

CHAPTER 16

INTRODUCTION TO SAMPLING ERROR OF MEANS

The message of Chapter 14 seemed to be that unsatisfactory sampling plans can result in samples that are unrepresentative of the larger population. Recall that it was stated that the major purpose of using a sample was to provide a practical means of estimating one or more parameters of the population to which you want to generalize your results. For example, perhaps it is the mean (\bar{F}) height of 4th grade youngsters that you want to estimate and to do that, you measure the heights of a sample of 4th graders in your state. The statistic \bar{X} (sample mean) will be used as your best estimate of the corresponding population parameter. But, it is highly unlikely that the mean in any one particular sample will be identical to the true population mean. Thus, regardless of the sampling method used, good or bad, there is likely to be some error in the sample statistic in representing the parameter. Error due to sampling is a fact of life and one can never eliminate that problem unless you have access to the total population on which to do your analysis. But, that is an improbable situation. For example, the United States census attempts to contact everybody (emphasis on body) to make counts of various things such as the amount of homelessness, number of senior citizens, and the like. However, even when the government tries to obtain parameter information, they are unable to do it. So, even they are faced with error due to imperfect data collection. The concept of sampling error will be explored in more detail in this section using the sample mean as the statistic. The sample mean is a relatively easy way to introduce the concept of sampling error since the error of sample means follows rather straightforward rules.

Sampling from Populations

The first thing you have to understand is that one could draw or take many different samples from a given population. Assuming that you are using a good sampling plan such as random sampling, different samples will include different subsets of elements from the population. Therefore, some samples may have more taller persons in them than others and others may have a greater number of shorter persons. In any case, if you were looking for average heights as depicted by the different samples, the means in the various samples are likely to be different. If the differences are very small from sample to sample, then sampling error is small and this issue is really not very important. However, if different samples tell you radically different things about "average heights", then sampling error is large and must be factored into any inference you make from the sample to the larger population.

As a simple illustration of how sample means can vary, consider the population of intelligence test scores (infamous IQ's) where the mean is supposed to be about 100 and the standard deviation is about 16. Keep in mind that there is nothing magical about 100 and 16; these figures are based on what the test publishers set as the mean and standard deviation. Think about having a batch of sampling fanatics who each go out and take their

own "random sample" and calculate their own "sample mean". Consider the following case where I have 100 of these sampling fanatics and each took their random sample from a population where the parameter (mean) is 100 and the parameter (standard deviation) is 16. Thus, all 100 are in fact sampling from a population where, by convenience for this illustration, we know the values of the parameters. Using the random command in Minitab, I can do this easily generating 100 samples where $n = 16$ in each case, and each sample has been drawn from the population where mean = 100 and standard deviation = 16.

Since the normal distribution is a continuous distribution, decimal values for sample observations are possible. Some of the data from some of the columns that I generated at random are printed below.

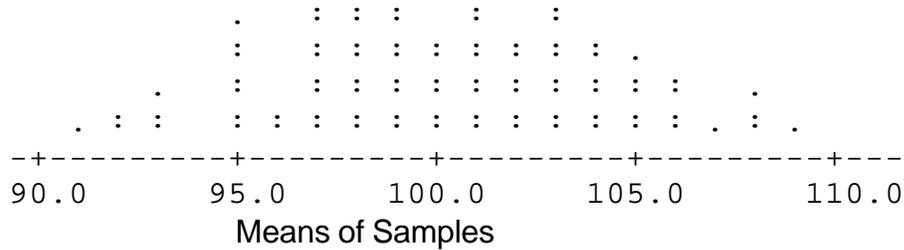
SAMPLE DATA GENERATED FROM NORMAL DISTRIBUTION

76.096	91.577	97.295	88.321	104.996	106.773
107.946	98.359	81.363	92.778	104.944	121.652
84.814	94.141	108.815	89.404	73.668	103.131
102.499	116.391	88.792	107.704	123.405	85.467
101.622	91.068	76.549	108.326	113.167	91.734
87.056	113.038	95.306	124.745	110.878	111.636
106.309	114.186	89.441	85.068	98.791	72.707
99.570	102.709	56.432	106.734	90.769	98.096
98.109	118.165	109.745	85.402	105.000	89.130
109.952	122.646	93.605	114.807	123.699	125.285

But, even though I am assuming that IQ scores are normally distributed in the population, and normal distributions are really continuous distributions (thus allowing for decimal values), I have taken each of the 16 values collected at random from each of the 100 data gathering fanatics, and rounded it off to the nearest whole number. This is more realistic when we are talking about IQ scores; ie, scores of 98 or 84 or 121.

