

The Central Limit Theorem and Inferential Statistics (Soc 301)

In the last notes, we discussed how, if you assume the distribution for a variable in the population is normal, then we can use a z table to deduce probabilities of obtaining a sample member with a value on the variable within any given range. The first step in the process of conducting statistical inference is to extend this idea of determining probabilities of obtaining a sample member with a particular range of values on the variable to determining probabilities of obtaining an entire *sample* of a given size with a given value of a *statistic* (e.g., the mean) from the population. For example, if we know the population mean, μ , is 10, we might want to know the probability of obtaining a sample that has a mean, \bar{x} , of 8 or less.

In order to understand this process, we must discuss properties of statistics (like the mean) that can be drawn from a population. One of the most important theorems in statistics, the Central Limit Theorem (CLT), states that, as sample sizes increase, regardless of the distribution of a random variable in the population, sample means (\bar{x} s) follow a normal distribution with a mean equal to the population mean (μ) and a variance equal to the variance in the population divided by the sample size used to compute the mean (σ^2/n). Equivalently, the standard deviation of the distribution of sample means is equal to σ/\sqrt{n} .

This theorem is conceptually difficult but forms the basis for statistical inference, and so some extended explanation is warranted. When you take a sample from a population, instead of thinking of this as a collection of individuals, consider that you are sampling a *mean* from the population. Obviously, just as there are an infinite number of samples of size n that you could draw from the population, there are an infinite number of sample means that you can draw (one for each of the samples of size $n!$). The CLT says that, if you put all of the means of all of the samples of a given size, and if that sample size is large enough, the distribution of these sample means will be normal. Furthermore, the mean of this distribution of sample means (called the sampling distribution) will be equal to the overall population mean, and the variance of the distribution of sample means will be equal to the variance of the population divided by the sample size.

In order to make these ideas concrete, I used the data on age from homework 2, treated this data as if it were a population, and from it I drew 1,000 random samples each of size $n=2$, $n=5$, $n=10$, $n=30$, $n=50$, $n=100$, and $n=500$. For each of the samples, I computed the mean. So, for samples of size $n=2$, I computed 1,000 means, for samples of size $n=5$, I computed 1,000 means, etc., etc. The following figure shows histograms of these collections of means by sample size.

The upper left plot is the histogram of the original “population” distribution of age. The upper right plot is the histogram of 1,000 sample means—called the sampling distribution—for the samples of size $n=2$. The sampling distribution for the mean when $n=2$ looks much like the original population distribution: it is skewed strongly to the right. However, as the figure shows, as the sample size increases, the sampling distributions become much

