The Bimodality Principle

Erhard Reschenhofer
University of Vienna

*Journal of Statistics Education* Volume 9, Number 1 (2001)

Copyright © 2001 by Erhard Reschenhofer, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the author and advance notification of the editor.

**Key Words:** Model selection; Selecting the level of significance; Testing.

**Abstract**

In statistics courses, students often find it difficult to understand the concept of a statistical test. An aggravating aspect of this problem is the seeming arbitrariness in the selection of the level of significance. In most hypothesis-testing exercises with a fixed level of significance, the students are just asked to choose the 5% level, and no explanation for this particular choice is given. This article tries to make this arbitrary choice more appealing by providing a nice geometric interpretation of approximate 5% hypothesis tests for means.

Usually, we want to know not only whether an observed deviation from the null hypothesis is statistically significant, but also whether it is of practical relevance. We can use the same geometrical approach that we use to illustrate hypothesis tests to distinguish qualitatively between small and large deviations.

**1. Introduction**

The histograms of many datasets occurring in practice have the appearance of a bell. They are symmetric about their means and tail off rapidly as we move away from the means. A typical example is shown in Figure 1a, which summarizes the mean temperatures in May recorded from 1845 to 1978 in St. Louis. (This dataset will be described in more detail in Section 3.) Of course, histograms can also look quite different from the distribution in Figure 1a. They may be skewed, have thick tails, or exhibit more than one peak. In this paper, we are interested in the last case, particularly that of two peaks. Histograms with two peaks are called bimodal, and those with only one peak are called unimodal. An example of a bimodal histogram is shown in Figure 2b, which summarizes a dataset containing mean temperatures observed in July and in September. Here bimodality is due to the fact that the dataset is heterogeneous. It could easily be dissected into two more homogeneous parts by studying the July temperatures and the September temperatures separately.
Clearly, bimodality does not always occur when we have a mixture of two sets of observations with different means -- the means must be sufficiently different. Consider, for example, Figure 2a, which summarizes a dataset containing mean temperatures observed in May and September. In this case, the difference between the mean of the May temperatures and that of the September temperatures is too small to cause bimodality. If we want to assess the difference between the means of two datasets, we could examine the shape of the histogram of the combined dataset. Bimodality of this histogram could serve as a qualitative indicator for a big difference between the means. This idea will be explained in more detail in Section 3.
Figure 2a. Histogram of the Mean St. Louis Temperature in May and September (1845-1978).

Figure 2b. Histogram of the Mean St. Louis Temperature in July and September (1845-1978).
A different question, namely whether or not an observed difference between two means is statistically significant, is discussed in Section 4. To answer this question, we must examine the distributions of the sample means rather than the distributions of individual observations. Again we might check for bimodality. But this time we must examine the combination of the distributions of the sample means. It turns out that bimodality occurs whenever the null hypothesis of identical means is rejected by a hypothesis test at an approximate 5% level of significance. Hence this approach provides a nice geometric interpretation of tests for differences between means at the 5% level.

The question of how the 5% level of significance was chosen as a standard is examined in Section 2. Finally, Section 5 discusses the usefulness of fixed-level significance testing versus the mere reporting of $p$-values, describes class reaction to the bimodality principle, and gives suggestions for covering this material with students.
2. The Origins of the 5% Level of Significance

A crucial problem in statistics is to discriminate between two or more competing hypotheses or models. The first problem of this kind faced by a beginner is that of testing the null hypothesis that the mean of a normal distribution is equal to a specified value \( c \). When the sample size is large, it is usually suggested that we reject this null hypothesis whenever the distance between \( c \) and the sample mean \( \bar{x} \) exceeds two standard deviations of the sample mean. A similar problem is that of testing the null hypothesis that the means of two normal distributions are identical. The latter null hypothesis is usually rejected whenever the distance between the two sample means, \( \bar{x} \) and \( \bar{y} \), exceeds two standard deviations of \( \bar{x} - \bar{y} \). In each of the two cases, the stated rejection rule guarantees that the probability of rejecting a true null hypothesis is only 5% (approximately).

