

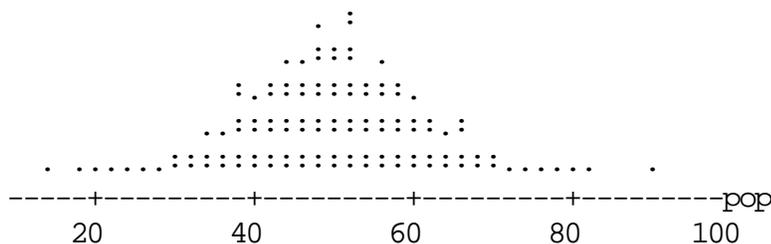
## SAMPLING DISTRIBUTIONS AND POPULATIONS

Dennis Roberts - Penn State

One of (many) the confusing points in statistics is the separation of the concepts of the target parent population and sampling distributions that result by sampling from those parent populations. This short handout is designed to show both the parent population and several sampling distributions (using varying n) that result.

### Population Distribution

Assume that we define a population that is normal with a mean = 50 and a standard deviation = 10. In normal distributions, you usually can go about 3 standard deviation units on both sides of the mean. So, with a mean of 50 and a standard deviation of 10, scores IN THE POPULATION would range from about 20 to 80. I have simulated what such a distribution would look like and have "dotplotted" it below.



```
MTB > desc c1
```

	N	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
pop	1000	49.798	49.908	49.853	10.245	0.324
	MIN	MAX	Q1	Q3		
pop	13.248	90.664	42.928	56.755		

Note that the simulated distribution (based on 1000 cases) does not have a mean exactly equal to 50 nor a standard deviation exactly equal to 10; but, they are fairly close. **PLEASE NOTE: THIS DOTPLOT REPRESENTS HOW THE ORIGINAL DATA POINTS, IN THE PARENT POPULATION, WOULD VARY AROUND THE TRUE POPULATION MEAN. THIS IS THE DISTRIBUTION OF X VALUES ... ORIGINAL DATA POINTS.**

### Sampling Distributions

Now, what if you wanted to take samples of varying sizes from this population, and then examine the sampling distribution of means in each case? Let's say that we take 200 samples, at random, from this population where ...

we first start with sample sizes of  $n = 64$  and then do the same thing using (different) sample sizes of 49, 36, 25, 16, 9, 4 and ... yes ... even  $n = 1$ . In case, we would be producing a sampling distribution of means based on varying  $n$ 's.

In the first case, we take 200 samples each time of  $n = 64$  and then examine the distribution of the 200 sample means. Theoretically, using the standard error of the mean formula, the standard deviation of this sampling distribution of means should be approximately equal to:

$$S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

For the case where  $n = 64$ , and the population standard deviation is about 10, this standard error should be about  $10 / 8 = 1.25$ . Sample means should average to about 50 in the long run with a standard deviation of about 1.25. The value of 1.25 is the standard error.

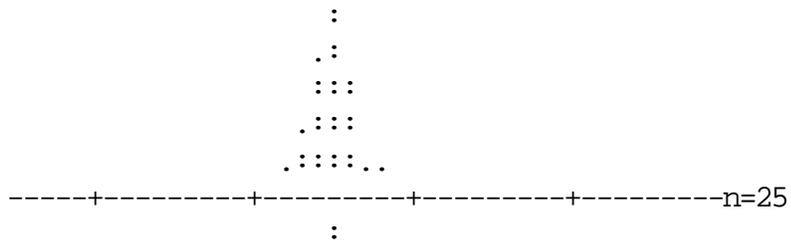
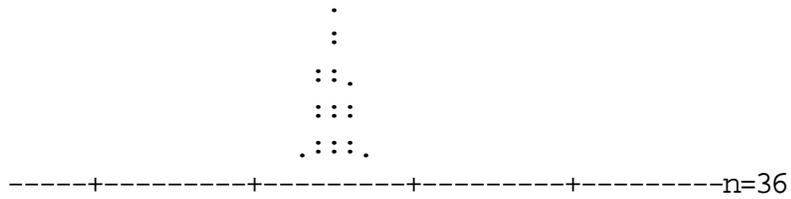
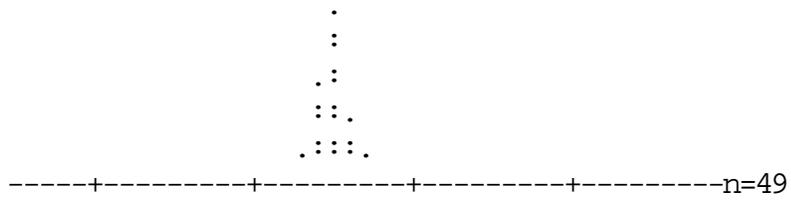
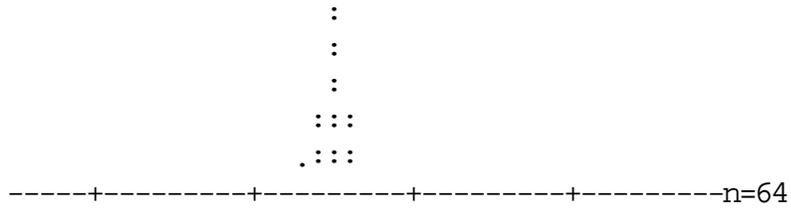
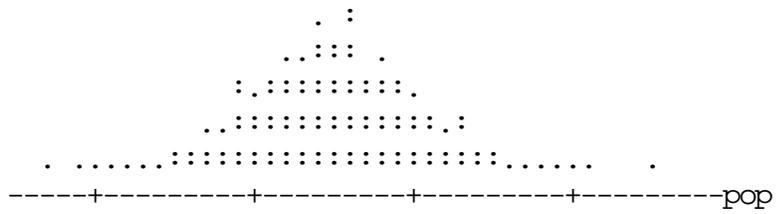
If we takes samples of  $n = 49$ , then the standard error of the mean should be about  $10 / 7 = 1.43$ .