Now, for each of the 100 samples, I calculated the sample mean. Thus, over this data gathering experiment, we have 100 sample means where each mean is based on a sample size of $n = 16$. To see what these means look like, I have simply made a graph or dotplot of the 100 values. See the top of the next page for this graph. I have also calculated some descriptive statistics on that set of 100 sample means and that information is also included below the graph.

	:		
.	:		
		:	:
		:	:



DESCRIPTIVE STATISTICS ON 100 SAMPLE MEANS

	# Means	MEAN	MEDIAN	STDEV	MIN	MAX
SampMeans	100	100.13	100.50	3.91	91	109

Thus, if you had been the one who obtained the sample where the mean was 91 and you used that as your best estimate of the population mean, you would have made an error of about $91 - 100$ or -9 ; ie, 9 IQ points too low compared to the truth. On the other hand, a person who obtained the "high" sample mean of about 109 would have made the same sized error but would have overestimated the truth about 9 points. Conceptually, any difference between the parameter and the statistic is an error and since different samples have larger or smaller errors, these errors are called **SAMPLING ERRORS** (makes sense!). By definition, sampling error in this case is as follows.

$$\text{SampErr} = (\text{Statistic} - \text{Parameter})$$

$$\text{SampErr}' (\bar{X} - \mu)$$

Anything we can do, by implementing a better sampling plan, to make the gap smaller (on the average) between the statistic and the parameter, the better off we will be.

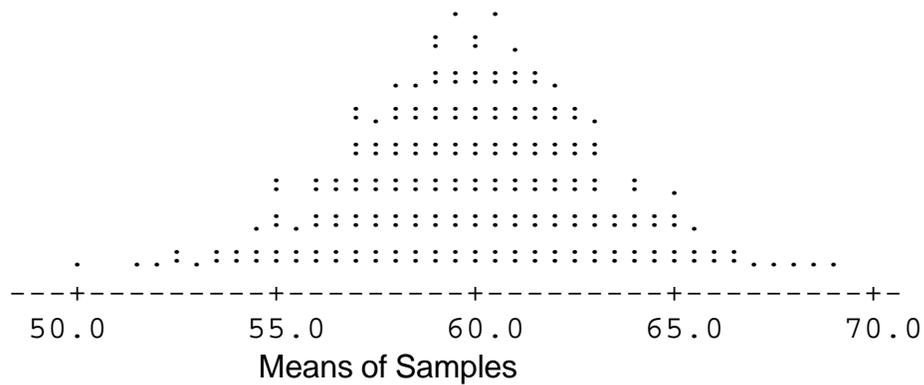
Sampling Distribution of Means

To explore the concept of sampling error further, consider that the population of typists tends to type at a rate somewhere between about 45 wpm to 75 wpm, which means that the overall mean would be about (halfway between) 60 words per minute. Note: I am not saying this is the true picture of typists! I am just using it as an illustration. What would happen if we took many samples (our sampling fanatics are at it again!) from this population of thousands and thousands of typists and examined the means of these samples?

The data below are based on randomly generating 1000 samples of $n = 9$, using Minitab, where we are sampling from a population where wpm values range from 45 to 75. For reference purposes, 10 of the samples, and their values and sample means, are shown below.

SAMPLE										MEAN
1	54	57	60	68	66	54	70	64	60	61.44
2	46	68	66	66	63	45	67	70	47	59.78
3	53	54	64	52	67	45	68	50	70	58.11
4	47	46	72	45	57	60	60	52	57	55.11
5	50	59	74	49	67	56	73	48	69	60.56
6	46	62	72	64	65	45	45	53	54	56.22
7	58	71	52	45	58	69	49	46	73	57.89
8	66	45	59	51	50	53	57	60	61	55.78
9	49	68	51	54	46	73	54	74	72	60.11
10	53	61	51	59	47	58	70	52	54	56.11

The dotplot and descriptive statistics of the 1000 means are as follows.



DESCRIPTIVE STATISTICS ON 1000 SAMPLES

# Samps	MEAN	STDEV	MIN	MAX
Samp Means	1000	59.963	3.101	50.222 68.778

Notice, in this case, that the sample means vary from about 50 to about 69. Thus, even though the population mean or parameter is about 60, sample means will vary around the parameter. About half of the 1000 means seem to be lower than 60 and the other half seem to be higher than 60. Also note that the shape of the distribution of sample means appears to resemble the normal distribution. That is, most of the sample means are close to the middle or the parameter of 60 and fewer and fewer of the sample means deviate larger distances from the central or middle point.

