

Simulation of Estimators

Suppose you want to estimate the center of the normal distribution. Should you use the sample mean, or the sample median? Let's try it and see. Let's simulate draws from a normal distribution with mean 10 and standard deviation 2.

```
MTB > random 500 c1-c7;  
SUBC> normal mu=10, sigma=2.  
MTB > print c1-c7
```

Row	C1	C2	C3	C4	C5	C6	C7
1	13.3323	12.0040	10.4347	8.7582	9.5735	12.3209	11.7769
2	10.2563	10.3434	10.7180	12.9429	10.8445	12.9942	9.2298
3	9.7769	7.5453	8.8069	11.2322	12.3766	8.9044	9.9838

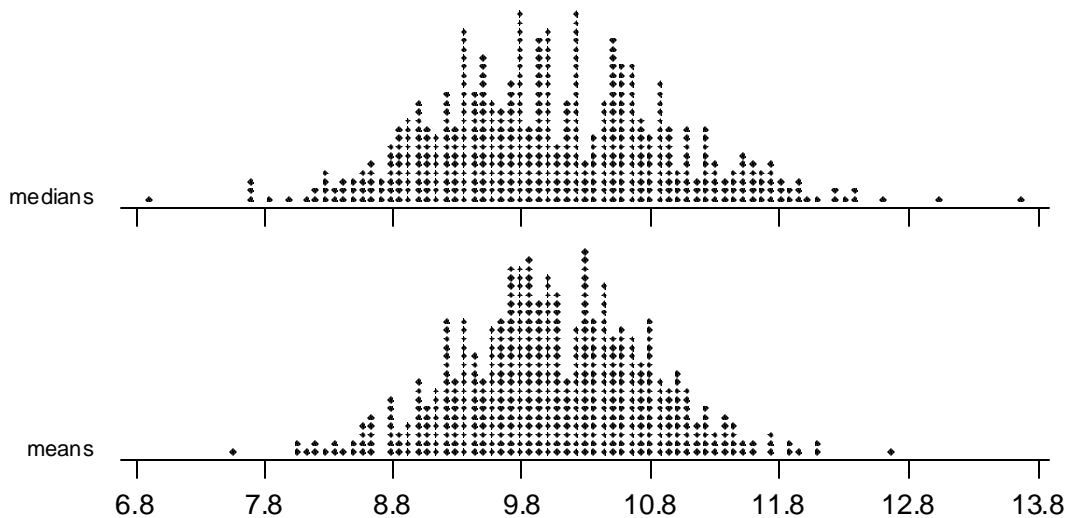
Here are the commands along with some of the data. Then we try to estimate the population mean both ways. Remember, the right answer is 10.

```
MTB > rmean c1-c7 c8  
MTB > rmedian c1-c7 c9
```

Here is the result:

By eyeballing it you can see that both measures seem to give “average guesses” of about 10,

Dotplot for means-medians



which is what we would expect from unbiased estimators. Also, the guesses for the median are more spread out . . . more likely to be far from the truth. We can see this in the descriptive statistics.

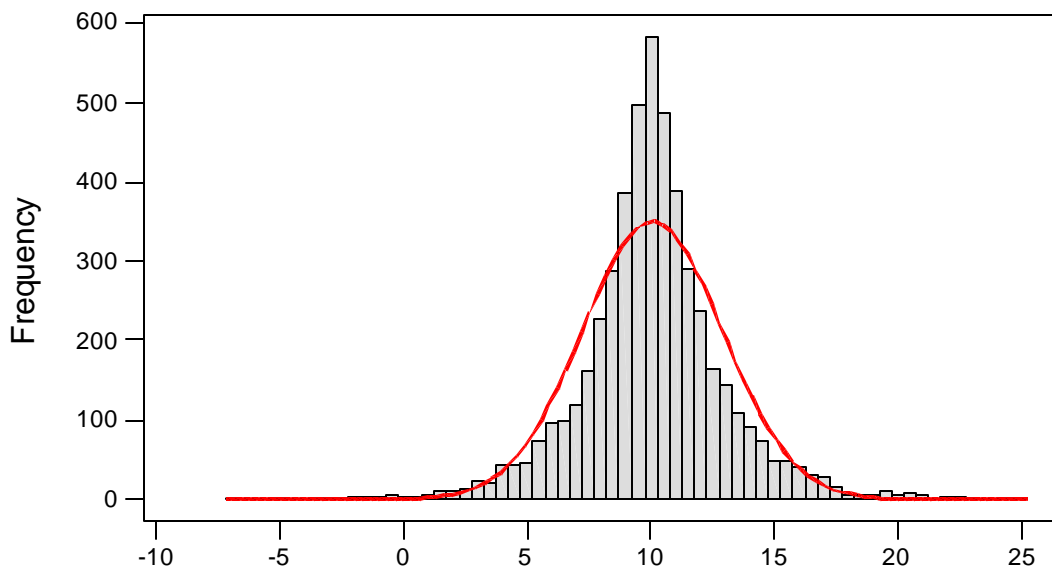
Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
means	500	10.035	10.021	10.036	0.758	0.034
medians	500	10.034	9.987	10.022	0.944	0.042

