

Sampling Distributions, Standard Errors, and the Central Limit Theorem

Imagine that a study with sample size N was replicated many times (infinity technically). For each replication of the study, sample statistics could be calculated. The collection of this large number of sample statistics based on a sample size N is called a **SAMPLING DISTRIBUTION**. The Mean of a Sampling Distribution is the Expected Value of the computed sample statistic in the Population. That is, the Mean of a Sampling Distribution is the Population Parameter. For example, if the sample statistic computed for each replication is the sample mean (\bar{Y}), then the Mean of the Sampling Distribution is the Population Mean (μ_Y). If the sample statistic computed for each replication is the sample correlation coefficient (r), then the Mean of the Sampling Distribution is the Population Correlation (ρ). In practice, the Standard Deviation of the Sampling Distribution is known as the Standard Error (Se) of the statistic computed. Standard Errors involve Standard Deviations and Sample Size in their computation. Standard Errors are used to compute Confidence Intervals and to calculate inferential test statistics such as the t -test.

If the sample statistic computed in each replication is the Sample Mean (\bar{Y}), then there are some interesting properties due to the Central Limit Theorem. Imagine the distribution for rolling one die with 6 sides. Because each side is equally likely, the Parent Population is non-normal. It is a discrete Uniform distribution with a Population Mean of $\mu = 3.50$, a Population Variance of $\sigma^2 = 2.917$, and a Population Standard Deviation of $\sigma = 1.707$ (see Fig1.)

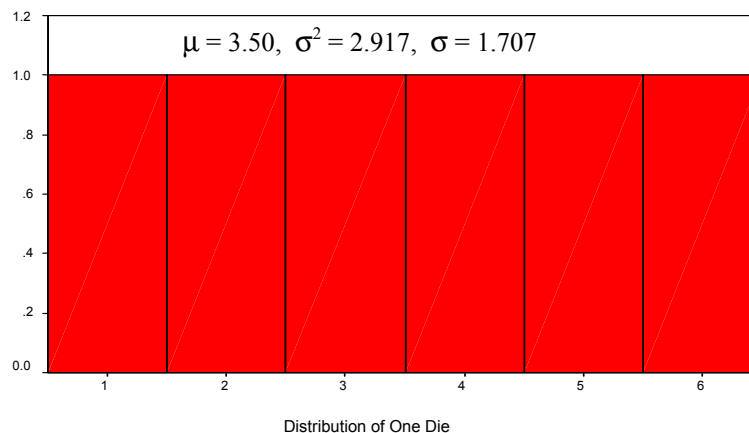


Figure 1.

Sampling Distribution of the Mean

Now imagine $N = 4$ people rolling one die each and then taking the Mean of those outcomes. Now consider if this process was replicated a large number of times. Occasionally, but very rarely, there would be a sample Mean roll of 6 (or 1), but most Mean rolls would be near the Expected Value of 3.5. Based on the Central Limit Theorem, the Sampling Distribution of the Mean for $N = 4$ people rolling one die would have a Mean of $\mu = 3.5$ and a Standard Deviation or

Standard Error of the Mean of $Se(\bar{Y}) = \sigma/\sqrt{4} = (1.71/2) = 0.854$. Note that the Standard Error of the Mean for 4 rolls ($Se(\bar{Y}) = 0.854$) is smaller than the Standard Deviation of the Parent Population ($\sigma = 1.707$). This is because for $N = 1$ person, rolling a 6 is not that rare, but for $N = 4$ people, obtaining a Mean of 6 is very rare.

Another interesting property of the Central Limit Theorem is that although the Parent Population (the roll of one die) is non-normal (i.e., Uniform, see Fig. 1), the Sampling Distribution of the Mean of $N = 4$ rolls is shaped somewhat like a Normal curve (see Fig. 2).