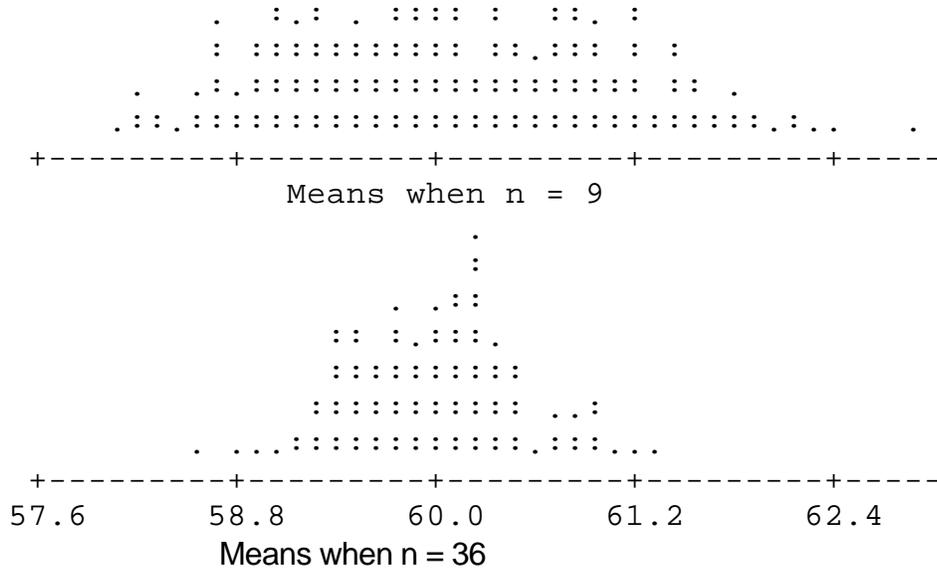
A distribution of sample means (assuming there are hundreds and hundreds of samples included) is called a **SAMPLING DISTRIBUTION** since the data in it came about due to taking many, many random samples from the population and making a distribution of the statistics (sample means) from those samples. In this case, since the distribution is

based on sample means, it is more technically called a **SAMPLING DISTRIBUTION OF MEANS**. Please keep in mind that this sampling distribution is based on chance which, stated differently, is simply the luck of the draw of random sampling in action. Sampling distributions of means tend to be normal shaped. Also notice that the standard deviation of this sampling distribution, 3.101 in this case, is an index number of how much the distribution spreads out, or is a measure of sampling error. Remember, in a normal distribution, one can mark off approximately 3 standard deviation units on either side of the mean. In this case, 3 units of about 3 would mean that sample means (when $n = 9$ from this population) would not be expected to vary by more than approximately 9 or 10 points on either side of the population parameter mean, or 60 in this situation. How much spread would you see in this sampling distribution if we had used either larger or smaller sized samples? That is, what would the standard deviation of the sampling distribution (3.101 in this case) be if samples had been a different size? We now turn to that question.

Sampling Error and Sample Size

To expand our discussion of sampling error of means, it would be helpful to explore the factors that impact on whether sampling error tends to be relatively large or relatively small. Consider the following. Continuing with our "typing speed" example, assume that out there in the larger population of typists, that the average wpm is about 60 ($\mu = 60$) and the standard deviation of the wpm values is about 3 ($\sigma = 3$). Assume that we randomly selected 200 samples of $n=9$ each and examined what the means looked like across all 200 samples. I have used Minitab to do these simulations. This would be similar to what was just presented in the section above; ie, samples have $n = 9$. However, after looking at those results, you could do the simulations but, in the process, change the sample size from $n = 9$ to $n = 36$; ie, make the size of the sample 4 times greater. In fact, in this example, I have done just that. I have generated a second set of 200 samples where the parameters have been kept the same (mean = 60, standard deviation = 3) but the size of the sample has been increased to $n = 36$. Thus, we have a real experiment here in that only one condition has been manipulated and that factor is the size of the sample. All other things have been held constant. The dotplots at the top of the next page show a graph of the 200 means that resulted in each of the 2 different simulation cases: top dotplot is when there are 200 samples where $n = 9$ in each case, and the second dotplot represents where there are 200 samples where $n = 36$ in each case. Keep in mind, size of samples is either 9 or 36, not 200. The value of 200 is how many samples have been generated.

