12-2001

On the comparison to EEG Norms : a new method and a simulation study

Marco Congedo

To the Graduate Council:

I am submitting herewith a thesis written by Marco Congedo entitled "On the comparison to EEG Norms : a new method and a simulation study." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Arts, with a major in Psychology.

Joel Lubar, Major Professor

We have read this thesis and recommend its acceptance:

S. Handel, M. Hash

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a thesis written by Marco Congedo entitled "On the Comparison to EEG Norms: a new Method and a Simulation Study". I have examined the final paper of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Arts, with a major in Psychology.

Joel F. Lubar

Joel Lubar, Major Professor

We have read this thesis
and recommended its acceptance:

Stephen Handel

S. Handel

M. Nash

Accepted for the Council:

Vice Provost and Dean of Graduate Studies

On the Comparison to EEG Norms: a new Method and a Simulation Study

A Thesis
Presented for the
Master of Arts
Degree
The University of Tennessee, Knoxville

Marco Congedo
December 2001

Ad Aaron,
ed alla sua pazienza.

Acknowledgments

The research presented in this thesis has been possible thanks to the constant supervision of Dr. Joel Lubar. His support and encouragement has been generous and constructive but never constraining, and has given the author the necessary motivation to pursue this project. The author also wish to express his gratefulness to Dr. Roberto Pascual-Marqui, whose work greatly influenced his studies, and whose comments gave him the idea underlying this thesis.

Leslie Sherlin patiently corrected the original manuscript. His comments were irreplaceable in smoothing the writing style of the author.

Abstract

Quantitative Electroencephalography (qEEG) as a tool for the diagnosis of neurological and psychiatric disorders is receiving an increased interest. While qEEG analysis is restricted to the scalp, the recent development of electromagnetic tomographies (ET) allows the study of the electrical activity of cortical structures. Electrical measures of a patient can be compared to a normative database derived on a large sample of healthy individuals. The deviance from the database's norms provides a probabilistic measure of the likelihood that the patient's electrical activity reflects normal brain functioning. The focus of this thesis is the method for estimating such deviance. The method currently employed estimates the mean and the standard deviation of the normative sample. The deviance is then expressed in terms of z-scores. This method is referred to as the parametric method. The accuracy of the parametric method relies on the assumption that the distribution of the normative sample is gaussian, but this assumption is not always fulfilled in real qEEG and especially ET data. A new method based on percentiles ("non-parametric") is proposed. The parametric and the non-parametric methods are compared using simulated data. The accuracy of both methods is assessed as a function of normative sample size and gaussianity for three different alpha levels. Results suggest that the performance of the parametric method is unaffected by sample size (bigger than 100), but that non-gaussianity jeopardizes accuracy even if the normative distribution is close to gaussianity. On the contrary the performance of the non-parametric method is unaffected by non-gaussianity, but is a function of sample size only. It is shown that, with n>160, the non-parametric method can be considered always preferable. Results are discussed taking into consideration technical issues related to the peculiar nature of qEEG and ET data. It is suggested that the sample size is the only constant across EEG frequency bands, measurement locations, and kind of quantitative measures. As a consequence, for a given database, the error rate of the non-parametric database is homogeneous, however the same is not true for the parametric method.

Table of Contents

List of Figures

Introduction

The comparison to quantitative Electroencephalography (EEG) norms is a very valuable tool in both the electrophysiological research and clinical practice. Typically, the individual's electroencephalogram is analyzed in the frequency domain by means of time series analysis techniques such as the Discrete Fourier Transform (DFT), or the computationally advantageous Fast Fourier Transform, also called FFT (Beauchamp, 1973; Brillinger, 1975; Lynn and Fuerst, 1989). A certain number of features are extracted from the Fourier cross-spectral matrix, each one describing a particular feature of the brainwaves in a specified frequency range. These may include univariate and multivariate measures of absolute power, relative power and mean frequency of each electrode, in addition to coherence, phase and asymmetry of each electrode pair. Each individual's quantitative feature is called a descriptor. Descriptors are compared to norms derived under the same conditions from a sample of healthy "normal" subjects, allowing the statistical estimation of the deviance from the population norms. For example, one is able to estimate the individual's deviance of the magnitude in the alpha frequency range, or the deviance in the phase relationship in the beta range between electrode pairs. A recent trend in the electrophysiological literature is the derivation of norms for electromagnetic tomographic data (Bosch-Bayard et al., 2001). Electromagnetic tomographies (ET) make use of the EEG potential difference recording on the scalp to estimate the current density within the brain. Functional images of the current density distribution are then superimposed onto MRI standard atlas anatomical images (Talairach and Tournoux, 1988), providing true neuroimaging of electromagnetic brain activity either in the time or in the frequency domain (Fuchs et al., 1999; Pascual-Marqui, 1995, 1999; Pascual-Marqui et al., 1994). The derivation of norms for current density data in analogous to the derivation of norms for qEEG. In the former, electrical activity is not measured on the scalp at the electrode level, but estimated within the brain in discrete cubic regions of arbitrary size called voxels. Since, typically, one defines thousands of voxels, but makes use of only tens of electrodes, the comparison to ET norms poses more stringent statistical problems then the comparison to the qEEG norms. In both cases the deviance from each norm is usually expressed in terms of z-scores. QEEG and ET data is log-transformed in order to approximate a gaussian distribution of the normative sample. Means and standard deviations of the normative sample are computed for each norm and used as parameters for the z-score standardization. This method assumes gaussianity of the sample distribution and hereafter will be referred to as “parametric”. The assumption of gaussianity is not always matched with real data. This is particularly true in the case of ET data where there is a very high chance that among the tens of thousands distributions of descriptors, some will not be gaussian. The aim of this thesis is to propose an equivalent non-parametric method based on percentiles for the estimation of the deviance from the norms. The method applies equally well to qEEG and to ET data. It is clear that in comparing to EEG norms we make several assumptions on the nature of human EEG. Essentially we assume that the human EEG is a stationary process with relatively high intra-subjects and inter-subjects reliability. Before any statistical concern on the method to estimate deviances it is important to assess the tenability of such assumptions. Most of the initial work in this respect has been done and integrated by E. Roy John and his associates (Ahn et al., 1980; John et al., 1977, 1980, 1988). First, it has been shown that quantitative EEG measures follow developmental equations, meaning that the frequency composition of the EEG reflects the age and the functional status of the brain. In other words, in normal condition, their normal values depend on, and can be predicted by age (Ahn et al, 1980; John et al., 1980; Matušek and Petersén, 1973). Usually the relationship is quadratic on the log of the age. For example, the dominant frequency power of the normal EEG increases during brain development and declines slowly after age thirty or so (Bosch-Bayard et al., 2001; John et al., 1987; Szava et al., 1994). As a result, data from a wide age range database is modeled by means of polynomial regression equations in order to take into account the age differences (John et al., 1980). There is now evidence suggesting that EEG norms may vary slightly as a function of sex and hemispheric dominance (Veldhuizen et al., 1993). If these results are replicated, corrections for these two factors should be applied as well. Second, it is well known that the intra-subject spectral descriptors of the EEG are consistent over short period of time, probably as a result of stable homeostatic regulations of the neurotransmitters (Hughes and John, 1999). This is particularly true for the EEG recorded during a resting state where the subjects have the eyes closed, and for relative power measures (John et al., 1987; Matušek and Petersén, 1973). Another advantage of relative measures is that they are independent from subjective factors such as skin and skull thickness, and from a global scale power factor that increases inter-subjects variability (Hernández et al., 1994). For these reasons qEEG normative databases are usually generated for the eyes closed resting state only, and relative power measures are preferred. Third, normative quantitative EEG (qEEG) descriptors were found to be independent from cultural and ethnic factors. High reliability was found in studies from Barbados, China, Cuba, Germany, Holland, Japan, Korea, Mexico, Netherlands, Sweden, the United States, and Venezuela (quoted and referenced in Hughes and John, 1999).

