DOCUMENT RESUME

ED 393 939 TM 024 976

AUTHOR	Breunig, Nancy A.
TITLE	Understanding the Sampling Distribution and Its Use in Testing Statistical Significance.
PUB DATE	9 Nov 95
NOTE	25p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Biloxi, MS, November 1995).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Computer Uses in Education; *Educational Research; *Hypothesis Testing; *Sampling; Statistical Distributions; Statistical Inference; *Statistical Significance; Test Results
IDENTIFIERS	Bootstrap Methods; *Central Limit Theorem

ABSTRACT

Despite the increasing criticism of statistical significance testing by researchers, particularly in the publication of the 1994 American Psychological Association's style manual, statistical significance test results are still popular in journal articles. For this reason, it remains important to understand the logic of inferential statistics. A fundamental concept in inferential statistics is the sampling distribution. This paper explains the sampling distribution and the Central Limit Theorem and their role in statistical significance testing. Included in the discussion is a demonstration of how computer applications can be used to teach students about the sampling distribution. The paper concludes with an example of hypothesis testing and an explanation of how the standard deviation of the sampling distribution is either calculated based on statistical assumptions or is empirically estimated using logics such as the "bootstrap." These concepts are illustrated through the use of hand generated and computer examples. An appendix displays five computer screens designed to teach these topics. (Contains 1 table, 4 figures, and 20 references.) (Author/SLD)

Understanding the Sampling Distribution and its Use in Testing Statistical Significance

Nancy A. Breunig

Texas A&M University

Paper presented at the annual meeting of the Mid-South Educational Research Association, Biloxi, MS, November 9, 1995.

Abstract

Despite increasing criticism of statistical significance testing by researchers, particularly in the publication of the 1994 American Psychological Association's style manual, statistical significance test results are still popular in journal articles. For this reason, it remains important to understand the logic of inferential statistics. A fundamental concept in inferential statistics is the sampling distribution. This paper explains the sampling distribution and the Central Limit Theorem and their role in statistical significance testing. These concepts are illustrated through the use of hand generated and computer examples.

Understanding the Sampling Distribution and its Use in Testing Statistical Significance

In recent years, statistical significance testing has been increasingly criticized by researchers. In fact, the Journal of Experimental Education has an entire issue dedicated to a discussion of statistical significance testing (Thompson, 1993a). Articles within the journal provide explanations of what statistical significance testing actually does and why people have persisted in using it (Shaver, 1993, p. 293). In addition, they present the reader with alternatives to statistical significance testing (Thompson, 1993b) or at a minimum suggest that effect size should be reported along with statistical significance (Carver, 1993). According to Thompson (1994b), as scientists, the questions that should be of concern when engaging in statistical significance testing are "(a) what the magnitude of sample effects are and (b) whether these results will generalize." (p. 6) Unfortunately, statistical significance testing does not answer either of these questions (Thompson, 1994a).

Despite the concerns raised about statistical significance testing by researchers, and the fact that the Publication Manual of the American Psychological Association (1994, p. 18) itself alerts the researcher of the limitations of statistical significance testing and encourages one to provide effect size information, statistical significance test results are still popular in journal articles. For this reason, it remains important to understand the logic of statistical significance testing.

The purpose of this paper is to explain the sampling distribution which is one of the fundamental concepts underlying all inferential procedures (Chalmer, 1987; Freund & Smith, 1986; Hinkle, Wiersma, & Jurs, 1994; Mohr, 1990). A definition and explanation of the sampling distribution and its relation to statistical significance testing will be provided. Included in the discussion is a demonstration of how computer applications can be used to teach students about the sampling distribution and the Central Limit Theorem. The paper concludes with an example of hypothesis testing and an explanation of how the standard deviation of the sampling distribution is either calculated based on statistical assumptions, or is empirically estimated using logics such as the "bootstrap".

Chain of Reasoning in Inferential Statistics

When conducting statistical significance testing, the researcher is trying to infer something from the sample being observed. This is why statistical significance testing is called inferential statistics. Thus, there are generally two tasks of inferential statistics. The first task is to test hypotheses about parameters. The second task is to use statistics (descriptive measures of a sample) to make statements about or to estimate parameters (descriptive measures of a population). The parameters are unknown and that is why inferences need to be made about them (Chalmer, 1987; Hinkle et al., 1994). For example, if a representative sample of undergraduate and graduate students at a major university spend an average of two hours per day during a semester in the student center, we might correctly infer that all students at the university spend approximately two hours per day per semester in the student center.

Hinkle et al. (1994) describe a chain of reasoning for inferential statistics which is illustrated in Figure 1. They state that the first step in inferential statistics is to draw a randomly selected (c. at least a representative) sample. A randomly selected sample is one in which

"...every member of the population has an equal chance of being selected." (Mattson, 1981, p. 75).

The sample needs to be a random sample because we are trying to make inferences about the population from the sample. If the sample is not randomly selected we may be introducing systematic bias into the sample, which can be either intentional or unintentional. A biased sample would not give us accurate information about the population and the population is what we are interested in (Mattson, 1981). For example, if you wanted a law to be passed that only English could be spoken in the classroom, you might intentionally choose to sample only those people that you knew did not support bilingual education. Thus, your sample results would make it appear that the majority of people in the United States supported your position and the law would be passed. Unintentional systematic bias could exist if you decided to sample your population by taking the first 200 people listed in the phone book. In this case, there would be many sources of potential bias, such as you're only accessing people who have telephones or who are listed in the telephone directory.