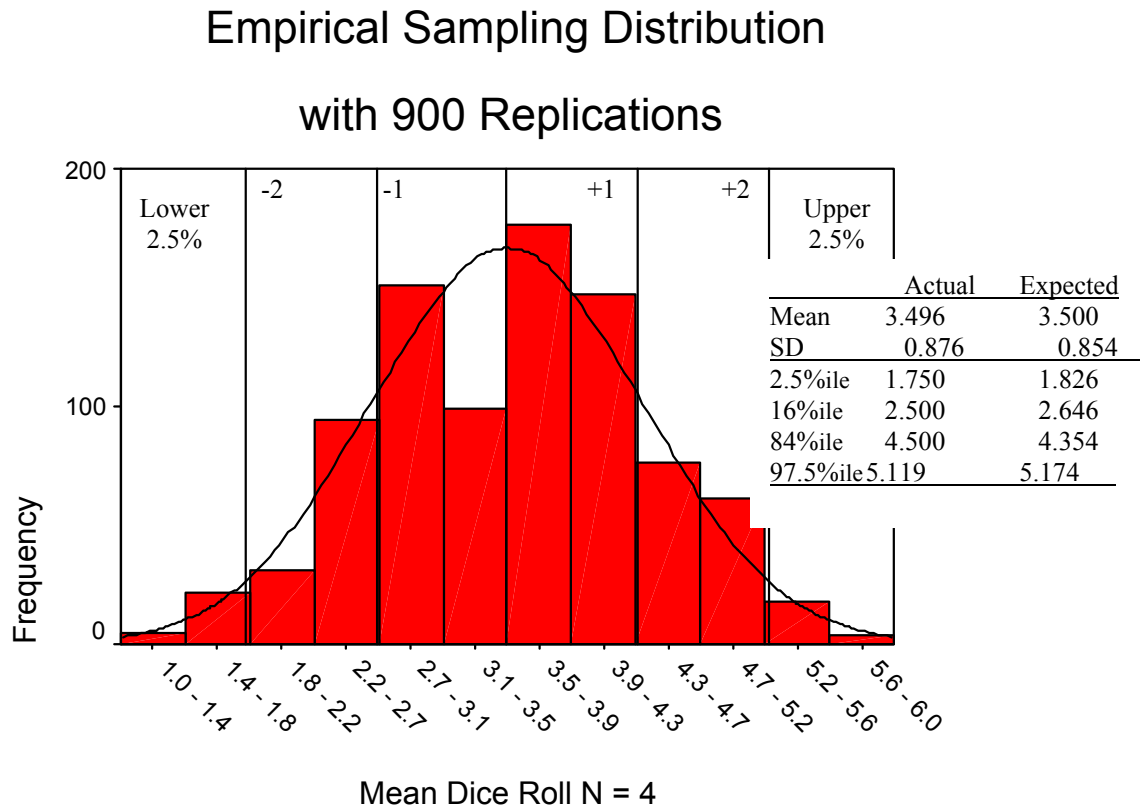


Figure 2.

Figure 2 and all Figures following show Empirical Sampling Distributions for various statistics with 900 replications. For example, for Figure 2, imagine that the $N = 4$ people rolled their dice 900 times and recorded the results. Of course, a simulation with 900 replications is NOT infinity, but it works reasonably well. These Figures also show the Actual Values for several points in the distributions as well as the Expected Values for these parameters based on statistical theory.

As the sample size N increases, the shape of the Sampling Distribution of the Mean becomes more normal. For example, with $N = 9$ people rolling one die each, the Sampling Distribution has a Mean of Mean of $\mu = 3.5$ and a Standard Deviation (i.e., Standard Error of the Mean) of $Se(\bar{Y}) = \sigma/\sqrt{9} = (1.707/3) = 0.569$.