The independence of the EEG spectrum from cultural and ethnic factors is a remarkable characteristic of the EEG. It has been suggested that it reflects the common genetic heritage of the mankind (Hughes and John, 1999). Fourth, qEEG norms proved to have high specificity and sensitivity. When subjects with no neurological or psychiatric dysfunction are compared to norms, only a few features show significant deviance (high specificity). On the contrary, when subjects with neurological or psychiatric dysfunctions are compared to norms, the number of significant deviant features greatly exceeds the number expected by chance alone (high sensitivity) (John et al, 1988). To date, comparisons to qEEG norms has been proven useful in the diagnosis of the attention deficit disorder with and without hyperactivity, learning disabilities, dementia, schizophrenia, unipolar and bipolar depression, anxiety disorder, obsessive-compulsive disorder, eating disorder, alcohol and substance abuse, head injury, lesions, tumors, epilepsy, and cerebrovascular diseases (For a review see: Hughes and John, 1999; Newer, 1988). For many other disorders and diseases, qEEG marks have been found, but additional research is needed to establish usefulness for diagnosis purposes. The four characteristics of the EEG power spectrum here mentioned can be considered the fundamental properties of qEEG, since they enable objective assessment of brain integrity in persons of any age, origin and background. It is important to keep in mind that EEG and qEEG are very peculiar functional neuroimaging techniques. For several reasons they cannot be simply replaced by more recent techniques such as Positron Emission Tomography (PET) and functional Magnetic Resonance Imaging (fMRI). For instance, EEG is truly non-invasive. Since recording units are portable and the recording process is inexpensive, EEG can be easily recorded for a long period of time and left in place, even in the operating room, in a ICU, or at the patient's bed. Finally, EEG has an unsurpassed temporal resolution, which may be decisive in analysis of brain dynamics, especially those that are very short in time. In summary: the comparison of quantitative electroencephalographic descriptors to databases derived from large samples of normal healthy individuals is a very valuable and powerful tool for the assessment of the integrity and changes of the brain's homeostatic regulation processes. The tenability of the assumptions underlying this procedure has been investigated for more than thirty years and is now well established. Some qEEG descriptors, namely relative power measures, have high test-retest reliability under eyes closed resting conditions. Under normal brain functioning conditions the EEG descriptors are dependent on age and other biological factors that can be taken into account, but they are independent of cultural and ethnic factors, hence enabling an objective assessment of brain integrity and changes over a lifetime. QEEG has been proven useful as a diagnostic aid in many neurological and psychiatric disorders or diseases. For many others promising research aiming to define qEEG marks is in progress. Since the spatial resolution of qEEG techniques and the nature of features that can be extracted from the EEG progress side-by-side with engineering, there is great hope in the electrophysiological community that EEG-related techniques of investigation will disclose more and more properties of the brain in the future.

Signal detection theory and Diagnostic Systems

In this section we briefly review some important concepts in the literature on signal detection theory. These concepts will be of great utility in discussing the weaknesses of the parametric and non-parametric methods for the estimation of deviances from the norms. These concepts will also provide us with a workable framework to compare the parametric and non-parametric methods, which is the aim of the simulation presented in this thesis. Normative databases are essentially diagnostic systems. The general task of diagnostic systems is to discriminate among possible states of the object under study, and to decide which one actually exists. In the case of normative databases, the task is to label the descriptor of the new individual as "normal" or "abnormal", or, using a more appropriate terminology, as "non-deviant" or "deviant". No diagnostic system is perfectly accurate, and all of them have to face the problem of the detection of the true signal from the actual signal received, which is a mixture of signal and noise. Modern detection theory treats the decision in probabilistic terms, according to which there are two statistical hypotheses. In the following discussion we will refer to a particular descriptor only. The arguments readily extend to whatever number of descriptors. The study of the accuracy of diagnostic systems sprang from the signal detection theory and is a common subject in the biomedical literature (Swets, 1988; Swets and Pickett, 1982). In comparing to norms, the system receives an input, the value of the descriptor, and takes one of two possible decisions. We will refer to the input, or actual status of the new individual, as the "event" (E). E can take on two mutually exclusive values. Let us label them as positive (+) or negative (-), which we will use hereafter instead of, respectively, "deviant" and "non-deviant". So, E+ is the event corresponding to a true deviance from the norms, and E- is the event corresponding to a true non-deviance from the norms. Notice that the status of the subject is given, and not observed. The system output is the decision taken. We will refer to this output, what the database decides, as the "diagnosis" (D). D also can takes on two mutually exclusive values. Following the same notation we will have D+ in the case of a positive decision (the new individual is decided to be deviant) and D- in the case of a negative decision (the new individual is decided not to be deviant). With two alternative events and two corresponding diagnosis, the data of a test of accuracy is conveniently summarized in a two-by-two contingency table (table1. All tables and figures are in the Appendix). We wish to obtain perfect correspondence between the events and the diagnosis, that is, we wish that the value of the descriptor for a new subject is labeled as deviant if it is in reality deviant and non-deviant if it in reality non deviant. These two outcomes correspond to the agreement (or concordance) between the input and the output of the diagnostic system, referred to in table 1 as true positive (TP) and true negative (TN). When there is no agreement then we have an error, which can be of two types: false positive (FP) and false negative (FN). If we consider proportions instead of raw frequencies of the four outcomes, then just two proportions contain all of the information about the observed outcomes (Swets, 1988). For instance we normalize each raw frequency in a cell by the column total. We have now:

$TP=TP/(TP+FN)$

$FN=FN/(TP+FN)$

$FP=FP/(FP+TN)$

$TN=TN/(FP+TN)$

In this way we obtain proportion estimations (analogous to probability values) bounded between zero and one and the following properties hold:

$TP+FN=1;$

$FP+TN=1$

In other words, the elements of the couples TP-FN and FP-TN are complement of each other and all the information about the observed outcomes can be obtained considering only one element for each couple. Furthermore, by normalizing the raw frequencies we obtain measures *independent of the prior probability of the event*, meaning that the estimation of errors will be independent of the proportions of positive events (E+) and negative events (E-) entered in the system (Swets, 1988). This is a fundamental property of any accuracy measure of diagnostic system. Figure 1 show these normalized measures in a different, albeit equivalent, perspective. Organizing the same data in a probability tree diagram we see that what we are computing, equivalently, are the probabilities to have positive or negative diagnosis (D+ and D-) *conditioned* on the probability that the event was positive or negative (E+ and E-). For example, the rate of normalized true positives is the probability to have a positive diagnosis given (conditioned on the fact) that the event was positive. In notation we write p(D+|E+). This quantity (normalized TP) is also referred to as 'Sensitivity' (SN) and is usually reported together with the normalized TN, or p(D-|E-), which is referred to as "Specificity" (SP). SN is a measure of the ability of the system to take a positive decision when it is indeed the case. Its complement is the normalized FN proportion. The SP is a measure of the ability of the system to take a negative decision when it is indeed the case. Its complement is the normalized FP proportion. According to what we have said before, SN and SP summarize the contingency table exhaustively. However, for the purpose of our simulation, a more complete depiction of the errors committed by a normative database is achieved considering two additional measures. These are the *inverse probability* of a true positive response and the inverse probability of a true negative response. Practically, what we want to know is the probability that a deviance exists when the system says it does, and the probability that a deviance does not exist when the system says it does not. These definitions are not just a play on the words (see previous definitions of SN and SP). We seek p(E+|D+) and p(E-|D-), which are, respectively, the *inverse* probability of SN and SP (To obtain those you need to invert the position of E and D). These probabilities are easily computed arranging the data as in figure 1 and using the formula defining the conditional probability or the Bayes' formula (Lipschutz and Lipson, 2000). The agreement E+D+ corresponds to the true acceptance of the alternative hypotheses "the new individual is deviant on that descriptor", while the agreement E-D- corresponds to the true rejection of this alternative hypotheses. Accordingly, we will refer to the quantity p(E+|D+) as "true acceptance" (TA) and to the quantity p(E-|D-) as "true rejection" (TR). For reasons that will be clear later, only considering together SN, SP, TA, and TR, will enable us to perform a complete and fair estimation of the systematic error rate for the parametric and non-parametric methods.

The parametric method based on z-scores

We are now ready to turn to the issue of deviance estimation. The steps required in order to build a normative database according to the parametric method (PM) and to the non-parametric method (nPM) are listed in table 2. The focus of this thesis is steps 5 and 7, and in fact, these are the only two steps where the procedures for the PM and the nPM differ. We are concerned here with the way in which the significance of the deviance is estimated. We will not discuss the sampling of the normative subjects (which determine the homogeneity of the normative sample), or the issue of multiple comparisons (which is essential to avoid false positives). To date, to my knowledge, all published normative databases estimated the significance of the deviance according to a parametric method based on z-scores (Bosch-Bayard et al., 2001; John et al., 1987; Thatcher, 1999; Veldhuizen, Jonkman, and Poortvliet, 1993). The work of John and his colleagues was decisive for the development and assessment of this statistical methodology (John et al., 1977). When z-scores at each electrode location are interpolated to construct brain topographical maps, the result is called "Significance Probability Mapping", or SPM (Duffy et al., 1981). In step 3 of table 2 we defined the descriptors of our database. According to the notation used in table 2, there are d=LxF descriptors for each normative subject, i.e, for each subject there is a descriptor for each combination of location (electrode for qEEG and voxel for ET) and feature (quantitative measure in a specified frequency range). For example, a descriptor is the relative power in the alpha range, and another descriptor is the relative power in the theta range. Thus, each descriptor can be conceived as a vector comprised of N values, where N is the number of subjects in the database. Let us call x_d the vectors of descriptors, where the subscript d denotes a particular descriptor in the LxF matrix. For each feature, the appropriate log-transformation is applied to all subjects (John et al., 1987). The resulting data distribution of the vectors x_d is approximately normal with mean y_d and standard deviation σ_d. In step 6 we considered the LxF matrix of descriptors referring to a new individual to be compared to the database. Notice that the LxF matrix for the normative database is a matrix of vectors, i.e., that is a 3-D matrix. Instead for any new individual the LxF matrix is a 2-D matrix of individual entries. Identical log-transformations are applied to this matrix as well. Let us call $\hat{y}_d$ each entry of the descriptor matrix for the new individual. The task is to obtain an estimation of the deviance, from the mean of the x_d, for each $\hat{y}_d$. Given gaussianity of the normative sample distribution, the deviance of the new individual for each descriptor d is estimated as

$z_d=(\hat{y}_d - y_d) / \sigma_d$ [1.0]

The z-scores computed with 1.0 are accurate if the normative sample distribution is normal (gaussian). The more the normative sample distribution deviates from normality, the less the z-scores will be accurate, leading to more and more false negatives and false positives as a function of the distribution skewness and kurtosis. Skewness refers to the third moment around the mean of a distribution and is a measure of asymmetry. For example, a chi-square distribution with one degree of freedom is said to be right-skewed. Kurtosis is the fourth moment around the mean and is a measure of the peakedness of the distribution. A "flat" distribution has higher kurtosis then a "peaked" one. A theoretical standard normal distribution has skewness=0 and kurtosis=3. Given an approximate gaussian distribution, the more these two values deviate from the theoretical values, the more the distribution deviates from gaussianity. The problem with the rate of false positives and false negatives in the case of non-gaussian distributions is a subtle one. We could tolerate it if we could assess, or at least estimate, the rate of false positives and false negatives and if these rates would be the same on the two sides of the distributions. Unfortunately this is not the case. Indeed, with estimation [1.0] we will obtain different rates of false positives and false negatives depending on the side of skewness (left-skewed or right-skewed distribution) and the side of the test (left-handed or right-handed test). Similar arguments apply to the amount of kurtosis. The effects of skewness and kurtosis on the rate of false positives and false negatives are easily captured in a graphical fashion (figure 2). This figure is crucial for the interpretation of the results of the thesis and should be analyzed carefully by the attentive reader. Figure 2a depict a normative sampling distribution very close to the theoretical gaussian. Suppose that distribution is indeed gaussian. With an alpha level of 0.05, the decision criterion of the database is to label as "deviant" all new observations with z-score >1.96 or <-1.96 (the area under the curve for z>1.96 or z<-1.96 equals 0.025, so their sum is 0.05). Let us consider the right-handed test first. A z-score exceeding 1.96 leaves on its right a proportion of the area under the curve less then 0.025. So the diagnosis will be positive (D+). By definition, a new individual's score with p<0.025 is positive (E+). The result is a concordance between the event and the diagnosis (true positive). Because of simmetricity, for a left-handed test the result will be the same. For all z-scores comprised between –1.96 and 1.96 both the event and the diagnosis will be negative (E- and D-), and we will have concordance again (true negative). We can see that if the normative sampling distribution is gaussian, the normative database will virtually commit no error. Figure 2b depict a normative sampling distribution right skewed. Notice that the mean of the distribution (blue line) is no longer at the peak of the distribution since the density on the right side of the distribution is bigger than the density on the left side. The two violet vertical lines delimitate the interval including 95% of the density (area under the curve). On the right of the right violet line the density is 0.025%, and so it is on the left of the left violet line. Let us consider the right-handed test first. Because of skewness, for some z slightly bigger than 1.96 (D+), the area under the curve on the right of the z-value is greater than 0.025 (E-). The diagnosis is positive (z>1.96), but the event was not (area>0.025). We have a false positive. In figure 2b, the right-sided z interval for which a false positive will happen is indicated in green. For the left-sided test the situation is opposite. Here for some z>-1.96 (D-) the area under the curve is already less than 0.025 (E+). The diagnosis is not positive, but the event was indeed positive. We obtain a false negative. In figure 2b, the left-sided z interval for which a false negative will happen is, again, indicated in green. If the distribution is left-skewed, we would have obtained mirror results, i.e., false negatives on the right side of the distributions and false positive on the left side. At this time it should be clear that with skewed normative sample distributions we obtain different types of error on the two sides of the distribution. This means that what in reality are equivalent, but opposite, deviances, are interpreted by the diagnostic system differently, according to the sign of the z-score. If the amount of error generated is not negligible, this property of the parametric method would constitute a serious problem. Therefore we need to estimate it, and this will be accomplished in the simulation we are going to present. Before that, let us introduce an alternative method for the estimation of the deviance, a non-parametric method based on proportions.