According to Hinkle et al. (1994), the second step in the chain of reasoning for inferential statistics is that "...the estimate from this sample must be compared to an *underlying distribution* of estimates from all other samples of the same size that might be selected from the population" (p. 147). An underlying distribution can be defined as, "...the distribution of all possible outcomes of a particular event" (Hinkle et al., 1994, p. 138).

The third step in inferential statistics involves making inferences based on the comparison and probability of the sample with statistics with the underlying distribution of the statistic when random sampling has been employed (Hinkle et al., 1994). The sampling distribution is this underlying distribution of the statistic (Hinkle, et al., 1994, p. 149).

The Sampling Distribution

A formal definition of a sampling distribution provided by Hinkle et al. (1994), "...is the distribution of all values of the statistic under consideration, from all possible random samples" (p. 149). The sampling distribution most commonly seen in textbooks is the sampling distribution of the mean; however, the reader should be aware that you can have a sampling distribution of any statistic such as the sampling distribution of the median or standard deviation.

Sampling distributions can be derived either empirically or theoretically. Most sampling distributions of a statistic have already been established theoretically; however, to understand the concept of a sampling distribution it is useful to demonstrate empirical methods of deriving these distributions (Mattson, 1981). The empirical methods can consist of hand calculations or, if this would be too lengthy of a process, which is often the case, computer applications can be utilized.

To illustrate the concept of a sampling distribution using hand calculations, consider constructing the sampling distribution for the mean of a random sample of size $\underline{n} = 2$, from the finite population of size $\underline{N} = 5$. The elements of the population will be the numbers 2, 4, 6, 8, and 10. The mean of the population is:

$$\mu = \frac{2 + 4 + 6 + 8 + 10}{5} = 6$$

and

$$\sigma = \sqrt{\frac{(2-6)^2+(4-6)^2+(6-6)^2+(8-6)^2+(10-6)^2}{5}}$$

$$=\sqrt{8}$$
$$=2.8$$

Taking random samples of $\underline{n} = 2$ from the population, there are 10 equally probable possibilities:

2 and 4, 2 and 6, 2 and 8, 2 and 10, 4 and 6

4 and 8, 4 and 10, 6 and 8, 6 and 10, 8 and 10

The mean of the first sample is: (2 +4)/2 = 3. The remainder of the means for each sample may also be calculated, yielding the following values: 4, 5, 6, 5, 6, 7, 7, 8, and 9. If sampling is random, so that each sample statistic has the probability 1/10 (each outcome [1] divided by the number of equally likely outcomes [10]), the sampling distribution of the mean would be as shown in Table 1.

This example illustrates two important points. First, the mean of the sampling distribution equals the mean of the population, $\mu_{\underline{M}}$ which equals $\mu = 6$. In addition, the standard deviation of the sampling distribution of the mean is smaller than the standard deviation of the population, $\underline{SDM}$ = 1.73 < σ = 2.8 (Hinkle et al., 1994). In this example, $\sigma_{\bar{x}} = 1.73$ is the standard error of the mean, i.e., the standard deviation of the sampling distribution.

This was just one example with a very small sample and a very small population. You can create many such examples yourself, and you will see that the expected value of all possible values of the sample mean from random samples of size $\underline{n}$ equals the mean of the population. That is, $\mu_{\underline{M}} = \mu$. In addition, the standard error of the mean (standard deviation of the sampling distribution of the statistic) is always less than or equal to the standard deviation of the parent population (Moore & McCabe, 1989).

Upon closer examination, one sees that the standard error of the mean increases as the variability of the population increases and decreases as the sample size increases. Thus, the standard error of the statistic is directly proportional to the standard deviation of the population. The formula for the standard error of the mean is (Freund & Smith, 1986, p.274):

$$\underline{SD_M} = \sigma / \sqrt{n}$$

Central Limit Theorem

Researchers do not normally calculate sampling distributions but instead use theoretical sampling distributions which are defined by mathematical theorems. One such theorem is the Central Limit Theorem (CLT). The CLT states that: given a population with a mean equal to μ and variance equal to σ^2, as sample size ($\underline{n}$) increases, the sampling distribution of the mean for simple random samples of $\underline{n}$ cases will approach the normal distribution (Hinkle et al., 1994; Howell, 1987). Thus, as is illustrated in Figure 2, with very small samples the shape of the distribution will depend on the shape of the parent population, but with samples of $\underline{n} = 30$, even a skewed parent population will result in a normal distribution (Harnett, 1975; Hinkle et al., 1994). Schulman (1992) refers to this phenomenon as the "magic of the normal distribution" (p. 19) and states that without this magic, most of statistics would be limited to applications where it could be demonstrated that the population had a normal distribution.

In order to discuss the CLT, the concept of the normal probability distribution must be understood. The normal probability distribution has three properties:

1. A normal distribution histogram is unimodal (has one mode or peak) and it is symmetrical (i.e., the part of the curve to the right of the mean is a mirror image of the part to the left). It's coefficients of skewness and kurtosis are both zero (Bump, 1991).

2. The normal distribution is continuous. This means that for every value of x there is a value for y and the total area underneath the curve is equal to 100 percent (Chalmer, 1987; Hinkle et al., 1994).

3. "The normal distribution is asymptotic to the X axis" (Hinkle et al., 1994, p. 88). The farther away from the mean the curve is, the more the curve approaches the X axis without actually ever touching it (Hinkle et al., 1994; Mittag, 1992).

Examples

Hand Calculated Example

To illustrate the concept of the CLT, we can refer to the example used above for the sampling distribution. As a reminder the population consisted of the numbers 2, 4, 6, 8, and 10. A histogram of the population is illustrated in Figure 3. As can be seen from the histogram the population is not a normal distribution. Now if we look at a histogram of the sampling distribution, which is illustrated in Figure 4, we see that it is approaching a normal distribution.