But how can the choice of the 5% level be justified? Cowles and Davis (1982) investigated the question of how the 5% level of significance was chosen as a standard. Examining early literature in probability and statistics, they found that Fisher (1925) was perhaps the first to formally mention the 5% level. In his book Statistical Methods for Research Workers, Fisher stated that deviations exceeding twice the standard deviation are regarded as significant. However, Cowles and Davis (1982) stressed that Fisher should not be credited with introducing the 5% level because the choice of this level by Fisher was not casual and arbitrary, but was influenced by previous scientific conventions. At the beginning of the 20th century, statements about statistical significance were still given in terms of the probable error, which was the nineteenth-century measure of the width of a distribution (see Porter 1986 and Stigler 1986). (The German astronomer Frederik Wilhelm Bessel appears to have coined the term ‘probable error’ or ‘der wahrscheinliche Fehler’ by 1815 (see Walker 1929, p. 186). The term ‘standard deviation’ was introduced almost 80 years later by Karl Pearson (see Stigler 1986, p. 328).) Deviations exceeding three times the probable error were considered significant (see, e.g., Student 1908). The probable error is defined as the median deviation from the mean. If the mean coincides with the median, which is the case for symmetric distributions, the probable error is just half the interquartile range. Observing that the upper quartile of a standard normal distribution lies between 0.66 and 0.67, we note that the probable error roughly corresponds to 2/3 of a standard deviation. In the normal case, a deviation of three probable errors therefore corresponds to a deviation of two standard deviations. Hence it seems that the 5% level has a longer history than is generally appreciated.

3. Using the Bimodality Principle to Distinguish Between Small and Large Location Differences

We use a simple meteorological example to introduce the bimodality principle. The variable of interest is the monthly mean temperature in St. Louis, Missouri. Data are available for the period from January 1845 to December 1978 (see Marple 1987). In view of the extreme unreliability of long-term weather forecasts, these measurements may be considered as roughly independent observations. This dataset is considered as a sample of 134 years from the population of all years. For each month we have \( n = 134 \) observations of mean temperatures. Suppose we want to compare the mean St. Louis temperature in May, \( M_i \), with that in September, \( S_i \). We will use the notation \( M_i \) to indicate the measurement of the variable \( M \) in the \( i \)th year. Analogously, \( S_i \) denotes the measurement of the variable \( S \) in the \( i \)th year. We start our analysis of the datasets \( M_1, \ldots, M_n \) and \( S_1, \ldots, S_n \) with a visual inspection of their histograms. (Minitab 12 was used to generate the histograms.) Figure 1a shows the histogram for the first dataset. The horizontal axis is divided into classes of width 1.5 degrees Celsius. The endpoints of the first class are 14.5 and 16 degrees Celsius, those of the second interval are 16 and 17.5 degrees Celsius, and so forth. (The mean temperatures are given to the nearest tenth (one decimal place), so that the first class actually covers the interval from 14.5 to 15.9, the second class covers the interval from 16.0 to 17.4, and so forth.) The vertical axis gives the proportion of the measurements that fall in each class. We can see that the mean temperatures in May are symmetrically distributed with a peak slightly below 20 degrees Celsius. The endpoints of the modal class are 17.5 and 19 degrees Celsius. The sample mean is 18.8 degrees Celsius, and the median is 18.9 degrees Celsius. The histogram for the second dataset is shown in Figure 1b. The mean temperatures in September appear to be symmetrically distributed with a peak slightly above 20 degrees Celsius. The endpoints of the modal class are 20.5 and 22 degrees Celsius. Both the sample mean and the median are 21.1 degrees Celsius. On average, the mean temperatures in September are slightly higher than in May.

Clearly, it depends on the circumstances whether or not this difference is considered as important. For an average citizen of St. Louis it may be insignificant, whereas for the operator of a solar power station it may be very important. A purely formal approach for assessing the size of this difference is to combine both samples into a single sample and then produce a histogram for the combined sample. If the distance between the means is large enough, this histogram will exhibit two peaks, each of which corresponds to a peak in one of the two original histograms. In our case, the difference is too small. The histogram for the combined dataset has only one peak (see Figure 2a); hence it does not indicate an important
difference. In contrast, if we compare the mean temperatures in July, $J_1, \ldots, J_n$, with those in September, we find two peaks in the histogram of the combined dataset (see Figure 2b). The bimodality of the histogram of the combined sample may be considered an indication of an important location difference between the two datasets. Indeed, the first peak is close to the mean of the September measurements (\( \bar{S} = 21.1 \)) and the second peak is close to the mean of the July measurements (\( \bar{J} = 26.3 \)).