The non-parametric method based on proportions

The parametric method relies on the assumption of normality of the distribution. In a one-sided testing framework (see figure 2a for a graphical representation), a z-score =1.645 means that, on the theoretical normal distribution, the 95% of the population fall below that value. In other words, only 5% of the population exibits a value equal or greater. The corresponding value on the other side of the distribution is –1.645, for which only 5% of the population exibits a value equal or smaller. A non-parametric method, to obtain a similar result, is by use of the sample proportion (sp). Sample proportions are analogous to percentiles and, like them, are obtained by sorting the sampling distribution values. The method is easily illustrated with an example. Refer first to a *right-handed test* with alpha=0.05. In this case we label a new individual as deviant if his/her value is large as compared to the normative database. For example, if the descriptor under analysis is the alpha relative power at the electrode O2, then a deviant subject will show a large power value as compared to the norm. Suppose our normative sample is comprised of 20 subjects (N=20). Let us sort the normative values referring to any descriptor *d* in ascending order to obtain the sorted x_d vector:

2 2.5 2.8 3.5 3.6 3.7 4 4.9 5.2 5.7 8.4 8.5 11.1 12.3 14.8 16.4 18.9 20 21 25.4

The 95th percentile is the value below which the 95% of the subjects fall. Values comprised between 21 and 25,4, leave on the right-side 5% of the observations (5% of 20= 1). A value equal to or bigger than 21 is associated with a p-value <0.05. We obtain a p-value with a counting random variable (Holmes et al, 1996). Let us define the discrete random variable (RV), sample proportion (Φ) as the *proportion of values in the x_d vector falling above the new individual's value*. Then, Φ is indeed a p-value, although it is discrete and not continuous. By definition, if no value in the x_d vector exceeds the new individual's value, then Φ=0. In this case in fact the new individual shows the most extreme value and this is as significant (unlike) as it can possibly be. With this definition the discrete RV Φ can take on N+1 values ranging from 0 to 1 and decreasing by multiples of 1/n. Φ =1 (20/20=1) means that all normative subjects exceed the new individual's value. In this case in fact the new individual's value is the smallest, and there is no evedence at all that the new individual's value is significant (keep in mind that if our test is right-handed we have to ignore the extreme values on the left of the distribution, no matter how extreme they are). Φ =0 means that the new individual exibits the most extreme value. Suppose our new individual's value for the descriptor *d* is 22.3. Comparing this value to the sorted vector above we see that 5% of the observations fall above this value, thus Φ is 0.05 (there is only 1 observation falling above the value 22.3; 1/20=0.05). Suppose the value is 1.8; Φ is 1 (20/20=1). Suppose it is 5.4; Φ is 0.55 (11/20). Φ = 0.05 can be considered deviant just like a z-score=1.645. Both correspond to a probability of 0.05, with the difference that in a non-parametric fashion the p-value is computed on the actual data and not as a result of the integrals of the theoretical normal distribution. The same method, reversed, is applied in the case of a *left-handed test*. In this case the discrete random variable (RV) sample proportion (Φ) is defined as the *proportion of values in the x_d vector falling below the new individual's value*. By definition, if all values in the x_d vector are bigger than the new individual's value, then Φ=0; In this case infact the new individual's value is the smallest and this provides the strongest evidence for his/her deviance on the left side of the distribution. With this reversed definition the discrete RV Φ still can take on N+1 values ranging form 0 to 1 and increasing by multiples of 1/n. Φ =0 means that all normative subjects exceed the new individual's value. Φ =1 means that the new individual's value exceed all normative subjects. Suppose again our new individual's value for the descriptor *d* is 22.3. For a left-handed test, comparing this value to the sorted vector above we see that 95% of the observations fall below this value, thus Φ is 0.95 (there are 19 observation falling below the value 22.3; 19/20=0.95). Suppose the value is 1.8; Φ is, by definition, 0. Suppose it is 5.4; Φ is 0.45 (9/20). If a two-tailed test is wished, then the median of the distribution is computed. If the new individual's value is on the right of the median then a right-handed test as described is performed. On the other hand, if the new individual's value is on the left of the median then a left-handed test is performed. Of course, for a two-tailed test we need to halve the alpha level at the two sides of the distribution, so that the total alpha level equals indeed alpha. For simplicity of display purposes we usually wish to convert the sp estimation in a scale comprised of both positive and negative values.

If Φ is the sample proportion (with 0< Φ <1) then this is accomplished, for example, by

$$\Phi' = (\Phi - 0.5) \mathrm{x} 2 \qquad (2.0)$$

As a result of the above tranformation the deviance Φ' will now be comprised between –1 and 1. (In fact, (0-0.5) x 2 = -1, and (1-0.5) x 2 = 1). After the tranformation, Φ' = –1 corresponds to Φ = 0, Φ' = 0 corresponds to Φ=0.5 (the median), and Φ' = 1 still corresponds to a Φ=1. The reason for this transformation is that maps generated are easier to interpret. To get the true sample proportion estimation (Φ) one can use the reversing formula:

$$\Phi = (\Phi'/2) + 0.5 \qquad (2.1)$$

Given the way in which the deviance is estimated with the transformed sample proportion Φ', the interpretation of maps is easier. Alternatively one can convert the p-value obtained in a z-score using the integrals of a theoretical normal distribution. The result would be equivalent and in both cases the transformation will help the interpretation of maps. The performance of the non-parametric method here described is not affected by non-gaussianity of the sampling distribution. However its performance is a function of the sample size. Considering sample proportions we define a discrete RV, but the underlying phenomenon is continuous, hence we loose "resolution".

In the following simulation we assess the amount of errors generated because of this loss of resolution and we compare it with the amount of error generated by the parametric method.