: : :
 : . : . : . : :



DESCRIPTIVE STATISTICS ON n = 9 AND n = 36 SAMPLES

	# Samps	MEAN	STDEV
n = 9 Samples	200	60.095	0.983
n = 36 Samples	200	59.990	0.508

Notice, of course, that the two sampling distributions tend to fit the normal pattern. In addition, when examining the descriptive statistics, note that the means of all 200 samples when n = 9 and also when n = 36 tend to be close to 60, which is the parameter or population mean in this case. But, just as obvious as that is the fact that when n = 9, the means disperse more and are more variable around the center value of 60. Thus, when random samples are smaller, the sample means vary more and that means there is more sampling error. This is evident not only by the dotplots but also by the standard deviation of the sampling distribution of means. When n = 9, the sampling distribution had a standard deviation of about .98 whereas the standard deviation of the sampling distribution when n = 36 was only about a value of .51 (about half the size). Note: in this case, even though the sample size was 4 times larger, the sampling error was only about 1/2 the amount. The rule here is simple: **AS THE SIZE OF THE RANDOM SAMPLE INCREASES, THE AMOUNT OF SAMPLING ERROR** (as measured by the standard deviation of the sampling distribution) **DECREASES**. Therefore, there is an inverse or negative relationship between sampling error and the size of the random sample: as one goes up, the other goes down. Look at it this way.

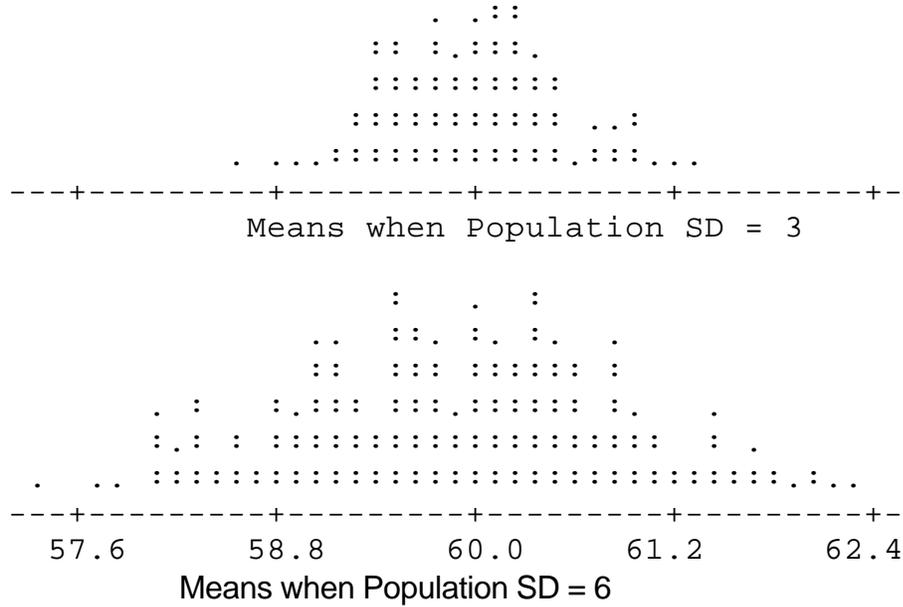
$$\frac{\text{Sampling error of means}}{1} = \frac{1}{\text{Sample size}}$$

Sampling Error and Population Variability

The comparison between sampling distributions when sample size is smaller ($n=9$) and when sample size is larger ($n=36$), only examined sample size and sampling error. I would like to be able to report to you that sample size is the only factor that impacts on sampling error of means but, alas, it is not. Another factor that has an impact on sampling error of means is the variability that exists in the population. As a very simple example, what if we had a population of students from a statistics course who, because they were very smart, scored on exams scores of 90% or higher. If you were to take samples from that population, with your different samples, you might get sample means around 93 or 97; such values would be possible. However, what if the class actually was more variable in its ability and, scores on exams ranged from about 65 to 100. If you were to take random samples from that population, you might still find sample means around 97 but, on the low side, you could possibly see sample means down in the low 70 area. Chance would allow both of these sample means (97 or say 73) to occur because the scores or elements in the population actually occur. The rule is simple: you can't get a sample mean unless there are elements in the population that have that value. Thus, in the first case where the elements were all high in the population, sample means could not vary very much, but in the second case, the sample means could vary much more because there is a much larger range of scores in the population. The variability in the population issue is what we now explore in a little more detail.

Recall in the previous simulations, I defined the parameters as being a mean of 60 and a standard deviation of 3. With a population standard deviation of 3, that means that the actual score numbers or typing wpm values in the population (since I assumed that the population was normally distributed) could have ranged from about 60 down 3 units of 3, up to 60 plus 3 units of 3 or from about 57 to 63. What might happen to sampling error of means if the original population had been wider for example? To test this out, I will need to keep my sample size or n constant, so as not to confound our new experiment, while allowing the population variability to vary. So, here is what I did.