Computer Examples

Computer programs have greatly advanced the teaching of statistical concepts (Freund & Smith, 1986; Mitaag, 1992; Schulman, 1992; Yang & Robinson, 1986). To better illustrate the concepts of the sampling distribution and CLT, a computer application can be used. The computer application to be used in this paper was developed by James Lang. Hypercard version 2.0 is needed to run the program on a Macintosh computer. Samples of the program are provided in the Appendix. In the first example shown in the Appendix, 200 samples were taken of a sample size of $\underline{n}$ = 10. The mean of the population (μ) was 4.5 with a standard deviation (σ) of 2.87. The mean of the sampling distribution ($\mu_{\underline{M}}$) was 4.46 and the standard error of the mean (SDM) was .88. This is consistent with CLT because the standard error of the mean should have equaled $2.87/\sqrt{10}$ and it did.

In the second example, 200 samples were taken with a sample size of $\underline{n}$ = 20. The population parameters remain the same ($\mu = 4.5$, $\sigma = 2.87$) but the mean of the sampling distribution ($\mu_{\underline{M}}$) was 4.44 and the standard error of the mean (SDM) was .64. Again, this result is consistent with CLT. The standard error of the mean (SDM) should have equaled $2.87/\sqrt{20}$ and it did.

In the third example, various sample sizes ($\underline{n}$) were chosen and then the computer took 500 samples of the selected sample sizes ($\underline{n}$) from the population. Means and standard deviations of the sets of sample means were generated. This example demonstrates the relationship of the mean ($\mu_{\underline{M}}$) and standard deviation of the sample means (SDM) to the mean (μ) and standard deviation (σ) of the population. As sample size increases, the mean ($\mu_{\underline{M}}$) better approximates the population mean (μ). At sample size $\underline{n}$ = 2, the mean of the sample means was 4.34 and the population mean was 4.5. At sample size $\underline{n}$ = 100, the mean of the sample means ($\mu_{\underline{M}}$) was 4.50104. As one would expect, the standard error of the means also decreased as sample size increased. At sample size $\underline{n}$ = 2, the standard error of the mean (SDM) was 2.11. Whereas, at sample size n = 100, the standard error of the mean (SDM) was .28.

CLT and Statistical Significance Testing

Overview

How do the concepts of the sampling distribution and the CLT relate to statistical significance testing? As was stated at the beginning of the paper, in statistical significance testing, the researcher is trying to make inferences about the population based on a random sample drawn from the population. When this random sample is drawn and a statistic such as the mean (M) is computed, the statistic represents both the parameter of the population and sampling error. Statistical significance testing involves determining the magnitude of the difference between the statistic and the hypothesized value of the parameter. Once the researcher determines the magnitude of the difference, he/she makes a judgment as to whether this difference is "statistically significant" or not. In otherwords, the researcher decides to either reject or fail to reject the null hypothesis (Hinkle et al., 1994).

Steps in Hypothesis Testing

In order to better understand the role of the sampling distribution in statistical significance testing, the actual steps of hypothesis testing will be summarized and an example will be provided. When engaging in statistical significance testing, the researcher first states the null hypothesis. The null hypothesis states that their is no relationship or difference (Hinkle et al., 1994). For example, if it is believed that the mean weight of male college professors is 170, the null hypothesis would be:

$\underline{H}_O$: μ = 170.

The second step in statistical significance testing is to set the criterion for rejecting the $\underline{H}_O$. In order to do this, the researcher must select a level of statistical significance which is the probability of making a Type I error. A Type I error is when a researcher rejects a null hypothesis that is actually true. The most common levels of significance selected are .05 and .01 (Hinkle et al., 1994). According to Hinkle et al. (1994), "The level of significance represents a proportion of area in a sampling distribution that equals the probability of rejecting the null hypothesis if it is true. This area of the sampling distribution is called the region of rejection" (p. 171). Using the above example that the mean weight of male college professors is 170, if we selected a random sample of $\underline{n}$ = 144 male college professors to test our hypothesis and the sample mean ($\underline{M}$)= 166, we would have to use the sampling distribution to decide whether the difference between the sample mean and the hypothesized population mean is large enough to reject the null hypothesis. The sampling distribution for this example is the theoretical distribution of all possible samples of size $\underline{n}$ = 144 of male college professors' mean weight. Due to the Central Limit Theorem, since the sample size is reasonably large, we know that the distribution of sample means for this example is normal. We would state that the mean of the distribution equals the population mean (μ = 170), and the standard error of the mean ($\underline{SDM}$) equals 1.67 if the population standard deviation (σ) is 20.

standard error of the mean ($\underline{SDM}$) $= \sigma / \sqrt{n} = 20 / \sqrt{144} = 1.67$

If the population standard deviation (σ) was unknown then the sample standard deviations would have to be used to estimate the population standard deviation (σ). Before the use of computers, researchers had to rely on statistical assumptions to calculate the standard error. Now, several microcomputer programs exist that allow the researcher to use bootstrap logic to estimate the standard error (Reinhardt, 1992). Conceptually, bootstrap methods copy the data set over and over again, infinitely many times, to create a mega data set. Resampling from the original data set with replacement occurs. Thus, large numbers of samples are drawn from the mega file with statistics calculated for each new sample and then all the statistics are averaged (Thompson, 1993b, p. 369). As Reinhardt states (1992), "computer-intensive bootstrap methods can provide estimates for the standard error of results by using the actual data, rather than relying on the assumption that the sampling error is normally distributed..." (p. 15).