Variable	Minimum	Maximum	Q1	Q3
means	7.516	12.641	9.551	10.542
medians	6.926	13.699	9.370	10.644

This result is not general. It holds for the normal because there is a relatively small chance of extreme outliers in the normal distribution. By contrast, consider the Laplace distribution. Here is what the Laplace looks like compared to the normal.

Comparison of the Laplace Distribution and the Normal.

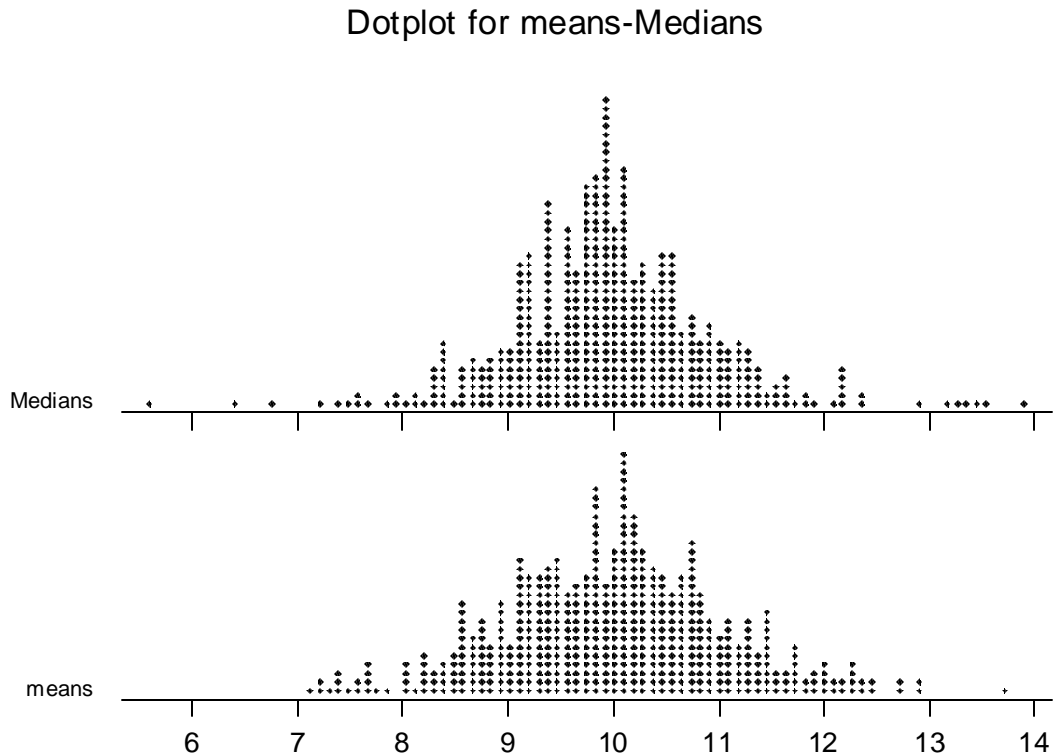


Fatter tails; more probability about the mean. Here is what happens if we simulate draws from the Laplace. (The true mean of this particular Laplace distribution is 10.)

```
MTB > random 500 c1-c7;
SUBC> laplace a=10 b=2.
MTB > rmean c1-c7 c8
```

```
MTB > rmedian c1-c7 c9
```

Both still seem unbiased, but the clear superiority of the mean is no longer evident. As a matter



of fact, we can look at the descriptive statistics to get a more precise view.