Simulation

Method

In order to perform a simulation aiming to evaluate the performance of a normative database we need to define uniquely positive events (E+) and negative events (E-), i.e., we need to delineate conditions under which a simulation entry is by definition deviant or non-deviant. Any particular method to take a decision about the deviance of the event will provide a diagnosis, either positive (D+) or negative (D-), according to its own procedure, and being unaware of the real status of the event. The agreement, or concordance, between the event and the diagnosis can then be estimated. By allowing a large number of events to enter the system we obtain reliable estimations of concordance and discordance. In order to define unambiguous positive and negative events we need to refer to theoretical distributions for which the "true" acceptance interval of the null hypothesis is known. For instance, let us set the type I error (alpha) as 0.05. For a random variable z distributed as a standard normal we accept the null hypothesis for $-1.96<z<1.96$. In words, if z is comprised between –1.96 and 1.96, we accept the null hypothesis. In terms of a normative database this means that the new individual is considered to be normal. In our simulations the normative sample of reference was emulated by means of normal distributions. New individuals were emulated as individual random samples distributed in the same way as the normative reference. For all practical purposes they constitute events for which the status (E+ or E-) is known a priori on the basis of the distribution of the normative reference. In the discussion that follows we will call each event submitted to diagnosis a *simulation entry*. As an example of the procedure followed to define simulation entries consider the following; given a normative reference sample distributed as a random normal, alpha=0.05, and a right handed test, we know a priori that any simulation entry with $p(z)<0.025$ is positive. For each simulation entry, we computed the database outcome (D+ or D-) with both the parametric and non-parametric method, independently one from the other. According to what is seen above, the parametric diagnosis is based on equation [1.0], and the non-parametric diagnosis is based on the RV sample proportion. For each simulation entry there will be a concordant or discordant outcome and this will add a raw frequency in a table just like table 1. This constitutes an outcome among four possibilities (table1). We submitted 100.000 simulation entries, under identical conditions, for each normative reference sample considered. This allowed reliable estimations of sensitivity (SN), specificity (SP), true acceptance (TA) and true rejection (TR). The evaluation of concordance was repeated varying sample size and gaussianity of the normative reference sample. This way we could assess the error rate of the parametric and non-parametric method in critical situations. In addition, we repeated the simulations for three alpha levels (decision criterion of the system). The latter variable must be included because all of the four measures of accuracy we chose depend on the decision criterion used (Swets and Pickett, 1982), therefore we need to monitor the error rate as a function of alpha. Finally, two simulations for all the above conditions are needed, with one evaluating the right-handed test, and the other evaluating the left-handed test. The reason for this further splitting is that, as we have shown above in the case of skewed distributions, the parametric method generates two different types of error at the two sides of the distribution and we do not want to confuse them considering the outcomes of a two-sided test. A total of 486 (9x9x3x2) simulations were performed, each one evaluating 100.000 simulation entries. The simulations were performed by a computer program written in Delphi Pascal (Borland Corporation). All together they required approximately 4 hours computation time on a Dell personal computer equipped with a 1.8 GHz Pentium 4 processor and 512 Mb of RAM. Normative samples, the x_d vector described above, were emulated by means of a gaussian random number generator function embedded in Delphi Pascal. The function (called randG) generates random samples gaussian-distributed with a specified mean and standard deviation. For all simulations we used mean=10 and variance=1. In this way all random samples were non-negative. This was required by the skewness manipulation we chose (performed by means of a power transformation as seen below). Each distribution actually employed in the simulation was computed as the (sorted sample-by-sample) average of 10,000 gaussian distributions generated with the randG function. This ensured that correspondent distributions were very similar across different conditions of the simulation.

Alpha level manipulation

The alpha level is the decision criterion employed in the normative database. It quantifies the amount of evidence requested by the system before a positive outcome is issued. Three alpha levels were considered: 0.05, 0.025, and 0.0125. Since all tests were one-handed, these three levels correspond to the two-handed test alpha levels 0.01, 0.05, 0.025. Keep this in mind while analyzing results. All published databases considered in our review (Bosch-Bayard et al., 2001; John et al., 1987; Thatcher, 1999; Veldhuizen, Jonkman, and Poortvliet, 1993) use the fixed alpha level 0.05. In our simulations this corresponds to alpha=0.025. In addition to this alpha level we considered a more stringent criterion (alpha=0.0125), and a more lenient criterion (alpha=0.05). The reason is that the measures of accuracy we used are independent of the prior probabilities of positive or negative events, but are not independent of the decision criterion (Swets and Pickett, 1982). Since we expect different error rates solely because the decision criterion is changed, we might want to monitor the behavior of our system as a function of the decision criterion.

Sample size manipulation

Nine sample sizes were considered, ranging from 80 to 720 with increment of 80 (80, 160, 240, 320, 400, 480, 560, 640, 720). The choice for the increment was contingent. It can be shown that the accuracy of the non-parametric method for the minimum alpha level we considered (alpha=0.0125) increases discretely in steps of 80 (sample size). The reason is intuitive. We shown that the RV sample proportion (Φ) can take on only discrete values ranging between 0 and 1 increasing by a factor of 1/N. Consider the alpha level alpha=0.0125. With N=80, the possible values that the RV Φ can take, sorting them in ascending order, are 0, 0.0125, ... 1. With N=160, they will be 0, 0.00625, 0.0125 .... 1. As you can see now, as soon as N reaches 160, the random variable Φ gains resolution, having the ability to take on three possible values less than the alpha level (p<alpha).

Gaussianity manipulation

Gaussianity was manipulated transforming the normal averaged distribution with a power function. For each level of gaussianity considered each sample of the normative distribution was raised to a fixed power. This resulted in a skewed distribution respecting the order of the original samples. Nine levels of gaussianity were considered, corresponding to nine different powers ranging from 1 to 3 with an increment of 0.25 (1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3). The first distribution always remained unchanged after transformation (power of 1) and constituted a true empirical random gaussian distribution. In this case the performance of the parametric method was expected to be excellent. The frequency polygons (analogous to the more common histogram) of the nine distributions actually employed for both the left-handed and right-handed tests, are shown in figure 3. More precisely, shown are the empirical distributions generated for the maximum level of sample size (N=720), removing the displacement along the x-axis due to the power transformation. Table 3 reports the mean and standard deviations of the skewness and kurtosis of the empirical distributions actually used in the left-handed test and right-handed test simulations. Mean and standard deviation were computed across the different sample sizes used in the simulations. From table 3 we can see that, because of the averaging procedure, the gaussian random distributions had all very similar skewness and kurtosis for all the levels of sample size (small standard deviation), yielding almost identical distributions to be used in the left-handed test and right-handed test simulations. Table 3 also shows how skewness deteriorates with higher powers. The same effect is seen graphically in figure 3. Figure 3 shows also that the actual empirical distributions used in the simulations were pretty smooth. We can now state more precisely how we defined positive and negative events and how we computed positive and negative diagnosis in our simulations. For power=1 the reference distribution remains unchanged. It will be normal with mean=10 and variance =1. Simulation entries are random samples of the same distribution. We need to define a confidence interval for a simulation entry to be defined positive (E+) or negative (E-). This is accomplished by means of integrals of standard normal distributions. For example, for alpha=0.05 and a right-handed test, the confidence interval is given by [x=-∞; x=(10+1.96)]. In other words, p<0.025 (E+) is obtained only for simulation entries which value is bigger than 11.96. With x<=11.96 the event is negative (E-). The diagnosis outcome for the parametric method is given by means of the theoretical probability associated with the z-score. The z-score is based on the empirical mean and standard deviation of the reference distribution ([1.0]). The diagnosis outcome for the non-parametric method is given directly by means of the random variable Φ, which express already a p-value. The diagnosis outcomes for the two methods results in either D+ or D-. They are computed independently from each other and independently of the real status of the simulation entry (E+ or E-). We obtaine in this way the status of the event and the status of the diagnosis according to both the parametric and the non-parametric method. Therefore a concordant or discordant outcome is established for each method and the respective counts are increased in tables like table 1. For higher powers the mechanism is the same but the confidence interval and the simulation entries are defined accordingly. For example, the confidence interval for power=1.25 is given by $[x=-\infty; x=(10+1.96)^{1.25}] = [x=-\infty; x\cong 22.2415]$. The simulation entries are now sampled with a normal distribution with mean 10 and variance 1 and raised to the power of 1.25, before a diagnosis is issued. The above examples can be easily extended to all simulations employed, i.e, for every level of gaussianity, sample size and alpha.