The last simulation from the previous section was the case of taking 200 samples where each sample was $n = 36$, and the population had mean = 60 and standard deviation = 3. So, now, I went through a new random generation of 200 new samples where n is still 36 but, I sampled from a widened population; ie, I made the standard deviation in the new population a value of 6. So, if the mean is 60 in the population and the standard deviation is 6, that would mean that individual wpm values could range from about 3 units of 6 on either side of 60, or roughly between about 40 and 80. Thus, this second sampling situation has held sample size constant but, nearly doubled the range or width of the parent population. Here is what the data looked like when I used Minitab to do this new simulation. Again, the comparison dotplots are shown below.



DESCRIPTIVE STATISTICS FOR MEAN WHEN SD = 3 AND SD = 6

	# Samps	MEAN	STDEV
Means SD = 3	200	59.990	0.508
Means SD = 6	200	59.897	0.996

Note: the top dotplot is the same as the bottom one from the previous graph. Again, both sampling distributions are approximately normal with the center values being near the population parameter of 60. Notice in this case however that there is less variation in the sample means (and hence less sampling error) when the population has less variability. When the population was wider to start with (in this case, with a standard deviation of 6), the sample means could also vary more widely which means more sampling error. This is also reflected in the standard deviations of the sampling distributions: for the population standard deviation of 3, the standard deviation of the sampling distribution was about .51 whereas when the population standard deviation was 6, the standard deviation of the sampling distribution distribution was about 1. Please note that you must make a distinction between the standard deviation of the larger population (from which you sample) and the standard deviation of the sampling distribution: that is the measure of error. Here we have a direct or positive relationship in that sampling error gets larger as the variability in the population increases. All this means is that if the population is homogeneous (for example, if values go from 9 to 11), sample means can only range (at a maximum) from 9 to 11 but if population variability is larger (for example, if values go from 5 to 15), then sample means have a greater range over which to vary. As a generic equation, look at the following.

$$\text{Sampling error} = \text{Population variability}$$

))))))))) =)))))))))

As population variability increases, so does sampling error.

Combining the two findings above, we have the following general rules related to sampling error of means.

1. As the **size** of the random sample **increases**, the amount of sampling error of means **decreases**.
2. As the **variability** in the population **increases**, the amount of sampling error of means **increases**.

Or, to come to the point, the overall amount of sampling error is influenced by 2 factors and is determined by the following generic formula.

$$\text{Sampling error of means} = \frac{\text{Population variability}}{\text{Sample size}}$$

So far, we have examined the impact of population variability and sample size on the relative amount of sampling error, which can be quantified by the standard deviation of the sampling distribution of means. Greater amounts of sampling error are reflected in a larger standard deviation of the sampling distribution and smaller amounts of sampling error are associated with sampling distributions with smaller standard deviations. However, in practical situations, one will never be able to do an experiment 200 or 300 times, or repeat the sampling plan (i.e., do a survey) hundreds of times. In most circumstances, the typical operating procedure is to do one experiment or one survey. Thus, if this is so, how can we estimate how much sampling error there is likely to be if only one sample is selected? Good question!

Standard Error of the Mean

From the generic formula above that indicates that sampling error is directly related to population variability and inversely related to sample size, we should be able to estimate how much sampling error there is or there is likely to be if we know what the sample size is and can also estimate how much variability there is in the population. When doing a study, you obviously will know how large the n or sample size value is so finding the denominator in the "sampling error" formula is a snap. So, the problem here is to be able to estimate what value seems reasonable to place in the numerator. Recall from descriptive statistics that one can estimate the population standard deviation by using the regular standard deviation formula and using n - 1 in the denominator (see page 27). The estimate of the population standard deviation, S, is used to provide us with a good estimate of the variability in the population. Combining the fact that you will know what the random sample size is with using S to estimate the population standard deviation, we have

the following formula specific way to estimate sampling error of means, given 1 sample.

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

$S_{\bar{X}}$, or the left side term, is the symbol for sampling error and gives us an estimate of the standard deviation of the sampling distribution of means, which is our quantitative measure of sampling error. In the literature, the standard deviation of the sampling distribution of means, our error factor, is called the **STANDARD ERROR OF THE MEAN**. Sampling error is estimated by using the standard deviation using the sample data (divisor = $n - 1$) divided by the square root of the sample size. You will just have to trust me that the denominator is the square root of n , and not simply n . Thus, from one random sample, we can estimate how much sampling error there is likely to be (how much the means will vary across repeated samples) by knowing the sample size and having an estimate of the variability in the population.