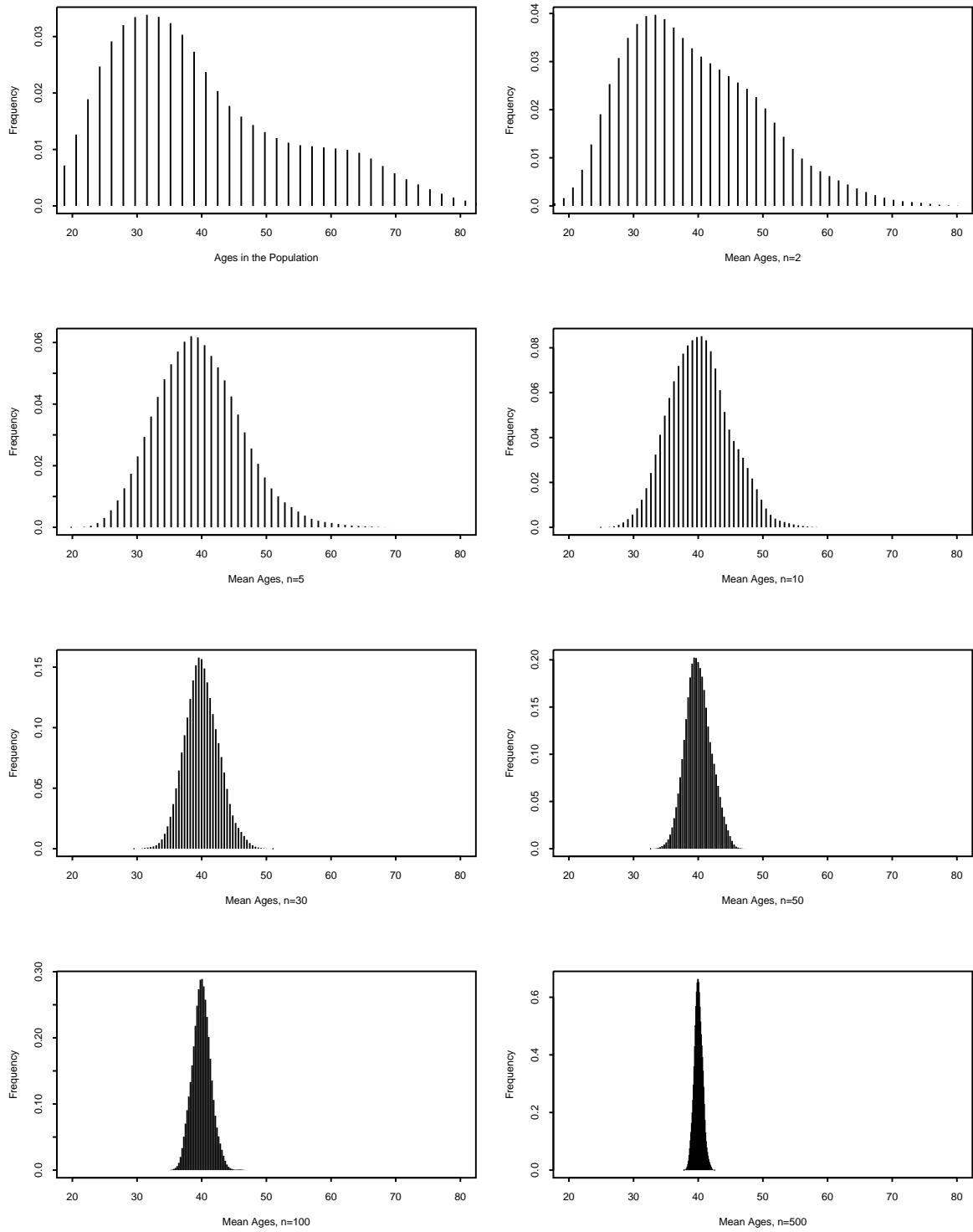


Figure 1: Histograms of Sampling Distributions for Various Sample Sizes

more symmetric (normal), and their variance decreases. Ultimately, when the sample size is $n=500$, the sampling distribution is very narrow.

The table below shows the means of the sampling distributions, as well as the observed standard deviations of these sampling distributions (called the “standard error”—abbreviated S.E.). The far right column shows the theoretical standard error based on the CLT. Notice that the means of the sampling distributions are very close to the mean of age in the population, and that, when the sample size is 30, the mean of the sampling distribution is consistently within rounding of the true mean. Also notice that the observed S.E. consistently matches the theoretical S.E. when the sample size is 100 or larger.

Sample Size	$\mu_{\bar{x}}$	Observed S.E.	Theoretical S.E. (σ/\sqrt{n})
(Actual Population)	40.0	14.0	14.0
2	39.4	10.2	9.9
5	39.7	6.3	6.3
10	40.2	4.5	4.4
30	40.0	2.5	2.6
50	40.0	1.9	2.0
100	40.0	1.4	1.4
500	40.0	.6	.6

Why do the means of the sampling distribution become normal as the sample size increases, and why does the variance of the sampling distribution decrease? If you recall from the previous set of notes, the joint probability of two independent events is the product of their respective probabilities. When we take a simple random sample from the population, we are taking a collection of independent draws from the population distribution. Thus, their joint probability can be computed as the product of each of their probabilities. When the sample size is small, it may not be unreasonable that we could draw two extremely rare observations—observations that are far from the population mean—and thus give us a sample mean that is far from the true mean. For example, the probability of obtaining two individuals whose values are rare enough that their probability of occurrence is .1 each is $.1 \times .1 = .01$. This is a small probability, but not incredibly small. However, it is extremely unlikely, if we draw a large sample, that we would draw a series of very rare values. For example, the probability of drawing five rare people like the ones we just discussed is $.1 \times .1 \times .1 \times .1 \times .1 = .00001$. This is an incredibly small probability. The implication is that it will be unlikely in a large sample to draw a large number of very rare individuals, and so our sample mean will be more likely to be close to the true mean.

The implication of these results for making inference about population means using sample means is that, if the sample size is large enough, we can very accurately estimate the population mean. Furthermore, we can quantify the probability of obtaining a sample with a given mean if we know the true population mean, again using a z table. Here, the calculation of z is only slightly different:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Since we know that the sampling distribution for \bar{x} is normal, and we know its standard deviation (more on whether we 'know' this later), we can also evaluate the probability of obtaining a sample with a given mean under an *assumption* about the population value of μ . This is the basis of hypothesis testing, which we will discuss next.