The third step is to compute the test statistic. The formula for the test statistic is:

test statistic = statistic - parameter/standard error of the statistic

In our example: test statistic = (166 - 170)/1.67. Thus, the test statistic is equal to -2.4. The test statistic indicates the number of standard errors the observed sample statistic (M) is from the hypothesized parameter (μ). This test statistic is then compared to the critical value found in the appropriate table. The critical values indicate the beginning values for the region of rejection of the sampling distribution. If the test statistic exceeds the critical value the null hypothesis is rejected (Hinkle et al., 1994). In our example, using the .05 level of significance for a twotailed test, the critical values are $\pm$ 1.96. Thus, the null hypothesis would be rejected because -2.4 exceeds the critical value of -1.96.

Conclusion

It is clear that the sampling distribution is a fundamental concept in statistical significance testing. Computer applications, such as the one illustrated in this paper, can be helpful in understanding the role of the sampling distribution and statistical assumptions such as the Central Limit Theorem in inferential statistics. In addition, computer-intensive bootstrap methods can be used to estimate the standard deviation of the sampling distribution when population parameters are unknown, using the actual data rather than having to rely on statistical assumptions (Reinhardt, 1992).

References

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington, DC: Author.

Bump, W. (1991, January). The normal curve takes many forms: A review of skewness and kurtosis. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio.

Carver, R. P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61, (4), 287-292.

Chalmer, B. J. (1987). Understanding statistics. New York, NY: Marcel Dekker.

Freund, J. E. & Smith, R. M. (1986). Statistics a first course (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Hinkle D. E., Wiersma W., & Jurs, S. G. (1994). Applied statistics for the behavioral sciences. Boston: Houghton Mifflin.

Howell, D. C. (1987). Statistical methods for psychology (2nd ed.). Boston: Duxbury Press.

Lang, J. (1991). Hypercard statistics stack [Computer program]. Orlando, Florida: Valencia Community College.

Mattson, D. E. (1981). Statistics: Difficult concepts, understandable explanations. St. Louis, C. V. Mosby.

Mittag, K. C. (1992, April). Using computers to teach the concepts of the central limit theorem. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Mohr, L. B. (1990). Understanding significance testing. Newbury Park, CA: Sage.

Moore, D. S., & McCabe, G. P. (1989). Introduction to the practice of statistics. New York: W. H. Freeman.

Reinhardt, B. M. (1992, April). Estimating result replicability using double cross-validation and bootstrap methods. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Schulman, R. S. (1992). Statistics in plain english. New York, NY: Van Nostrand Reinhold.

Shaver, J. P. (1993). What statistical testing is, and what it is not. Journal of Experimental Education, 61, (4), 293-316.

Thompson, B. (Ed.). (1993a). Statistical significance testing in contemporary practice. Journal of Experimental Education, 61, (4).

Thompson, B. (1993b). The use of statistical significance tests in research: Bootstrap and other alternatives. Journal of Experimental Education, 61, (4), 361-377.

Thompson, B. (1994a, April). Common methodology mistakes in dissertations, revisited. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Thompson, B. (1994b). The concept of statistical significance testing. Measurement Update, 4 (1), 5-6.

Yang, M. C., & Robinson, D. H. (1986). Understanding and learning statistics by computer. Philadelphia, PA: World Scientific.

Appendix

x sampling distribution

The purpose of this program is to let you watch the random process that leads to the sampling distribution for the sample mean. You may also compare the results of the program to the theoretical statement called the Central Limit Theorem.

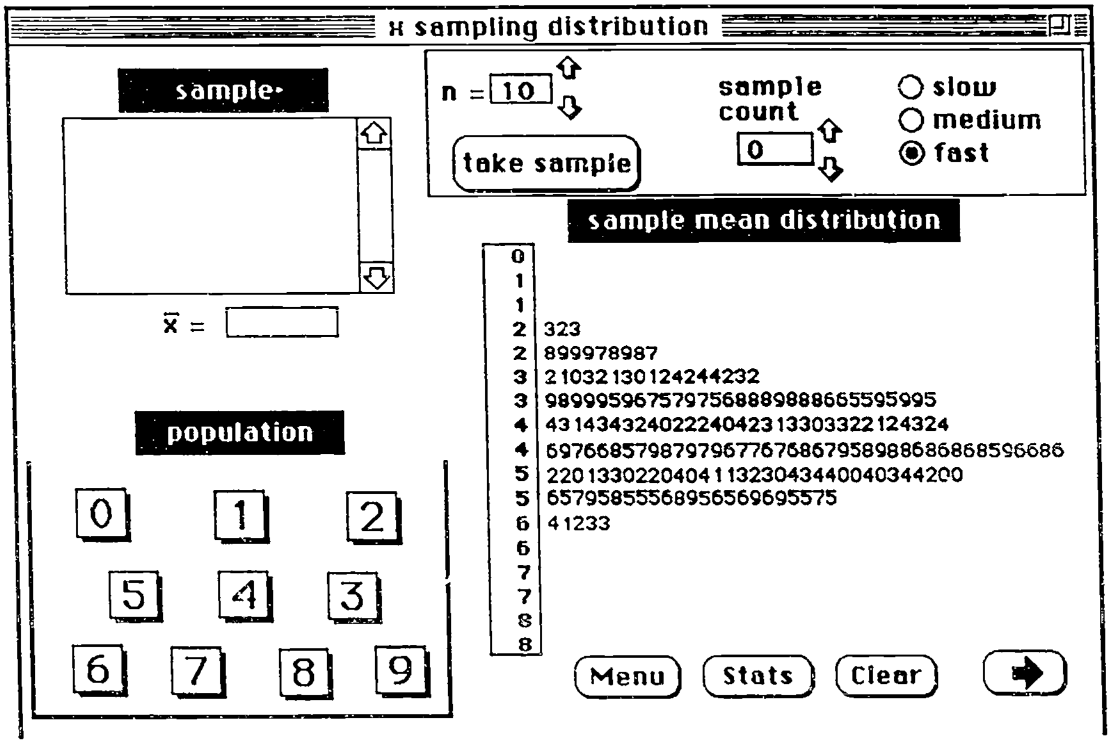
Click on the population below that you want to sample.