The above procedure for distinguishing between important and unimportant location differences is not completely objective because it contains a subjective component, namely the choice of the classes used for the construction of the histograms. Unfortunately, this choice strongly influences the appearance of the histogram. Choosing the width of the class intervals too small could give rise to spurious peaks (see Figure 2c). On the other hand, genuine bimodality could be concealed by choosing the width of the class intervals too large (see Figure 2d). An obvious way to get rid of this subjective component is to use another graphical tool for the description of the data instead of the histogram. The probability distribution of a continuous random variable like the air temperature is characterized by its probability density function. The probability that the random variable takes on a value in the interval from \( a \) to \( b \) is just the area under the graph of the probability density function between \( a \) and \( b \). A histogram can be regarded as an estimate of the probability density function. Many continuous random variables occurring in practice have bell-shaped probability density functions. Figures 1a and 1b suggest that this might be true also for our random variables \( M \) and \( S \). For both datasets, neither the Kolmogorov-Smirnov test nor the Anderson-Darling test detects any deviation from normality at the 10% level of significance. We may therefore assume that their probability density functions are of the normal type. Normal probability density functions are completely determined by two parameters, the mean and the standard deviation. Clearly, we do not know the means and the standard deviations of the random variables \( M \) and \( S \), but we can use estimates instead. Estimates of the means and the standard deviations of \( M \) and \( S \) are obtained by calculating the sample means,

\[
\bar{M} = \frac{1}{n} \sum M_i \quad \text{and} \quad \bar{S} = \frac{1}{n} \sum S_i ,
\]

and the sample standard deviations,

\[
s_M = \sqrt{\frac{1}{n-1} \sum (M_i - \bar{M})^2} \quad \text{and} \quad s_S = \sqrt{\frac{1}{n-1} \sum (S_i - \bar{S})^2} .
\]

(Being mean temperatures, the \( M_i \) and the \( S_i \) are averages themselves.) The sample means differ slightly (\( \bar{M} = 18.835 \), \( \bar{S} = 21.093 \)), but the agreement between the sample standard deviations is striking (\( s_M = 1.782 \), \( s_S = 1.782 \)). (Minitab 12 was used for all calculations.) Plugging these estimates into the equation of the normal probability density function

\[
f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( -\frac{(x - \mu)^2}{2\sigma^2} \right) ,
\]

we obtain estimates of the probability density functions of \( M \) and \( S \), namely \( f(x \mid \bar{M}, s_M) \) and \( f(x \mid \bar{S}, s_S) \). (Note that we use \( \mu \) and \( \sigma \) as the parameters of \( f \) rather than \( \mu \) and \( \sigma^2 \).) The graphs of these functions are shown in Figures 3a and 3b, respectively. (DERIVE 4 was used to generate the plots of the probability density functions.) The first graph summarizes the dataset \( M_1, \ldots, M_n \), and the second one summarizes the dataset \( S_1, \ldots, S_n \).
Figure 3a. Estimated Probability Density Function for the Mean Temperature in May.
What we need next is a graphical summary of the combined dataset $M_1, ..., M_n, S_1, ..., S_n$. Unfortunately, since all normal probability density functions are unimodal, they are of no use for the description of the combined dataset. Instead, we try to construct a summary for the combined dataset by combining the summaries of the two original datasets. To see how this can be done, we again consider histograms. The histogram for the combined dataset can be constructed either directly from the data $M_1, ..., M_n, S_1, ..., S_n$ or indirectly from the histograms of the original datasets. In the latter case, we must use the same classes for all histograms. For example, consider the class with endpoints 16 and 17.5 degrees Celsius. Twenty-four (17.9%) of the 134 measurements $M_1, ..., M_n$, 3 (2.2%) of the 134 measurements $S_1, ..., S_n$, and 27 (10.1%) of all 268 measurements $M_1, ..., M_n, S_1, ..., S_n$ fall in this class. The proportion of all measurements falling in this class is just the average of the other two proportions. This is a consequence of the fact that the original samples are of the same size. Each measurement $M_i$ or $S_i$ represents $(1/134) \times 100\%$ of the measurements in one of the original samples and exactly one half of this percentage, namely $(1/268) \times 100\%$, in the combined sample. Because probability density functions can be regarded as approximations of histograms with very small class intervals, it is natural to combine the probability density functions describing the original samples in the same way as we combined the histograms, i.e., by taking averages.

