size	Percentile Solution	$(n+1)/n$ Solution
Estimate	373	247	539	830	873	466	528	497	528
Error	-127	-253	39	330	373	-34	28	-3	28

**Table 3:** Estimates and Errors for the second sample 458, 5, 43, 124, 145, 69, and 462.

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?

We will use a calculator program to compute, for each of 200 sets of 7 random integers, estimates using all 9 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. (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.) The program for the TI-83 is given below, with comments on the side in *italics*:

```

ClrAllLists
SetUpEditor A,B,MEAN,MED,MIQR,           Set up lists to use
TIQR,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-Q3)-500→MIQR(X)               Median + IQR solution
  2*(Q1-Q3)→TIQR(X)                    Twice IQR solution
  (Q1+1.5(Q3-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)/.93-500→PER(X)                  Percentile solution
  (8/7)max(LA)-500→N(X)                   (n+1)/n solution
End

```

### Results of 200 Simulations

We are interested in the mean of the 200 errors for each of the 9 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 below:

Method	Twice Mean	Twice Median	Median + IQR	Twice IQR	Outlier Method	Symmetric Range	Average Gap Size	Percentile Solution	$(n+1)/n$ Solution
Mean	1.78	-7.78	-1.98	6.26	255.0	3.06	2.04	-27.5	2.18
St Dev	117.3	176.4	127.8	166.9	191.6	83.4	61.6	57.9	61.6

A second run of the program gives the following results:

Method	Twice Mean	Twice Median	Median + IQR	Twice IQR	Outlier Method	Symmetric Range	Average Gap Size	Percentile Solution	$(n+1)/n$ Solution
Mean	-9.11	-9.61	-1.78	15.11	250.5	-6.76	-3.49	-32.7	-3.35
St Dev	100.9	155.4	116.0	164.5	180.4	76.1	60.9	57.3	60.9

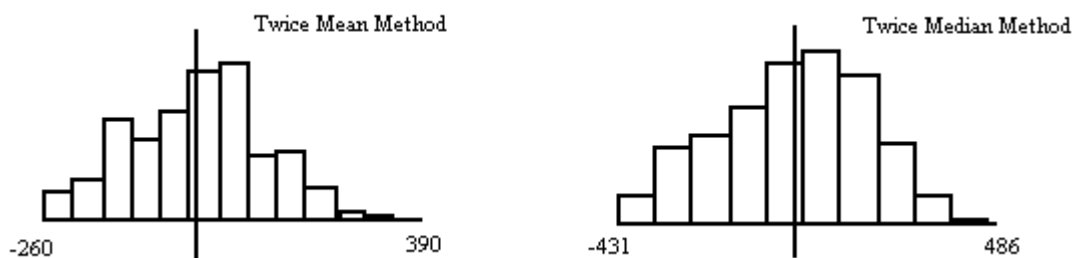
And a third run of the program illustrates the consistency of the mean and standard deviations for these estimates.

Method	Twice Mean	Twice Median	Median + IQR	Twice IQR	Outlier Method	Symmetric Range	Average Gap Size	Percentile Solution	$(n+1)/n$ Solution
Mean	12.89	19.14	-0.615	-7.67	233.3	11.11	3.77	-25.9	3.91
St Dev	108.8	164.0	118.7	172.0	180.5	75.3	54.8	51.5	54.8

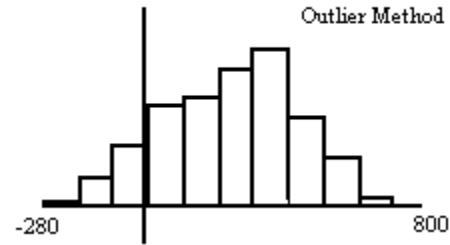
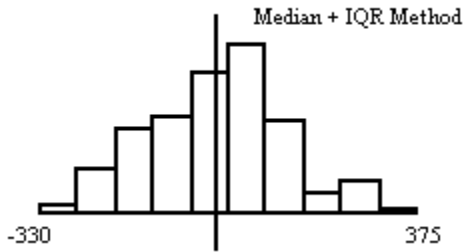
Notice that the Outlier Method over-estimated  $N$  by a large amount each time, while the Percentile method under-estimated each time. Even though there are problems with the mean and median methods, they are, on average, fairly accurate. The standard deviation of the median method appears to be about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

### Graphical Analysis

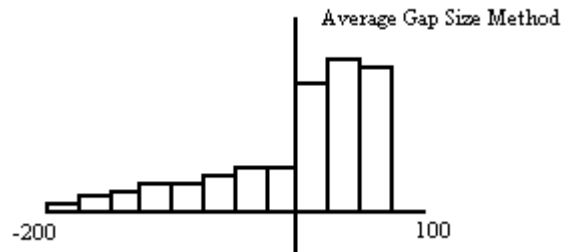
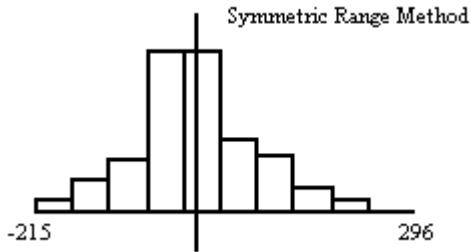
The histograms of the first 200 sample estimates are shown below:



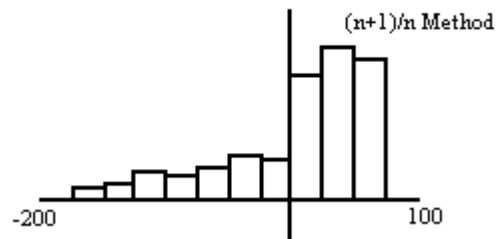
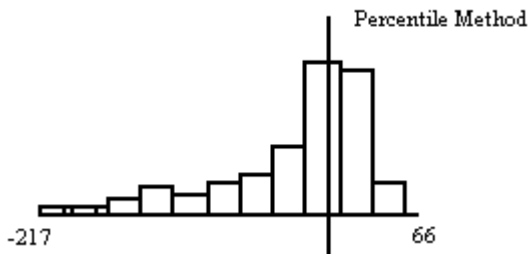
Both the mean and median methods have fairly symmetric distributions. They appear to have 0 near the center of their distribution, suggesting that they are unbiased. They overestimate as often as they underestimate the size of the population. Notice that the mean has a much smaller spread and so is a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



The median + IQR method also appears to be unbiased, but the Outlier method is clearly biased, consistently over-estimating 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 over-estimates. The procedure is more likely to over-estimate the value of  $N$ , but the over-estimates are considerably smaller than the underestimates.



Both the Percentile and  $\frac{n+1}{n}$  Procedures have distributions similar to that of the Gap Procedure. It should be easy for students to understand why the distributions of the final two procedures are similar. In one case, we multiply the maximum value by 1.08 and the other by 1.14, so they should have similar distributions. It is less clear why the distributions of the Average Gap Size and the  $\frac{n+1}{n}$  Procedures have exactly the same distribution.

### Conclusion

After all this discussion, which method should students choose and why? What are the important features of the estimator? Do we want approximate normality? mean of zero? small variance? This activity opens the door to our convention of preferring unbiased minimum variance estimators.

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