To see that data from one sample can, in fact, estimate the amount of sampling error for repeated samples, we would need to take a "typical" sample, look at the sample size, and see if we could estimate the population standard deviation from the sample. Recall that I have gone through 3 different simulations above in route to showing the relationship between sampling error, and each of factors of sample size and population variability. In each of these simulations, recall that I had 200 samples: 200 samples when $n = 9$ and population SD = 3, 200 samples when $n = 36$ and population SD = 3, and 200 samples when $n = 36$ and population SD = 6. For each of these 3 cases, our sampling distribution dotplot contained 200 mean values. But, for each of those samples, I could have calculated a standard deviation value (using $n - 1$ as the divisor) as an estimate of the population standard deviation. From each of those simulation situations, I could take an average of the 200 standard deviation estimates as my "ballpark" estimate of the population standard deviation. This is what I found when I did just that.

Simulation Condition	# SD Estimates	Average Value
$n = 9$, Pop SD = 3	200	2.89
$n = 36$, Pop SD = 3	200	2.97
$n = 36$, Pop SD = 6	200	5.92

Note that in all cases, where the actual population SD's are 3, 3, and 6 respectively, a "typical" standard deviation estimate from the batch of 200 samples (2.89, 2.97, and 5.92) was fairly close to the real SD value in the population. Therefore, when taking a random sample of some size, our estimate of the population SD is probably reasonably close to the actual parameter SD value.

Given our typical estimates of the corresponding population SD's, the value that is needed in the numerator of the standard error of the mean formula, what would a "typical" standard error of the mean, or $S_{\text{sub } X \text{ bar}}$, be? To get a typical estimate, we could divide each of the typical SD estimates by the square root of the sample size. See each calculation below.

Sim Condition	Typical SD	Exp StanErr of Mean	SD of Means
n = 9, Pop SD = 3	2.89	$2.89/\text{Sqrt } 9 = .96$.983
n = 36, Pop SD = 3	2.97	$2.97/\text{Sqrt } 36 = .50$.508
n = 36, Pop SD = 6	5.92	$5.92/\text{Sqrt } 36 = .99$.996

In the first simulation, using our typical estimate of the SD in the population, a typical $S_{\text{sub } X \text{ bar}}$ or standard error of the mean estimate would be about .96. Compare that to the actual SD of the means from the relevant sampling distribution and you will see that our typical value was very close. In the second simulation, our typical standard error estimate was .50 when the actual sampling distribution SD was .508. Again, pretty close. Finally, in the last simulation, our typical standard error estimate was .99 when the actual SD of the sampling distribution was .996. Again, very close. Thus, the bottom line is that when you take a random sample of fixed size from the population, we are in a fairly good position to estimate how much sampling error there might be in the long run, even though you have taken only 1 sample. The next Chapter will go into more detail as to how this estimate of sampling error of the mean is used.

As one last example in this extended discussion of sampling error of means, it may be helpful to look at a "typical" one sample case that resembles what a data gatherer may really do. Look at the following. Assume that most golf scores for 18 holes range from about 80 to 110 strokes (honestly reported let's assume!) with a mean of about 95. Also assume for a moment that this represents population data. What if you took a random sample of 100 golfers from that population and examined this set of their typical golf scores. The data and descriptive statistics might look like as follows.

RANDOM SAMPLE OF 100 GOLF SCORES

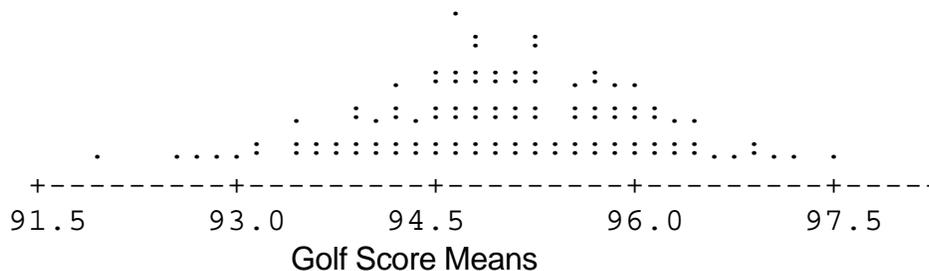
81 101 91 105 103 103 110 95 101 110 82
 83 89 98 101 99 101 96 85 107 95 100
 85 91 91 93 107 101 94 92 87 109 98
 91 88 98 80 81 85 81 96 96 104 84
 102 107 80 105 102 84 80 99 83 92 82
 105 100 84 80 101 92 82 106 105 101 93
 107 87 100 91 91 97 84 87 101 110 81
 92 101 85 85 89 109 109 108 97 84 102
 98 104 85 107 103 107 100 96 91 89 86
 94