Empirical Sampling Distribution with 900 Replications

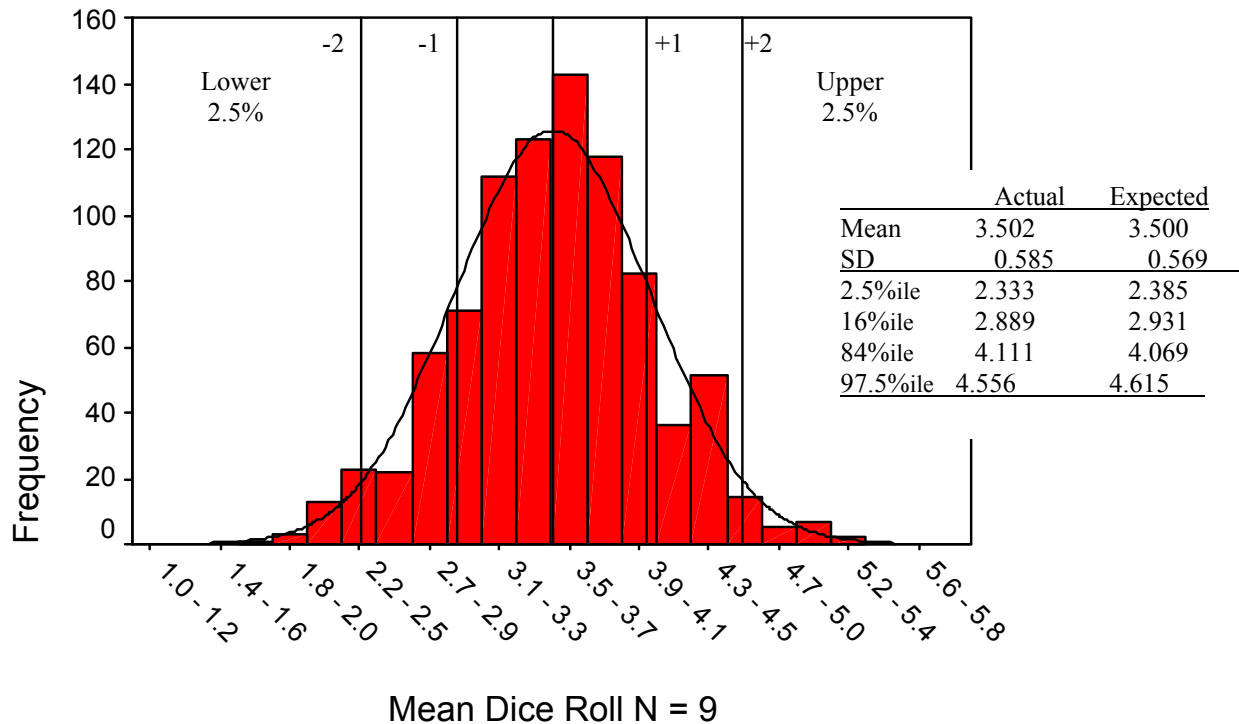


Figure 3.

Sampling Distribution of Mean Differences and Independent-Samples t -test

Now imagine that $N = 9$ people were randomly assigned to 2 groups of $n_B = 4$ and $n_G = 5$. The $n_B = 4$ people in Group B each roll one BLUE die. The $n_G = 5$ people in Group G each roll one GREEN die. One would expect that each group would have the same average roll. That is, one would assume a NULL HYPOTHESIS of NO DIFFERENCES in MEANS ($\mathbf{H}_0: (\mu_B - \mu_G) = \mu_D = 0$). Also, many statistics could be computed: (1) a Mean for each group (\bar{Y}_B and \bar{Y}_G), (2) a Mean Difference ($\bar{Y}_D = \{\bar{Y}_B - \bar{Y}_G\}$), or (3) an inferential statistic such as the Independent-Samples t -test. Although a rather large discrepancy in averages (i.e., Mean Differences) could occur, it is not likely given the Null Hypothesis ($\mathbf{H}_0: \mu_D = 0$). That is, for the statistics computed in this scenario, the Sampling Distributions are derived under a condition where the Null Hypothesis ($\mathbf{H}_0: \mu_D = 0$) is TRUE.