Results

To capture the essence of our results we need to consider again figure 2b. Let us anticipate the results for the parametric method. For a right-handed test, since the distribution has positive skewness, we expect three possible outcomes: E+D+ (red area on the right of the distribution), E-D+ (green area on the right of the distribution), and E-D- (all the area left). The only discordant outcome (error) is the E-D+ pairing. These are false positives. The error is due to the fact that although the area on the left of the observation is bigger than alpha (E-), the z-score computed with [1.0] is bigger than 1.96, leading to a p-value less than alpha (D+). Since this error happens on the right side we wish to compare it to the TP proportion. In other words (referring to figure 2b), we wish to compare the green area (error) with the red area on its right. We will show now that the specificity measure (SP) does not give us this information, but the true acceptance measure (TA) does. Remember that SP has been defined as TN/(FP+TN). Remember also that TN= p(D-|E-) and FP=p(D+|E-). In our simulations most entries are negative events. In fact the simulation entries were always random samples of the normative sample distribution. Hence (1-alpha)% of them is by definition a negative event and will fall in the E-D- (TN) category. The remaining will include E+D+ and E-D+ outcomes. Even if the FP proportion is large as compared to the TP proportion (the green area is big as compared to the red area) the specificity will be excellent, since it does not compare FP with TP, but FP with TN. On the other hand TA, defined as p(E+|D+), has as complement p(E-|D+). Its value is the right estimation of errors for this simulation, i.e., it compares the FP proportion to the TP proportion. This is the information we need; it is telling us among the events with positive diagnosis (green area +red area), how many, in proportion, were in reality positive (TP: red area) as compared to negative (FP: green area). Consider next the left-handed test. Refer again to figure 2b. Here we expect three different possible outcomes: E+D+ (red area on the left of the distribution), E+D- (green area on the left of the distribution), and E-D- (all the area left). The only discordant outcome (error) is the E+D- pairing (false negative), which is different from the type of error found on the right side. Here the error arises because although the area on the left of the observation is less than alpha (E+), the z-score computed with equation [1.0] is bigger than –1.96 (non-significant), leading to a p-value less than alpha (D-). We obtain some false negatives. Again, we wish to compare them to the TP proportion, and not to the TN proportion. In this case the sensitivity measure (SN) will give us this information. Remember that SN has been defined as TP/(TP+FN). Remember also that FN= p(D-|E+) and TP=p(D+|E+). For a left-handed text, (1-alpha)% of the outcomes will fall in the E-D- category (notice that on this side of the distribution errors (FN) come at the expenses of the TP proportion and the TN proportion is exactly (1-alpha)%). The remaining 5% will include E+D+ and E+D- outcomes. SN compares indeed TP to FN. It is telling us among the positive diagnosis, how many, in proportion, were in reality positive events (TP) as compared to negative events (FN). Errors with the non-parametric method follow a different pattern. For this method the best measure of accuracy turns out to be the true acceptance (TA) for tests on both sides of the distribution. This means that for both the right-handed and left-handed test, the non-parametric method results in only three outcome pairings: the two concordant pairs E+D+, E-D-, and the discordant pair E-D+. In other words, the non-parametric method tends to issue positive diagnosis when it is not the case. In summary, considering that real normative distributions can be both left and right skewed, with the parametric method we expect both FP and FN errors depending on the side of the test and on the side of the skewness. With the non-parametric method we expect FP only, regardless the side of the test and the side of the skewness. We now show quantitative results of these errors. As expected, the accuracy of the parametric method was found to be the same (with little random error) at different sample sizes (N>100) for all levels of non-gaussianity and alpha. Thus it will be shown as a function of non-gaussianity and alpha levels only. The accuracy of the non-parametric method was found to be the same (with little random error) at different non-gaussianity levels for all levels of sample size and alpha. Thus it will be shown as a function of sample size and alpha levels only. In every simulation performed, two out of the four measures of accuracy employed in this thesis always display a value of 1.0 (perfect accuracy) for all levels of the manipulated variable, i.e., they do not constitute a valuable test at all. The reason why this is the case has just been discussed. For example, for a right-handed test we do not expect false negatives at all for both methods, regardless the gaussianity, sample size, and alpha. The remaining two measures are reported in the results section. They both always displayed values of accuracy less then or equal to 1.0 and changed monotonically across the levels of the manipulated variables. However only one of these two measures provides us with a useful estimation of inaccuracy. We have just seen that those are either the SN or the TA for the parametric method, and the TA for the non-parametric method.

Right-handed test

Results for the right-handed test are reported in figure 4. 4a and 4b refers to the parametric method (PM), while 4c and 4d refers to the non-parametric method (nPM). The blue lines indicate the 0.95 level of a measure of accuracy. This level of accuracy can be considered excellent for any diagnostic system. The red lines indicate the 0.85 level of a measure of accuracy. This level of accuracy can be considered the minimum required for a normative database. Figure 4a reports the PM true acceptance (TA) proportion as a function of gaussianity of the normative reference sample (x-axis) for the three alpha levels employed. As explained in the above discussion, this is the critical test for the parametric method for a right-handed test when the reference distribution is right skewed. The TA is excellent in the case of normality of the reference distribution (power of 1) and deteriorates rapidly as the power increases; for power>1.5 the TA proportion for the usual alpha level (0.025) is unacceptable (<0.85). Figure 4b reports the specificity (SP) of the parametric method. For all alpha and gaussianity levels the SP is excellent. We commented before that this last result is meaningless, and the figure is here included for completeness (hereafter we will describe critical tests only). The critical test of the nPM method under identical conditions is shown in figure 4c. This graph plots the TA proportion as a function of the sample size. As expected, the performance of the nPM increases monotonically with N. For the usual alpha level (0.025), the performance is acceptable (TA>0.85) for N=160, and excellent (TA>0.95) for N=400 or more. Figure 5 reports the uncollapsed data of the critical tests for the two methods with alpha=0.025. For the PM method, data has been expanded over the nine levels of sample size, while for the nPM method data has been expanded over the nine levels of distribution gaussianity. These graphs do not add any information not already shown, but depict the rationale for collapsing data across sample size levels for the PM, and across gaussianity levels for the nPM. Corresponding graphs showing collapsibility in the case of the left-handed test will not be shown, being practically identical.