MEAN OF SAMPLE = 94.90
SD OF SAMPLE = 9.04

In this case, the sample size is 100 and the estimate of the population standard deviation is about 9.04. To find our estimate of sampling error or the standard error of the mean, we need to divide the estimate of the population standard deviation value of 9.04 by the square root of the sample size which would be square root of 100 or 10. In this case, since we are dividing by 10, all we have to do is to move the decimal place one spot to the left. Thus, our estimate of the standard error of the mean, based on this one random sample, is:

$$\text{Estimate of Standard Error of Mean} = 9.04/10 = \mathbf{.904}$$

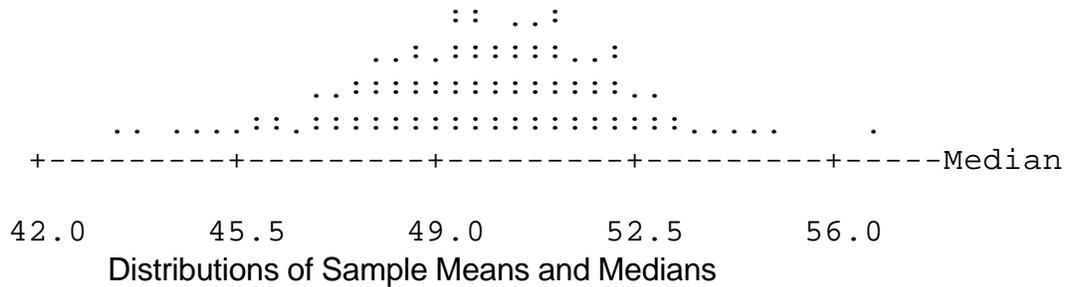
If in fact, the sampling distribution of means was centered around 95 as a score, with a standard deviation (standard error of the mean) of about .9, then that distribution of the potential variation in means would be similar to the following.



Although the "typical" sample mean should be close to 95, which is the parameter, any individual mean could vary from about 92 up to about 98 (approximately 3 units of .9 on either side of the parameter of 95). So, even though the true mean is 95, sample means will not always show that. Of course, if we had taken larger samples, the sample means would be closer to 95 since there is less sampling error. However, if samples had been smaller, then the means would have varied more around the true value of 95.

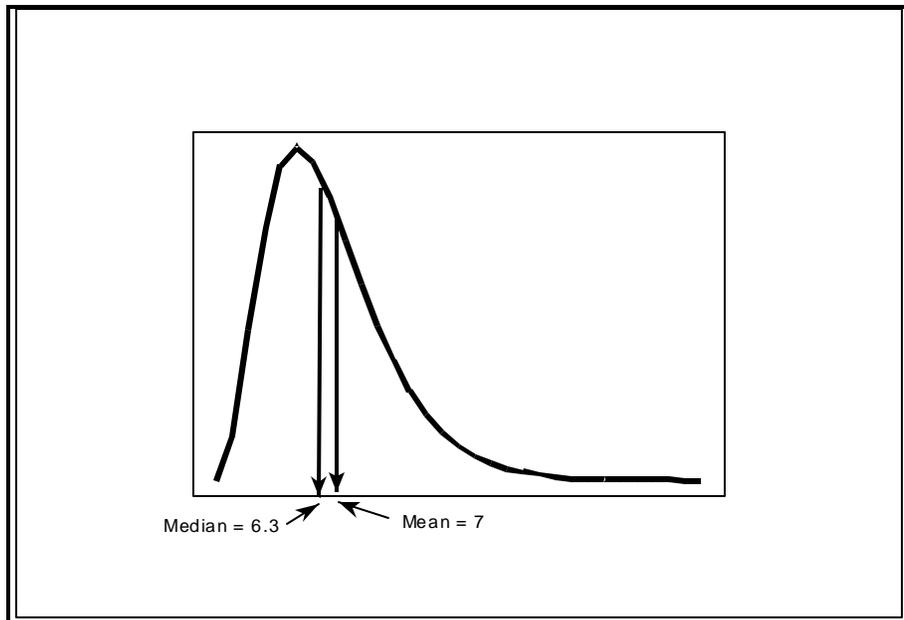
Statistics as Estimators of Parameters

In this Chapter, we have used the mean (statistic) of a sample to estimate the mean (parameter) of the population. Without too much thought, this seems to make sense: use the statistic that is the same thing as the parameter we are trying to estimate. However, in the case of trying to estimate the population mean, one may opt for (either deliberately or because there is no other way) using a different statistic to estimate the population mean. In the case of central tendency, recall that in some distributions, the measures of central tendency may all have the same value. Thus, if you wanted to know the mean of this set of data, you may - as an alternative - use the median or the mode as your best estimate of the mean of the distribution. In some cases, using the median and/or mode may not be much different than using the mean.