The table below shows what the approximate standard errors of the mean should be given a  $\sigma = 10$  and varying  $n$  cases.

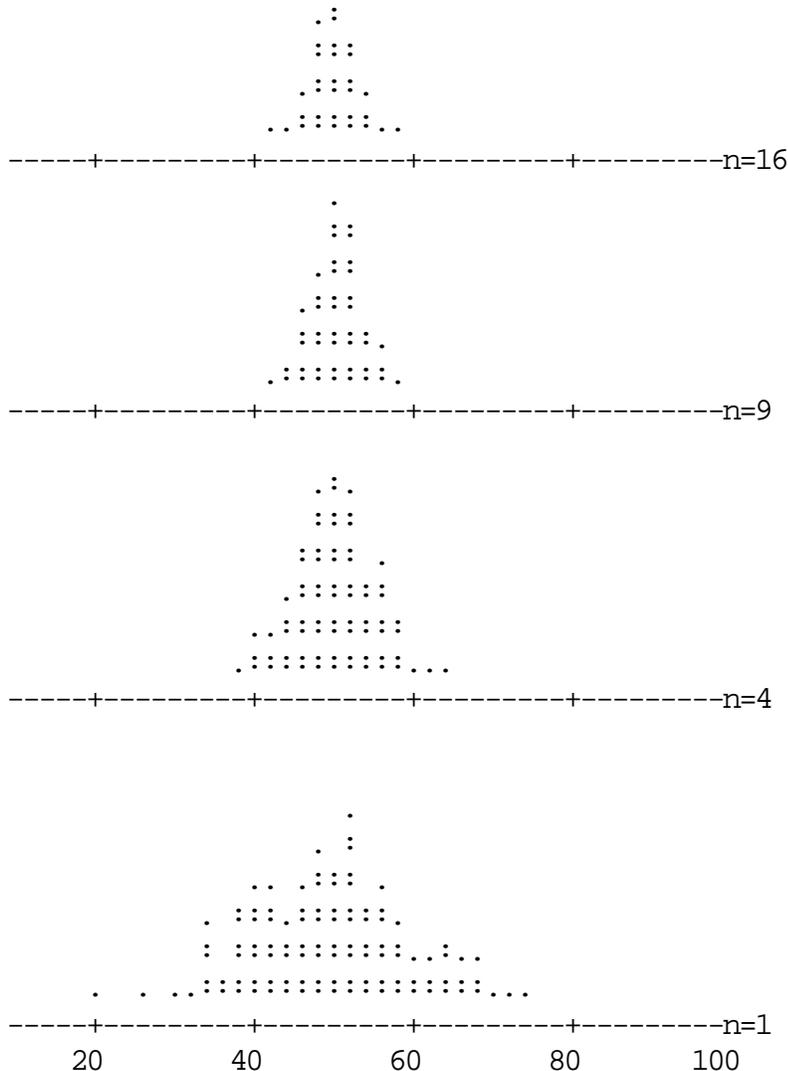
Sample Size	Standard Error
64	1.25
49	1.43
36	1.67
25	2.00
16	2.50
9	3.33
4	5.00
1	10.00

To show how the sampling distributions of MEANS compare to the original parent population of SCORES, I have made dotplots of, first, the simulated parent population, and then each sampling distribution of means starting with  $n = 64$  and moving down to  $n = 1$ . NOTE: For comparison purposes, you will have to flip down to the last dotplot to see the actual baseline scale.

```
MTB > dotp c1 c65-c72;
SUBC> same.
```



:



When  $n = 64$ , the means from random samples fit compactly around the population mean value of 50, and the low and high means in this sampling distribution are only about 3 points below and above 50. That is, when  $n = 64$ , it is highly unlikely that any sample mean will be more than about 3 score points away from the population mean. Please note that the  $n = 64$  distribution is a distribution of means, not original score values. Think about how the sampling distribution of means fits into the parent population distribution of scores. The parent population has 1000 cases, but each sampling distribution has only 200 cases. So, the sampling distributions in these simulations are smaller "volume" distributions.