Left-handed test

Results for the right-handed test are reported in figure 6. 6a and 6b refers to the parametric method (PM), while 6c and 6d refers to the non-parametric method (nPM). The blue lines indicate the 0.95 level of a measure of accuracy. This level of accuracy would be considered excellent for any diagnostic system. The red lines indicate the 0.85 level of a measure of accuracy. This level of accuracy can be considered the minimum required for a normative database. For reasons that should be clear now, we will describe graphs in figure 6a and 6c only. Figure 6a reports the PM Sensitivity (SN) proportion as a function of gaussianity of the normative reference sample (x-axis) for the three alpha levels employed. As explained in the above discussion this is the critical test for the parametric method for a left-handed test, when the reference distribution is right skewed. The SN is excellent in the case of normality of the reference distribution (power of 1) but deteriorates rapidly as the power increases. The decline is faster for the left-handed test than for the right-handed test (compare with figure 4a). This phenomenon can be easily captured inspecting the two tails of the distribution in figure 2b and considering the definition of SN and TA. In fact on the left side the distribution the proportion of errors (green area) grows at the expenses of the true positive proportion (red area), while on the right size the proportion of error (green area) grows at the expenses of the true negative proportion (all the area remaining on the left of the green area), hence the true positive proportion (red area) here remains unchanged. For power>1.25 the SN proportion for the usual alpha level (0.025) is already unacceptable (<0.85). The critical test of the nPM method under identical conditions is shown in figure 6c. This graph plots the TA proportion as a function of the sample size. Like for the right-handed test, the performance of the nPM increases monotonically with N. For the usual alpha level (0.025), the performance is acceptable (TA>0.85) for N=160, and excellent (TA>0.95) for N=480 or above. Allowing little random errors, these results for the nPM are comparable to those obtained for the right-handed test. In fact the nPM performs equally at both sides of the distribution, no matter what the skewness is.

p-values histograms

Further insight in the behavior of the two methods is provided by the histograms of the p-values. Each simulation entry leads to a binary decision of the system. However the system bases its decision on a continuous random variable, which is the p-value associated with the simulation entry. If this p-value exceeds a certain threshold, then the null-hypothesis is rejected and the diagnosis is positive. For a one-handed test (see figure 2a) the p-value of a z-score is the cumulative distribution function (cdf) of the standard normal distribution. Consider a right-handed test. The cdf corresponds to 1 minus the integral form –infinity to z, or, 1 minus the area under the curve from x=-infinity to x=z (remember that the total area under the curve equals 1). If we keep sampling from a standard normal distribution we obtain a uniform distribution of the p-values associated with the samples, with all p-values comprised between 0 and 1. Our simulation entries are nothing more than repeated samples of a transformed normal distribution. When the power equals 1, of course, the transformation does not alter the distribution at all. Figure 7 shows the histograms of the p-value computed from the two methods for both the right-handed and left-handed tests, in the case of alpha=0.0125, sample size=320, and power tranformation=2. This simulation refers to a strict alpha level, an intermediate sample size, and a moderate-high amount of skewness. Notice that while the non-parametric method yields a uniform distribution of p-values, the parametric method yields an irregular distribution, with the frequency of extreme p-values overestimated at one tail of the distribution and underestimated at the other. The left-handed and the right-handed test lead to mirror results. The reason why it happens is the distribution skewness (see the distribution corresponding to power of 2 in figure 3), which makes the density of the curve to be different at the two tails. The interested reader will find an explanation for the irregularity of the p-values distribution produced by the parametric method, analyzing and comparing figure 2b and figure 7. The interpretation of the phenomenon is as follows; regardless the distribution skewness, submitting random simulation entries to a non-parametric database will return a uniform distribution of p-values, as it should be. In the case of skewed distributions the parametric method, instead, will judge the significance differently, overestimating or underestimating entries falling at the very end of two tails of the distribution. The more the skewness, the more important the error in the estimation.

Conclusions

A total of 486 simulations were performed in order to compare two methods for the comparisons to EEG norms. The parametric method is based on z-scores and has been employed so far. The non-parametric method is based on sample proportions, or, equivalently, percentiles, and has been proposed in this thesis to overcome some problems related with the use of the parametric method. Each simulation estimated the error rate in the diagnosis of the two methods for both left-handed and right-handed test. Variables manipulated included the decision criterion of the normative database (alpha level), sample size, and non-gaussianity of the normative reference sample. For each combination of the side of the test and the method employed, the critical test was individuated. This was one of the four accuracy measures considered in this study (Sensitivity (SN), Specificity (SP), True Acceptance (TA), and True Rejection (TR)). The performance on the critical tests provided a framework for comparing the two methods. The performance of the parametric method (PM) was found to be unrelated to the sample size, given that N is not too small. With N=80 the performance of the method starts deteriorating (figure 5), therefore we conclude that this independence is true for N>100. The performance of the parametric method was found related to the non-gaussianity of the normative sample distribution. Typical distributions for which the parametric performance can be considered acceptable are shown in figure 8. Notice how close to gaussianity (black distribution) the two skewed distributions (green and red) for which the error rate was found to be acceptable. For a left-handed test the situation is even worse, with the simulation referring to the red distribution already leading to an unacceptable error rate (figure 6; see parametric sensitivity for power of 1.5). The performance of the non-parametric method was unaffected by the non-gaussianity of the normative reference distribution but was affected by the sample size. Acceptable (>0.85) accuracy (enough resolution) can be attained with N=160. Excellent accuracy (>0.95) can be attained with no less than around 440 subjects. This result contradicts the common notion that non-parametric statistics "should be used with a small sample size". For both methods, for both the right-handed and left-handed tests, the critical tests results in worse accuracy the smaller the decision criterion (alpha level). This important result contradicts the intuitive notion that reducing the alpha level would lead to a smaller rate of false positive. This is definitely not the case. Indeed alpha affects positively all measures of accuracy proportionally to its value; the bigger the alpha level, the better the accuracy. This result is here explained with a specific example. The reasoning extends readily to all possible situations. Consider the left-handed test for the parametric test. The critical test for this situation is the sensitivity (SN). Remember the SN is defined as TP/(TP+FN) and that under these circumstances the database is going to issue only TP, FN and TN outcomes. SN increases proportionally as TP increases and as FN decreases. Refers to figure 2b and look at the left tail of the distribution. This figure refers to a one-handed alpha level equals 0.025. Imagine we halve the alpha level. Both the green area under the curve (FN) and the red area under the curve (TP) will decrease (they will be displaced on the left and here the height of the curve is smaller). However the red area will decrease more than the green one, the reason being that the curve is shorter at the left extremity. As a result, the ratio TP(TP+FN) will be smaller, i.e., sensitivity will se smaller. Doubling the alpha level, on the contrary, will result in a sensitivity increase.