Combining the two probability density functions depicted in Figures 3a and 3b we obtain the function

$$\frac{1}{2}(f(x \mid \bar{M}, s_M) + f(x \mid \bar{S}, s_S)),$$

the graph of which is shown in Figure 4a. Analogously, Figure 4b shows the graph of

$$\frac{1}{2}(f(x \mid \bar{J}, s_J) + f(x \mid \bar{S}, s_S)).$$

Again we note that the difference between $\bar{J}$ and $\bar{S}$ is large enough so that the average of $f(x \mid \bar{J}, s_J)$ and $f(x \mid \bar{S}, s_S)$ has two peaks, whereas the difference between $\bar{M}$ and $\bar{S}$ is too small. The presence of two peaks (bimodality) in the combined probability density (mixture density) may therefore serve as an indicator for a large difference in the means. But how far apart must the means of two normal distributions be for the mixture density to be bimodal? We try to answer this question for the important special case where the standard deviations are identical. (We do not consider the interesting, but tricky, case of unequal standard deviations in this tutorial note. Bimodality criteria for this case are given in Robertson and Fryer 1969.) Consider Figure 5, which shows several mixture densities. It suggests that there are two peaks in the mixture density only if the distance between the means is greater than two standard deviations. The criterion of two standard deviations is established in a formal way using derivatives. But since calculus is not a prerequisite for all undergraduate statistics courses, the proof is given in the Appendix only. However, students who have completed their first calculus course are encouraged to work through this proof. It relies only on the combination of two probability density functions, taking derivatives, and checking concavity.
Figure 4a. Combination of the Estimated Probability Density Functions for May and September.
4. Using the Bimodality Principle To Illustrate Hypothesis Tests

In statistics, the difference between two sample means is usually assessed in two ways. First, the size of the difference is judged by its practical importance. In most applications, this can easily be accomplished without sophisticated decision rules. Only if the investigator has absolutely no clue which differences should be considered important, he/she might have recourse to a formal rule like the one based on a bimodality check. According to this rule, called the bimodality principle, a location difference between two (estimated) normal probability density functions is regarded as large (or important) if their mixture density is bimodal. In the previous section, we have applied this principle to distinguish between small and large location differences.

The second interesting question regarding the difference between two sample means is whether it is large enough to indicate that the population means also differ. In our example, we could wish to determine whether the overall mean temperature in May differs significantly from that in September. This question may be answered by applying a 5% level hypothesis test. In the second part of this section, we will show how the bimodality principle can be used to illustrate this test. But first we consider the one-sample case.

Suppose we are given a sample $x_1, \ldots, x_n$ from a normal distribution with mean $\mu$ and standard deviation $\sigma$. We formulate a simple null hypothesis, $H_0$, and an appropriate alternative hypothesis, $H_A$:

$$H_0: \mu = c, \quad H_A: \mu \neq c$$

The null hypothesis states that the mean $\mu$ is equal to a specified value $c$, and the alternative hypothesis states that the mean
differs from this value. It is natural to test the null hypothesis by calculating the sample mean and rejecting the null hypothesis whenever the discrepancy between $\bar{x}$ and $c$ is too large. The significance of any discrepancy depends on the reliability of the sample mean. To assess the reliability of a sample mean, we may consider its sampling distribution. The sampling distribution of $\bar{x}$ is normal with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$. An estimate is given by $f(x \mid \bar{x}, s_x/\sqrt{n})$. Under the null hypothesis, $\bar{x}$ should be close to $c$, hence the two probability density functions $f(x \mid \bar{x}, s_x/\sqrt{n})$ and $f(x \mid c, s_x/\sqrt{n})$ should not differ too much (see Endnote). The null hypothesis could be rejected if their mixture density is bimodal. Recalling from Section 2 that the mixture density of two normal probability density functions is bimodal if the difference between the means exceeds two standard deviations, we note that in this case the bimodality principle rejects the null hypothesis if $|\bar{x} - c| > 2s_x/\sqrt{n}$. Thus the bimodality principle makes the same decision as the standard large sample significance test at the 5% level. (Actually a large sample $t$-test rejects the null hypothesis if $|\bar{x} - c| > 1.96s_x/\sqrt{n}$.)