If you work your way down through the dotplots for sample sizes from  $n = 64$  to  $n = 4$ , you notice clearly that as the sample size gets smaller, the width of the sampling distribution gets larger and larger. That is, the range of the possible sample mean values gets larger and larger as the  $n$  gets smaller and smaller. All this indicates is that there is more sampling error when

random samples are smaller; i.e., there can be bigger gaps between the true population mean and any given sample mean, as n gets smaller.

### Sampling Distributions for n = 1

Finally, you should see that the last sampling distribution is the case where n = 1. What about this situation? Well, if you take samples of n = 1, you will eventually get in the samples you select each and every one of the original numbers or scores that existed in the target parent population. So, if that is the case, the mean of an n = 1 sample IS that value, and therefore, the means from n = 1 samples will eventually range across the entire spectrum of the population score values. In effect, the sampling distribution of means when n = 1 is exactly like the parent population. In this case, the standard error of the mean is identical to the actual standard deviation in the population. Whenever sample sizes are larger than 1, the standard error of the mean will be less than the population standard deviation.

### Statistics on Sampling Distributions

Below, I have listed the descriptive statistics on the dotplots above. Note that as sample sizes decrease, the standard deviations (**which are the standard errors of the means for these simulations**) get larger and larger and get closer to the value of 10, which is the population standard deviation. Note that these values are fairly close to the theoretical ones shown in the previous table.

```
MTB > desc c65-c72
```

	N	MEAN	MEDIAN	TRMEAN	STDEV	= Stan err
n=64	200	49.983	49.912	49.979	1.221	
n=49	200	49.886	49.940	49.881	1.341	
n=36	200	49.938	50.007	49.933	1.622	
n=25	200	49.750	49.855	49.756	1.950	
n=16	200	49.692	49.684	49.672	2.422	
n=9	200	50.163	50.280	50.165	3.277	
n=4	200	50.077	50.159	50.127	4.887	
n=1	200	49.320	49.443	49.315	9.750	
	MIN	MAX	Q1	Q3		
n=64	46.521	52.947	49.100	50.816		
n=49	46.118	54.646	48.941	50.756		
n=36	46.296	54.443	48.615	50.938		
n=25	44.219	55.658	48.518	51.083		
n=16	42.466	57.121	47.872	51.100		

n=9	42.075	58.450	47.738	52.482
n=4	37.540	63.165	46.898	53.376
n=1	20.411	74.146	42.473	55.583

So, what do we learn from this? First of all, it is important to realize that we are examining TWO distributions; the population distribution of SCORES and the sampling distribution of MEANS. As can be seen with the dotplots, the sampling distribution of means sits inside of the population distribution. While we can use the same baseline to place both distributions on, we need to keep in mind that the numbers along the baseline refer to mean values when looking at the sampling distribution, and to score values when looking at the population distribution.

Also, it is helpful to think about the widths of the sampling distributions of means as being some function of the population standard deviation. The worst case scenario would be if sample means vary as much as the actual scores in the population. Therefore, we could look at sampling error in light of how big the standard error of the mean compares to the actual population standard deviation. For example, when  $n = 100$ , then the sampling distribution has a width approximately equal to  $S$  over 10, or ... a standard error of the mean that is about  $1/10$  the size of the population standard deviation. For  $n = 25$ , the standard error of the mean is now about  $S$  over 5 or about  $1/5$  the size of the population standard deviation. Obviously, as  $n$  gets smaller, the standard deviation of the sampling distribution gets proportionately closer to the size of the actual standard deviation of the population. Or, in reverse, the larger the random sample, the smaller is the proportion of the standard error compared to the overall standard deviation of the population of scores. It would tend to make sense to assume that our goal would be to select a sample size that produces a sampling distribution of means that is substantially narrower than the full population.



DENNIS M. ROBERTS

Penn State University  
208 Cedar Building  
University Park, PA 16802 USA

**EDUCATIONAL PSYCHOLOGY**

AC: 814-863-2401  
Bitnet: DMR @ PSUVM  
Internet: DMR @ PSUVM.PSU.EDU

---

—