Empirical Sampling Distribution with 900 Replications

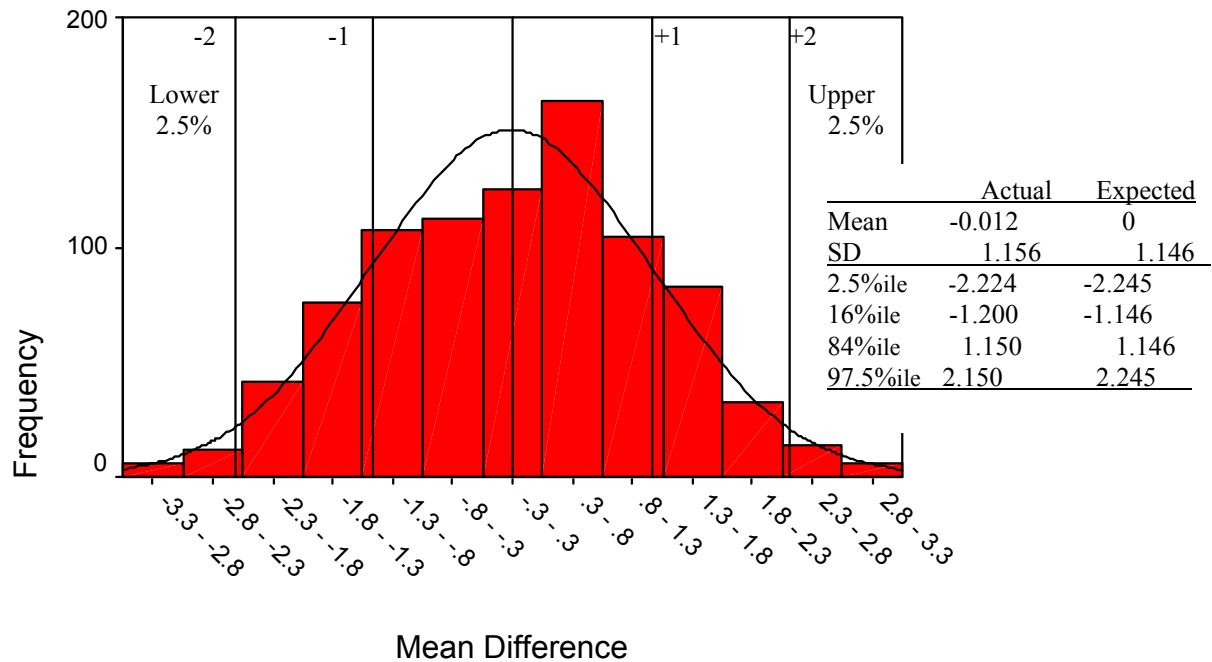


Figure 4.

Now imagine that these people roll their dice a large number of times. We could collect these statistics and have a Sampling Distribution for each of them. An important point to consider is that for a very large (infinite) number of replications, large Mean Differences WILL OCCUR BY CHANCE, but Mean Differences near the Expected Value of the Null Hypothesis ($H_0: \mu_D = 0$) will occur more frequently. The Mean of the Sampling Distribution of Mean Differences is 0 in this case. The Standard Deviation of this Sampling Distribution (i.e., Standard Error of Mean Differences) is expected to be $Se(\bar{Y}_D) = \sqrt{(\sigma^2/4) + (\sigma^2/5)} = \sqrt{(2.92/4) + (2.92/5)} = 1.146$. (see Fig. 4)

Figure 5 shows an Empirical Sampling Distribution for the t -test based on the dice rolling scenario under the condition that the null hypothesis of no mean differences in the population ($H_0: (\mu_B - \mu_G) = \mu_D = 0$) is TRUE. It should be noted that the Expected Values for the 2.5th and 97.5 percentiles were obtained by finding the two-tailed critical values for t with $(N-2) = 7$ degrees-of-freedom (df) for $\alpha = 0.05$.

Empirical Sampling Distribution with 900 Replications

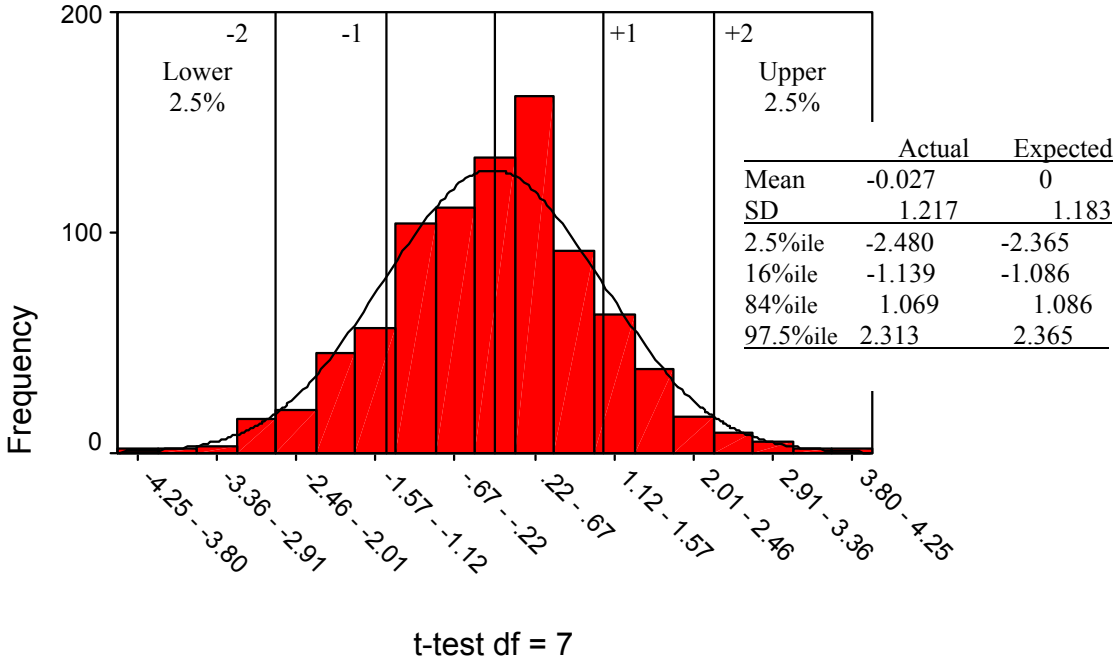


Figure 5.