As you can see from the dotplot of the medians (compared to the one for the means) and the descriptive data, the sample medians on average would also have provided a good estimate of the population mean. In this case, it appears that if you want to estimate the population mean, you could use either the sample mean or the sample median to make that estimate. So, if either one could be used, how do you decide which one? Well, remember that the concept of sampling error referred to the width of the sampling distributions. Wider sampling distributions means more sampling error. Certainly, it would seem to make sense to use the statistic from the sample that tends to produce less sampling error when estimating the population parameter. Look at the standard deviations from the two sampling distributions above. For the means, the standard deviation of the sampling distribution or our standard error of the mean is about 1.58 whereas that same value for the sampling distribution of the medians is about 2.02. As a rough guide, the sampling distribution in this case is only about $1.58/2.02 = 78\%$ of the width of the one for the median. Thus, sample means will tend not to vary as much around the population mean as will medians. While the dotplots don't show this too well, the standard deviation of the sampling distribution of medians is larger indicating more sampling error. Thus, while in the long run we could use the sample median as an estimate of the population mean, the sizes of the sampling errors would tend to be larger in these cases. In the literature, this ratio of the smaller standard error distribution to the larger standard error distribution is referred to as the **RELATIVE EFFICIENCY** of the statistic when estimating the parameter. In this case, we would opt for the sample mean as a better estimator of the population mean since it more efficiently (tends to be closer to) estimates the parameter of interest.

Another issue of concern here is called **BIAS** in the estimator. In the example above, our expectation is that the long run average of the sample median values will be equal to the true population mean value, just as will be the long run average of the sample mean values. In this case, neither statistic - sample mean nor sample median - would be expected to exhibit bias when estimating the population mean. However, recall that when the simulations were done, I sampled from a population that was normally distributed. In this case, both the parameter mean and the parameter median were the same values. But, what if the population happened to have different values for the mean and median? For example, what if the population were actually positively skewed? See below.



Recall that when there is skewness present, the mean and median will be offset from one another. In the case of a positive skew (say the population mean = 7 and the population median = 6.3), where the trailing off side is to the right, the mean is more pulled to the right compared to the median. Thus, in this case, the median will tend to be a lower value than the mean. What would happen in this case if we decided to use the medians from random samples as our best estimates of the population mean? If the population is positively skewed, then samples will also tend to be positively skewed. Thus, medians in samples will also tend to be less than the mean values in these same samples, and on average, would tend to center around 6.3. If we use the median values from these samples to estimate the population mean, we will systematically tend to underestimate the population mean. This is when we would say that the statistic (median in this instance) is negatively biased since it would tend to underestimate the parameter. If the population had been negatively skewed to start, medians would then tend to be positively biased. In either case, the median would not provide us with as accurate estimations as would the sample means. Certainly, in this instance, we would prefer to use the sample mean as our estimator of the parameter because the sample mean will be unbiased. Therefore, the bias (if any) of an estimator of a parameter should be considered when making a decision about which specific statistic to use as your estimator.

Finally, while we have only examined the issue of estimating the mean in a population (using either the sample mean or sample median), there are many other parameters that we may want to estimate. For example, if you were interested in the correlation between IQ and school performance in the population, you may use the sample correlation as the estimator. Or, what if you were interested in the variance of IQ scores in the population? In this case, you may use the variance of your sample as your estimator. In

either case, one needs to ask about relative efficiency or bias if there are competitor estimators available.

Practice Problems

1. If your stat package allows it, put 100 values generated at random in each of 9 columns where the population is normal with mean = 50 and standard deviation = 10. Print some of the values to see what random sampling brings. Then, using rows as samples, find the means of each of the 100 samples and put the mean in another column. What does the distribution of 100 means look like? What are the descriptive characteristics of that mini-sampling distribution? What would have happened to that distribution if n had been larger than 9 and the standard deviation had been smaller than 10?
2. Assume that you took a random sample of 30 cars as they passed by stop signs and measured the speed (MPH) with which they went through the sign. The data are as follows.

4, 5, 1, 0, 5, 8, 3, 0, 0, 3, 4, 8, 2, 10, 3, 4, 5, 8,

1, 0, 0, 3, 4, 7, 2, 3, 5, 7, 0, 3

Based on these data, find the standard error of the mean. How widely are the means of other random samples of 30 cars likely to vary around the true stop mean?

3. Assume that the correlation between X and Y , out in the larger population, is approximately .8. You take random samples from that population and then calculate the sample correlation between X and Y in each of these samples. In this situation, would you expect there to be any bias in the sample correlations, and if so, what type?