We now return to our meteorological hypothesis testing problem which involves two samples, $M_1,...,M_n$ and $S_1,...,S_n$. Assuming that the standard deviations are identical, i.e., $\sigma_M = \sigma_S$, we want to test the null hypothesis that the means are also identical, i.e., $H_0: \mu_M = \mu_S$. But we will not simply examine the combination of the two estimated sampling distributions obtained from the samples $M_1,...,M_n$ and $S_1,...,S_n$. Instead, we will use the more orthodox approach of examining the combination of two estimates obtained from the combined sample under the null hypothesis and under the alternative hypothesis, respectively. Under the null hypothesis, all $2n$ measurements come from a normal distribution with mean $\mu_M = \mu_S$ and standard deviation $\sigma_M = \sigma_S$. Hence, the sample mean of the combined sample, $(M + S)/2$, is normally distributed with mean $\mu_M = \mu_S$ and standard deviation $\sigma_M/\sqrt{2n} = \sigma_S/\sqrt{2n}$. The mean and the standard deviation of this normal distribution can be estimated by $(M + S)/2$ and $\sqrt{\frac{1}{2}(s^2_M + s^2_S)}/(2n)$, respectively. Note that the estimate of the standard deviation is meaningful both under the null hypothesis and under the alternative hypothesis because it is based on the deviations from $M$ and $S$, respectively, rather than from $(M + S)/2$. Thus we can also choose the same standard deviation for the second normal probability density function. Next we have to specify the mean of the second probability density function. The quantity of interest is the distance between the mean of the first distribution, $(M + S)/2$, and the mean of the second distribution. Under the alternative hypothesis, the measurements in the combined sample do not come from a single normal distribution, but from two normal distributions with different means. However, the distance between $(M + S)/2$ and $M$ equals that between $(M + S)/2$ and $S$. Both are given by $|M - S|/2$. Thus we may choose either $\bar{M}$ or $\bar{S}$ as the mean of the second normal distribution. In any case, the difference between the means of the two probability density functions will be $|\bar{M} - \bar{S}|/2$. The null hypothesis will be rejected if the mixture density is bimodal or, equivalently, if

$$\frac{|\bar{M} - \bar{S}|}{2} > 2\sqrt{\frac{1}{2}(s^2_M + s^2_S)}/(2n).$$

Rewriting this inequality as

$$|\bar{M} - \bar{S}| > 2\sqrt{s^2_M/n + s^2_S/n},$$

we notice immediately that our two-sample test based on a bimodality check agrees with the standard large sample test for comparing two means if the 5% level of significance is chosen for the latter test.

In our example,
and hence the hypothesis of identical means is rejected at the 5% level of significance. Correspondingly, the combination of the probability density functions \( f\left(\bar{M}, \sqrt{\frac{1}{2} \left(\sigma_M^2 + \sigma_S^2\right)}/(2n)\right) \) and \( f\left(\frac{1}{2} \left(\bar{M} + \bar{S}\right), \sqrt{\frac{1}{2} \left(\sigma_M^2 + \sigma_S^2\right)}/(2n)\right) \) exhibits two peaks (see Figure 6).

Figure 6. Combination of the Probability Density Functions \( f\left(\bar{M}, \sqrt{\frac{1}{2} \left(\sigma_M^2 + \sigma_S^2\right)}/(2n)\right) \) and \( f\left(\frac{1}{2} \left(\bar{M} + \bar{S}\right), \sqrt{\frac{1}{2} \left(\sigma_M^2 + \sigma_S^2\right)}/(2n)\right) \).

5. Discussion

Today's statistical software calculates \( p \)-values automatically; hence the practice of fixed-level significance testing is no longer dictated by the availability of tables. Of course, stating whether a hypothesis is rejected or not at some level of significance is not as informative as giving the \( p \)-value itself. Reporting the actual \( p \)-value indeed makes it much easier for the reader of a report to judge the significance of a result. Nevertheless, there are still situations, e.g., in economic forecasting, where statisticians must decide for or against some hypothesis before they can carry on with their work. Ideally, if a statistician is going to make such a decision, he/she should take the consequences of his/her decision into account in choosing the level of significance. Unfortunately, this often cannot be accomplished in an objective and verifiable way. At best, it will only be possible to decide whether the 10% level is more appropriate than the 1% level, but certainly not whether the 4% level is more appropriate than the 6% level. Hence it still makes sense to have standards like the 1% level, the 5% level, or the 10% level. The mere existence of such standards already makes cheating more difficult. Clearly, if
someone reports that he/she has rejected a hypothesis at the 6% level, the reader of the report will check suspiciously whether there are good reasons for using just this level of significance.

I have used the bimodality principle to illustrate 5%-level hypothesis tests in introductory statistics courses for science, education, and engineering students. However, I did not explain all the details and omitted the proof. I just showed the figures and used approximately half an hour to explain them. Student reaction was mixed. Only a few students, particularly those who frequently asked questions, explicitly appreciated the explanation. The majority never did question the use of the 5% level and therefore felt no need for an illustration. In my explanation I focussed on the coincidence that, on the one hand, the critical value of a large sample $t$-test at the 5% level is approximately 2, and, on the other hand, the mixture density of two normal probability density functions is bimodal if the difference between their means exceeds two standard deviations. This material (including the proof and all details) could possibly be appropriate for an investigation involving extra effort outside of a typical class. This may be an honors project associated with a class or even a senior project.