Statistical Hypothesis Testing and Referent Distributions

Fortunately, the Sampling Distributions of most statistics do not have to be derived empirically. That is, a research does not have to repeatedly sample data sets of size N in order to know the Standard Error (i.e., Standard Deviation of the Sampling Distribution) for the statistic of interest. These values are derived mathematically from statistical theory.

The *t*-test is an inferential statistic used to evaluate the likelihood of the data (results in one sample) assuming the Null Hypothesis ($H_0: (\mu_B - \mu_G) = \mu_D = 0$ in the dice rolling example) is TRUE. This process of (1) analyzing data from one sample, (2) computing an inferential test statistic, and (3) making a decision concerning the tenability of the null hypothesis is called STATISTICAL HYPOTHESIS TESTING. Statistical Hypothesis Testing involves calculating a test statistic (e.g., *t*-test *F*-ratio, chi-square test of independence) from actual data and comparing this computed value to the referent distribution of that statistic (e.g., *t* distribution, *F* distribution, χ^2 distribution). Typically, the referent distribution for a statistical test is developed theoretically, based on an infinite sample of test statistics with a few properties of the actual data (i.e., numbers of predictors, groups, and subjects). As stated before, the collection of an infinite number of statistics is called a *sampling distribution*. Every statistic has a sampling distribution. When a sampling distribution is used for Statistical

Hypothesis Testing, it is called a *referent distribution*. The concept of infinite sampling is for mathematical convenience, and thus, theoretical sampling distributions are derived mathematically. The derivation usually comes from some form of the normal curve (another mathematically derived function); thus, no actual data are used to develop these distributions. Therefore, it is important to recognize that inferential statistics and hypothesis testing in their most common forms are based on *sampling distributions* of statistics not on the actual data at hand. Moreover, because the referent distribution is the sampling distribution of the test statistic, the Central Limit Theorem applies in many cases.

The Central Limit Theorem states that a sampling distribution of means (or sums) from random samples of N observations approximates a normal distribution regardless of the shape of the parent population as N becomes large. This has particular importance because it implies that even if data are sampled from a non-normal distribution, the sampling distribution of the mean (or sum) is normal as N becomes large (Hays, 1994). Thus, “normal theory” test statistics regarding means (or sums) can be applied if sample sizes are “sufficiently large.” Given that most research hypotheses in educational research concern means or mean differences (Olejnik, 1987; Thompson, 1994), normal theory (parametric) statistics can be applied in many research situations with relatively large samples. Although there is considerable debate over the use of normal theory statistics when data are non-normal (Bradley, 1968; Cliff, 1996), Glass and Hopkins (1996) review this perspective and provide a compelling argument for the use of parametric statistics.

In practice, a researcher computes a test statistic (i.e., t -test, F -statistic, chi-square test) from his data and compares it to the appropriate referent distribution. If the computed test statistic exceeds the $100(1 - \alpha)$ percentile of the referent distribution, then results are said to be “statistically significant at the α level.” A similar approach that is more common today is to find the exact percentile rank of the computed test statistic in its referent distribution and report $p = (1 - \text{percentile rank})$ as the probability of a Type I error (i.e., a p -value). If $p \leq \alpha$ then the results are statistically significant. The most common level of statistical significance is $\alpha = 0.05$. Therefore, directional, one-tailed hypotheses, test statistics are most often compared to the 95th percentile of their referent distribution. For two-tailed tests and symmetric Confidence Intervals, the $100(1-\alpha/2)$ percentile is used as the critical value. Thus, for $\alpha = 0.05$, the computed test statistic is compared to the 97.5th percentile of the referent distribution. For a 95% Confidence Interval, the 97.5th percentile is multiplied by a calculated Standard Error to derived the interval’s width. Given the growing speed of statistical software that “spit-out” one-tailed p -values, two-tailed p -values, and/or 95% Confidence Intervals within seconds, the concepts underlying referent distributions have been de-emphasized in statistics education. Knowledge of referent distributions helps one better understand the test statistics themselves because whenever p -values are reported, referent distributions are involved.

References

- Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice-Hall.
- Cliff, N. (1996). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research*, 31, 331-350.
- Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn & Bacon.
- Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace.
- Olejnik, S. (April, 1987). Teacher education effects: Looking beyond the means. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.
- Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, 54, 837-847