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
means	500	9.9913	10.0191	9.9888	1.0722	0.0479
Medians	500	9.9517	9.9124	9.9400	0.9887	0.0442

Variable	Minimum	Maximum	Q1	Q3
means	7.0938	13.6904	9.2926	10.6702
Medians	5.5691	13.8597	9.3841	10.4830

The medians actually do better, because they have a smaller standard deviation!

Now lets compare two ways of estimating the population variance: one dividing the sum of squared deviations by n, and one dividing the sum of squared deviations by n-1.

```
MTB > random 300 c1-c3;  
SUBC> normal 10 2.  
MTB > rmean c1-c3 c4
```

```

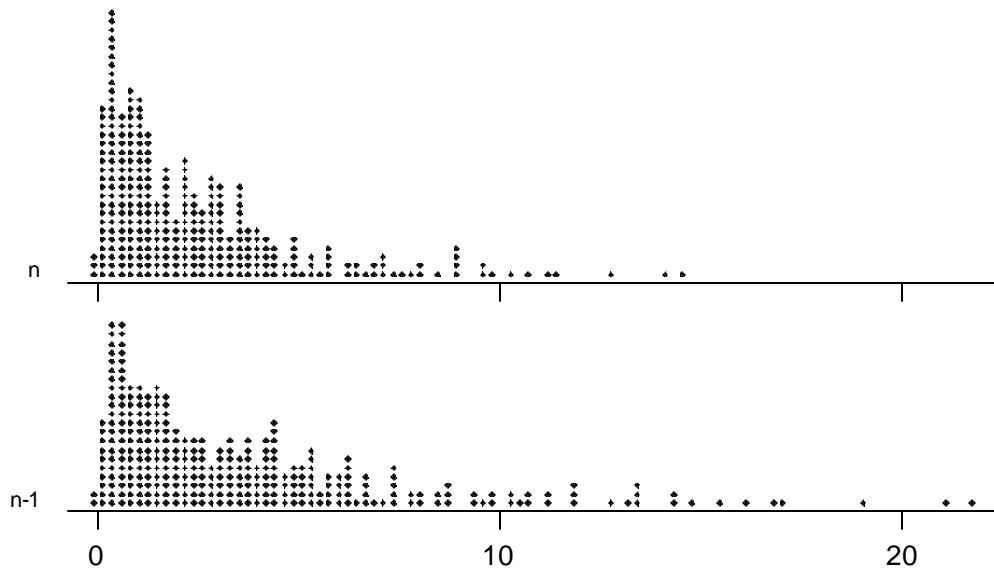
MTB > let c5=(c1-c4)**2
MTB > let c5=(c1-c4)**2 + (c2-c4)**2 + (c3-c4)**2
MTB > let c6=c5/2
MTB > let c7=c5/3

```

Note that the observations we simulated came from a normal with a true population standard deviation of 2, so that 4 is the true population variance. Here is how our estimates came out.

These are harder to judge by eye because the distribution is not symmetric, but you can see that

Estimating the Population variance



dividing by n tends to give an answer that is too small. This is clearer in the descriptive statistics.

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
n-1	300	3.829	2.650	3.368	3.904	0.225
n	300	2.553	1.767	2.246	2.603	0.150

Variable	Minimum	Maximum	Q1	Q3
n-1	0.015	21.666	1.057	5.194
n	0.010	14.444	0.705	3.462

Dividing by n-1 results in an average guess of 3.829, which is pretty close to 4. In fact, it can be shown that using the n-1 divisor makes the estimate unbiased. Using the divisor of n, however, results in an average guess of 2.553, well below the true value of 4.

However, the standard deviation is smaller using a divisor of n. Which is the more efficient estimator measured in MSE?

$$MSE \equiv E(\hat{\theta} - \theta)^2 = \text{var}\hat{\theta} + (\text{bias}\hat{\theta})^2$$

Applying this to the estimator with the n-1 divisor, we can approximate the MSE by using the results of our simulation.

$$MSE = (3.904)^2 + (3.829 - 4)^2 = 15.27$$

Applying this to the estimator with the n divisor, we can approximate *its* MSE.

$$MSE = (2.603)^2 + (2.553 - 4)^2 = 8.87$$

For efficiency, measured by MSE, we actually did better dividing by n than dividing by n-1. As a matter of fact, there is a theorem that says using a divisor of n+1 actually gives the best MSE when the draws come from a normal distribution. This is a good example of a case where our criteria diverge: using a divisor of n-1 gives an unbiased estimate, but using a divisor of n (or n+1) gives a more efficient estimate.