**Acknowledgments**

I wish to thank the Editor, the Associate Editor, and the Referees for helpful comments. This paper was written at the Sultan Qaboos University, Oman.

**Endnote**

Note that the sample standard deviation is a reasonable estimate of $\sigma$ under both $H_0$ and $H_A$. Under $H_0$, we could use a different estimate of $\sigma$ obtained by replacing $\overline{x}$ by $c$ in the definition of the sample standard deviation. However, this would have the unpleasant consequence that the two normal probability density functions would have unequal standard deviations. In case of a large discrepancy between the standard deviations, bimodality could occur even if there were no significant discrepancy between the means. In addition, the case of unequal standard deviations is technically demanding. Finally, it is quite common in situations where we must distinguish between different hypotheses (or models) to use the estimate of the nuisance parameter obtained under the weakest hypothesis (or with the largest model) also for the assessment of the stronger hypotheses (smaller models).

**Appendix**

**Theorem:** The mixture density

$$f_m(x | \mu_1, \mu_2, \sigma) = \frac{1}{2} \left( f(x | \mu_1, \sigma) + f(x | \mu_2, \sigma) \right)$$

of two normal probability density functions with the same standard deviation, $\sigma$, but with different means, $\mu_1$ and $\mu_2$, respectively, is bimodal if and only if $|\mu_2 - \mu_1| > 2\sigma$.

**Proof:** Depending on the distance between $\mu_1$ and $\mu_2$, the mixture density

$$f_m(x | \mu_1, \mu_2, \sigma) = \frac{1}{2} \left[ \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(x - \mu_1)^2}{2\sigma^2} \right) + \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(x - \mu_2)^2}{2\sigma^2} \right) \right]$$

will have either a maximum at $x_0 = \frac{1}{2} (\mu_1 + \mu_2)$ (the unimodal case) or a local minimum at $x_0 = \frac{1}{2} (\mu_1 + \mu_2)$ (the bimodal case). Indeed, $x_0$ is a stationary point because

$$f_m'(x_0) =$$
Now we must check the second derivative to see whether a maximum or a minimum occurs.

\[
f_m''(x_0) = \frac{1}{2\sigma^2 \sqrt{2\pi \sigma^2}} \left[ - \exp \left( \frac{(x_0 - \mu_1)^2}{2\sigma^2} \right) + \left( x_0 - \mu_1 \right)^2 \exp \left( \frac{-(x_0 - \mu_1)^2}{2\sigma^2} \right) \right.
\]
\[
- \exp \left( \frac{(x_0 - \mu_2)^2}{2\sigma^2} \right) + \left( x_0 - \mu_2 \right)^2 \exp \left( \frac{-(x_0 - \mu_2)^2}{2\sigma^2} \right)
\]
\[
+ \frac{1}{\sigma^2} \exp \left( \frac{- (\mu_2 - \mu_1)^2}{2\sigma^2} \right) \left[ -1 + \frac{(\mu_2 - \mu_1)^2}{\sigma^2} \right]
\]
\]
\[
= 0,
\]

if \((\mu_2 - \mu_1)^2 > 4\sigma^2\) or, equivalently, if \(|\mu_2 - \mu_1| > 2\sigma\). Thus, a minimum occurs only if the distance between the two means exceeds two standard deviations.

References

Cowles, M., and Davis, C. (1982), "On the Origins of the .05 Level of Significance," *American Psychologist*, 5, 553-558.

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd.

Marple, S. L., Jr. (1987), *Digital Spectral Analysis*, Englewood Cliffs: Prentice Hall.

Porter, T. M. (1986), *The Rise of Statistical Thinking 1820-1900*, Princeton, NJ: Princeton University Press.

Robertson, C. A., and Fryer, J. G. (1969), "Some Descriptive Properties of Normal Mixtures," *Skandinavisk Aktuarietidskrift*, 69, 137-146.

Stigler, S. M. (1986), *The History of Statistics*, Cambridge, MA: The Belknap Press of Harvard University Press.

Student (W. S. Gosset) (1908), "The Probable Error of a Mean," *Biometrika*, 6, 1-25.
Walker, H. M. (1929), *Studies in the History of Statistical Method*, Baltimore: Williams & Wilkins. Reprinted 1975, New York: Arno Press.