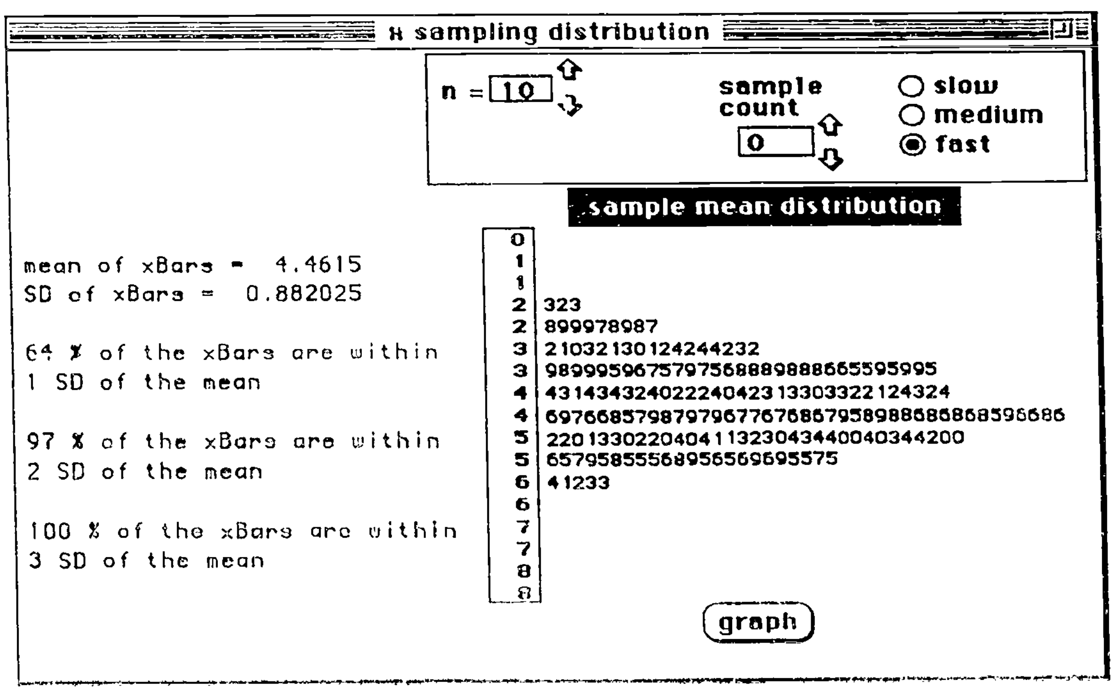
Population: 10,11,...29

Population: 0,1,2, ... ,9

Menu

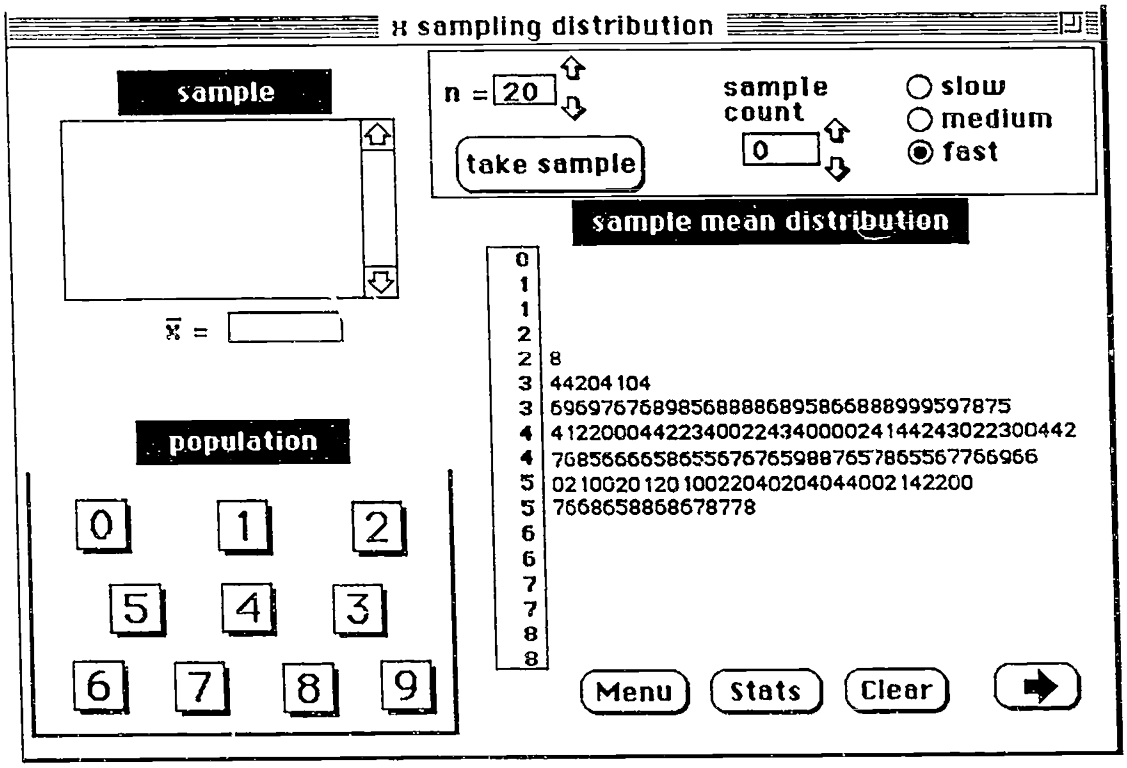
Example 1

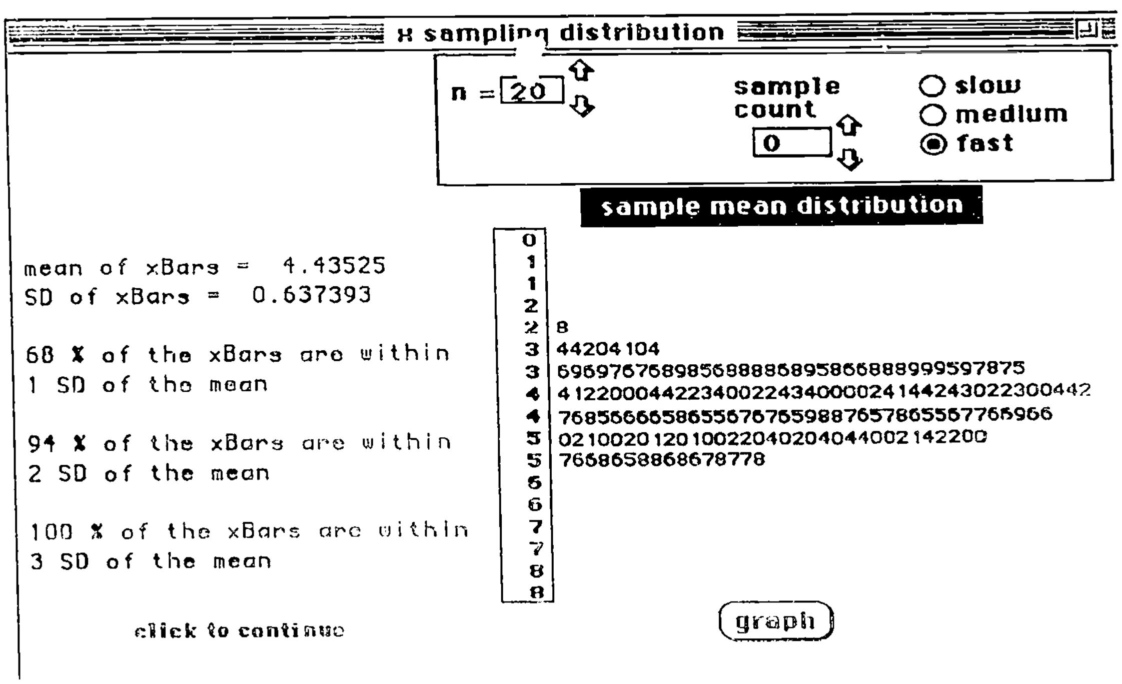




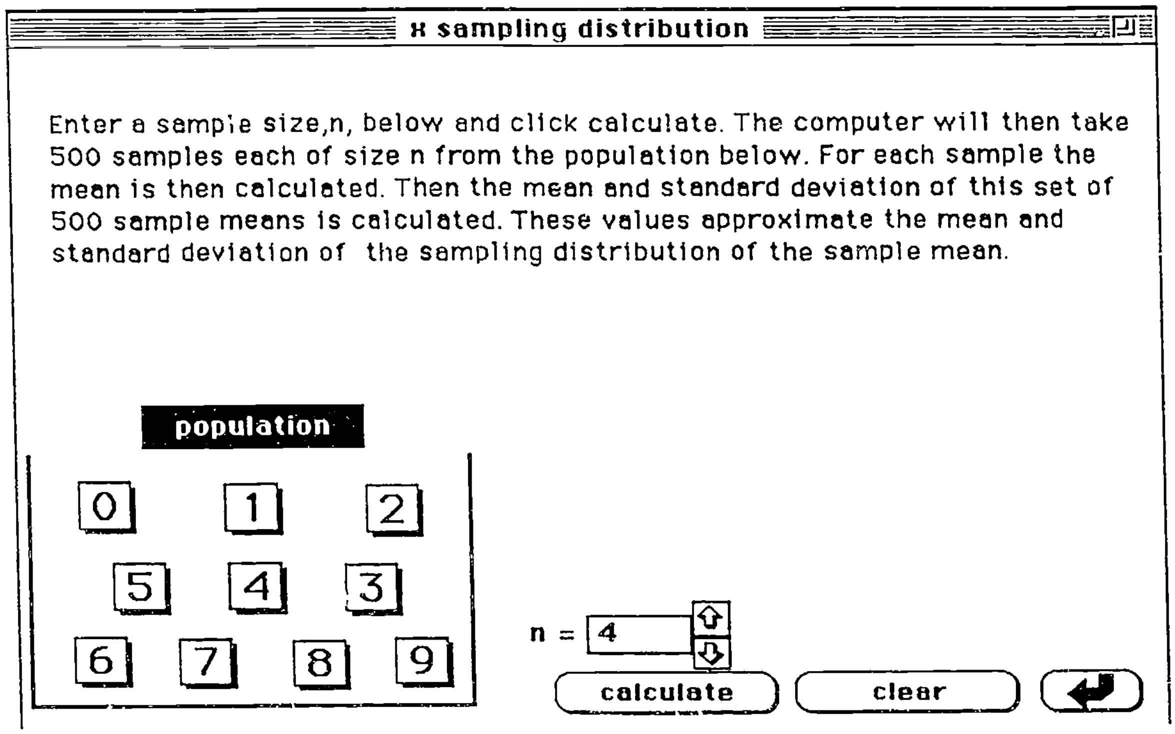
BEST COPY AVAILABLE

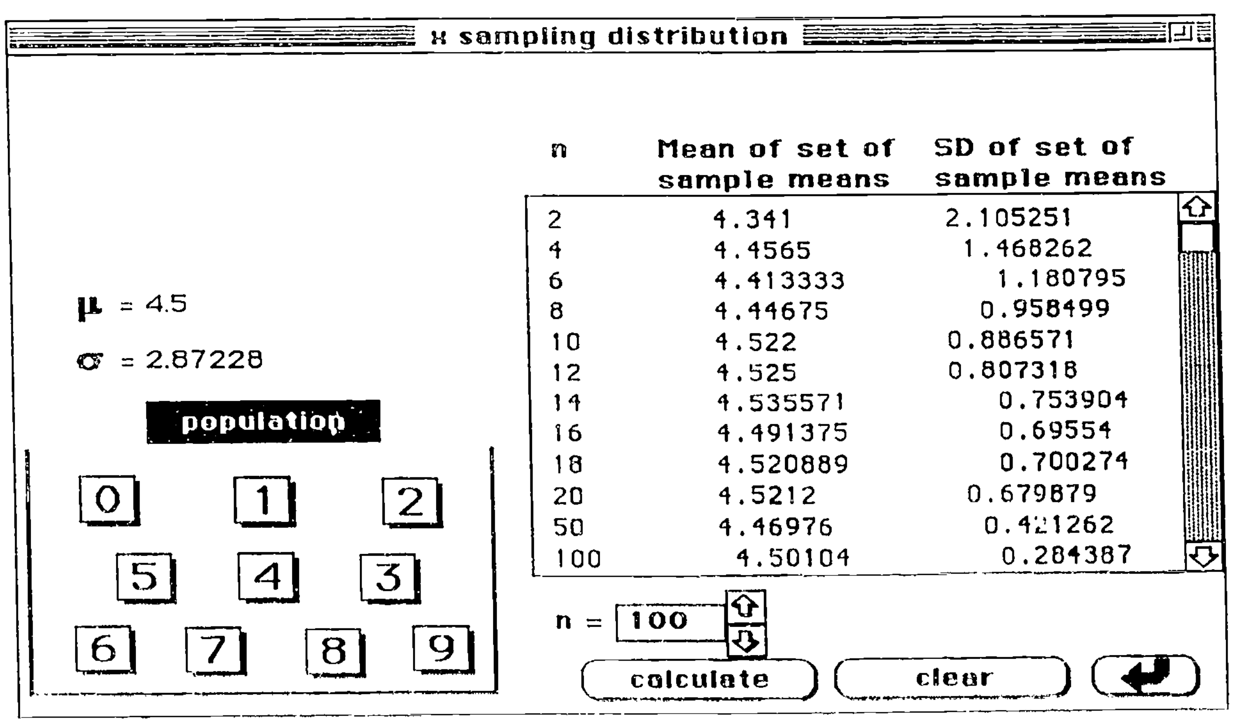
Example 2


x sampling distribution
sample
n = 20
take sample
sample count
0
slow
medium
fast
sample mean distribution
x̄ =
population
0 1 2
5 4 3
6 7 8 9
0
1
1
2
2 8
3 44204 104
3 6969767689856888868958668889995 97875
4 4 1220004422340022434000024 144243022300442
4 7685666658655676765988765786556776 6966
5 02 10020 120 10022040204044002 142200
5 7668658868678778
6
6
7
7
8
8
Menu
Stats
Clear


x sampling distribution
n = 20
sample count
0
slow
medium
fast
sample mean distribution
mean of xBars = 4.43525
SD of xBars = 0.637393
68 % of the xBars are within
1 SD of the mean
94 % of the xBars are within
2 SD of the mean
100 % of the xBars are within
3 SD of the mean
click to continue
0
1
1
2
2 8
3 44204 104
3 6969767689856888868958668889995 97875
4 4 1220004422340022434000024 144243022300442
4 7685666658655676765988765786556776 6966
5 02 10020 120 10022040204044002 142200
5 7668658868678778
6
6
7
7
8
8
graph

Example 3





n	Mean of set of sample means	SD of set of sample means
2	4.341	2.105251
4	4.4565	1.468262
6	4.413333	1.180795
8	4.44675	0.958499
10	4.522	0.886571
12	4.525	0.807318
14	4.535571	0.753904
16	4.491375	0.69554
18	4.520889	0.700274
20	4.5212	0.679879
50	4.46976	0.421262
100	4.50104	0.284387

Table 1

Sampling Distribution of the Mean

$\bar{X}$	Probability
3	1/10
4	1/10
5	2/10
6	2/10
7	2/10
8	1/10
9	1/10

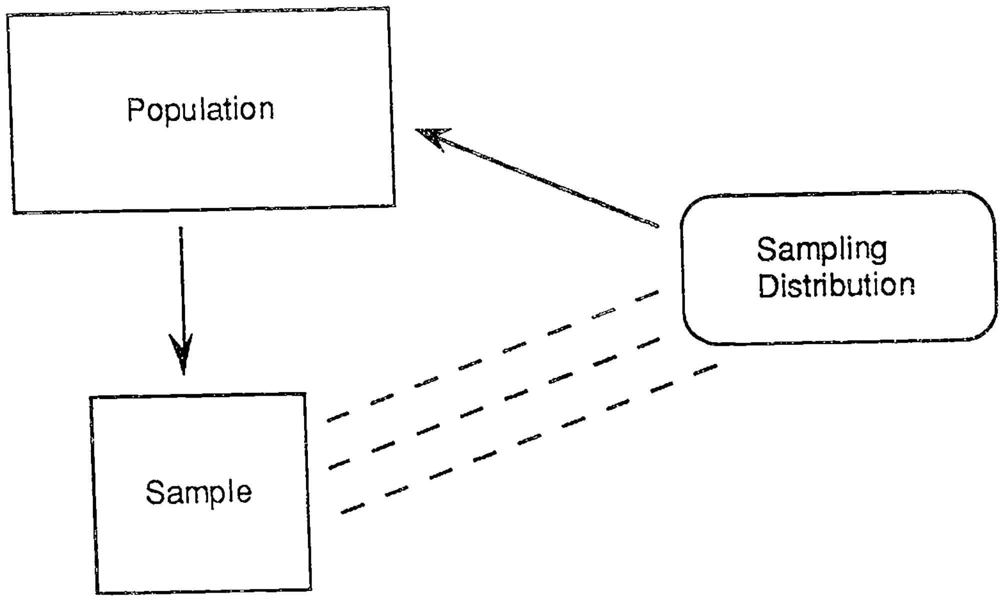
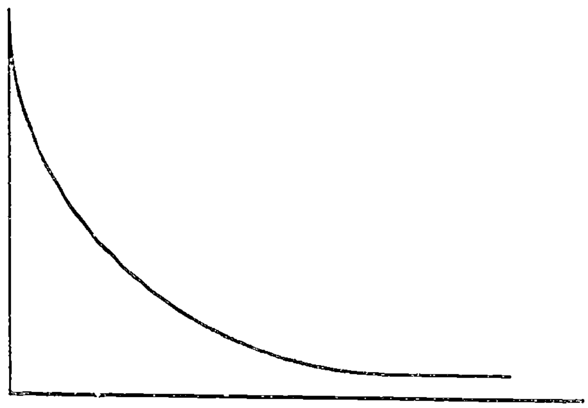
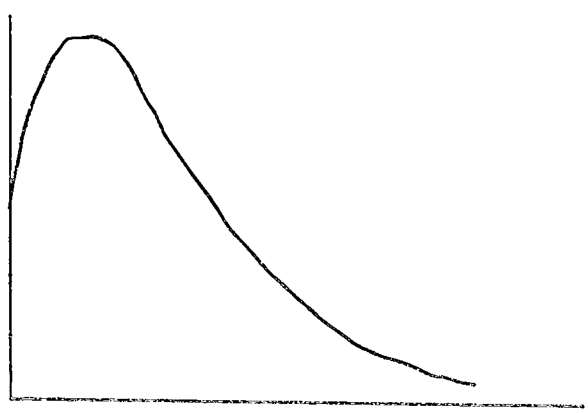


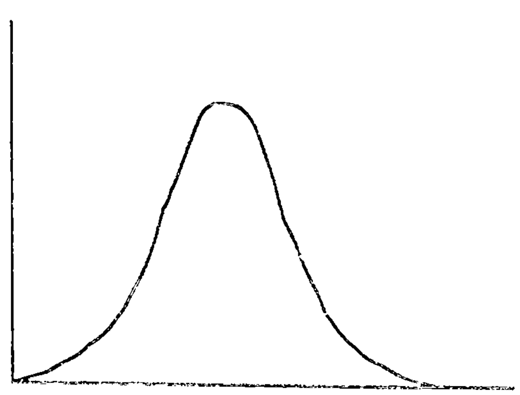
Figure 1. Chain of reasoning for statistical significance testing.



Parent Population



Sampling Distribution of the Mean for Sample Size $\underline{n} = 2$



Sampling Distribution of the Mean for Sample Size $\underline{n} = 30$

Figure 2. Sampling distributions of the mean for a skewed parent population.

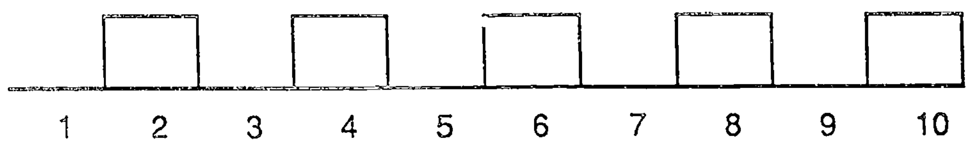


Figure 3. Histogram of the population.

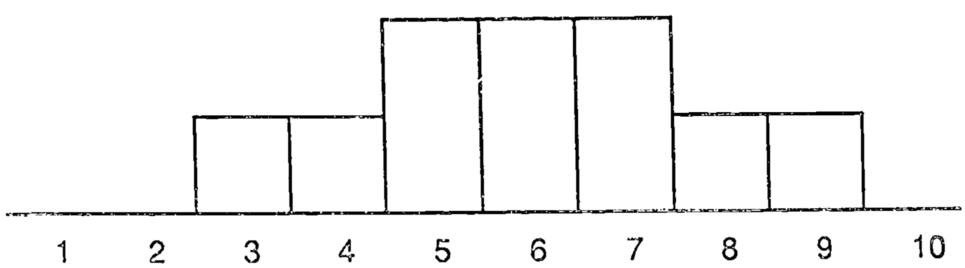


Figure 4. Histogram of the sampling distribution of the mean.