Discussion

We have been shown by means of simulations that the performance of the parametric test is impaired as a function of skewness. Non-gaussianity due to high or low kurtosis is known to affect the test even more (Pollock et al., 1990). These results are not a surprise. The problem is to assess how good the approximation to gaussianity for qEEG and ET data is, and to evaluate the advantages acquired by using an alternative method. About the approximation to gaussianity the literature is scattered and inconsistent. Only a few studies have been investigating specifically the gaussiantity approximation for qEEG data, and none, to my knowledge, investigated the gaussianity approximation for electromagnetic data. Nonetheless the same transformations applied for qEEG measures have been recently applied to this kind of data to generate a normative database (Bosch-Bayard et al., 2001). Electroencephalographic data in the frequency domain is markedly non-gaussian. Each measure is distributed in a particular way and the theoretical studies on their distribution are not exhaustive. For example, the power spectrum (absolute power) is distributed approximately as a chi-square (Beauchamp, 1973; Brillinger, 1975). The degrees of freedom (df) are a function of the EEG recording length (number of epochs), the FFT frequency resolution, wideness of the frequency band considered, the time-domain tapering employed, and other technical factors. One should take into consideration all these factors in estimating the df associated with a power spectrum chi-square distribution. It is in principle possible to adjust the power spectrum estimate so as to have a large number of df and the same number of df for all electrodes, frequency bands, and subjects, but this is elaborate and it is not usually done. Even doing so, one should then model the data by means of the chi-square distribution, and not by means of a normal distribution. For another measure, say the relative power, one should try to model the distribution of the measure with some theoretical distribution and adjust the FFT algorithm to obtain estimations conforming to that distribution consistently across electrodes, frequency bands, and subjects. At the time when the databases were first developed (1970's) a more simple approach was decided. For each measure a suitable data transformation (based on the log) was employed to approximate gaussianity. The idea was to allow a general method for the assessment of the deviance form the norms (the parametric method described above) and also to allow parametric statistics to be employed in research comparing groups. It has been this way for the past several years. A few specific studies provided evidences of the appropriateness of these transformations (Gasser et al., 1982; Oken and Chiappa, 1988; Pollok et al., 1990). Other evidence has been provided in papers describing the construction of normative databases but they are not as stringent studies from a statistical point of view (e.g.: John et al., 1988). A review of the literature convinced us that the gaussian approximation is not good enough to allow the use of parametric statistics. All specific studies found that the log-based transformations approximate fairly well gaussianity, but all of them found exceptions. Gasser et al. (1982) found exceptions in Delta, Theta, Beta 1 and Beta 2 for the absolute power measures. Oaken and Chiappa (1988) found approximately 1/8 of the descriptors for absolute power to be still non-gaussian after transformation. Relative power behaved a little better. Pollock et al. (1990) found the transformation of amplitude (square root of absolute power) to be excellent in all frequency bands but in theta. While John and his colleagues (1987, 1988) insist on data transformation, Thatcher (1998) found that for all measures, with the exception of phase, the untransformed data approximated gaussianity better than the transformed data, contradicting all previous results. It is worth noting that the sample size used in John's and Thatcher's studies was similar, so the unreliability of results cannot be explained by means of "deus ex machina" such as the central limit theorem. Furthermore, all of these studies used different montages, electrode reference, age range of subjects and even different measures. Finally, if in the case of qEEG a few proportions of departure from gaussianity can be ignored, for ET data it cannot be done so lightly. Before compiling a parametric database one has to check that the distribution for all descriptors is approximately gaussian. In the case of ET data this involves tens of thousands of checks. With such a large number and all the variability of EEG data, many of them will not pass the tests. The question is how one should behave with them. Should the non-gaussian descriptors be excluded from the database? Even ignoring this problem, we will be left with a normative database in which accuracy is different for each descriptor. In fact we shown that the accuracy is a function of skewness and each approximation to gaussianity will lead to different skewness levels. Furthermore, the outcome of the normative database will be different on the two sides of the distribution. These are not desirable characteristics for a normative database. One may jump over all these problems using a non-parametric approach, given that the sample size is large enough. It is fortunate that normative databases of clinical usefulness are constructed on the basis of large samples. Actually the sample sizes commonly employed are so large (500-600) that they would lead to more then 96% accuracy if the non-parametric method described in this thesis would be employed. Furthermore, the validity of results would be the same for the right-handed and the left-handed test, for all electrodes, frequency band and for whatever measure employed, regardless of its distribution. In fact the sample size is the only true constant across descriptors and we have shown in this thesis that the accuracy of the non-parametric method, given a fixed alpha, depends solely on sample size. This is a distinct advantage of the non-parametric method. In addition the extension of the method to ET data and to new electroencephalographic measures is straightforward. It should be pointed out also that developmental equations and other kind of between-subject differences can be taken into account while compiling a non-parametric normative database. For instance, polynomial regression equations based on age can be computed. Each descriptor value can be normalized over its regression predicted value to remove any unwanted trend in the data. Back in the 1970's it was not easy to perform a non-parametric test. Computers were slow and the computations required could take hours. Today they would take a minute. Another possible reason why non-parametric methods have not been employed yet is that they require more intense computer programming. However one does not have to check data gaussianity, does not have to transform the data and does not have to be concerned about the distribution of new measures any longer. In perspective, by using a non-parametric method one would actually save time.

References

Ahn, H., Prichep, L. S., John, E. R., Baird, H., Trepetin, and Kaye, H. (1980); Developmental Equations reflect Brain Dysfunctions. *Science,* **210**, 1259-1262.

Beauchamp. K. G., (1973); Signal Processing Using Analog and Digital Techniques. *George Allen & Unwin,* London

Brillinger, D.R, (1975); Time Series. Data Analysis and theory. *Holt, Rinehart, and Winston,* New York.

Bosch-Bayard, J., Valdés-Sosa, P., Virues-Alba, T., Aubert-Vázquez, E., John, E.R., Harmony, T., Riera-Diaz, J., and Trujillo-Barreto, N. (2001); 3D Statistical Parametric Mapping of EEG Source Spectra by Means of Variable Resolution Electromagnetic Tomography (VARETA). *Clinical Electroencephalography,* **32**, 47-61.

Duffy, F.H., Bartels, P.H., and Burchfiel, J.L. (1981); Significance Probability Mapping: An Aid in the Topographic Analysis of Brain Electrical Activity. *Electroencephalography and clinical Neurophysiology,* **51**, 455-462.

Fuchs, M., Wagner, M., Köhler, T., & Wischmann, H. A., (1999); Linear and Nonlinear Current Density Reconstructions. *Journal of Clinical Neurophysiology,* **16**, 267-295.

Gasser, T., Bächer, P., and Möcks, J. (1982) Transformation towards the normal distribution of broad band spectral parameters of the EEG. *Electroencephalography and Clinical Neurophysiology.* **53**, 119-124.

Hernández, J.L., Valdés, P., Biscay, R., Virues, T., Szava, S., Bosch, J., Riquenes, A., and Clark, I. (1994); A Global Scale Factor in Brain. *International Journal of Neorosciences,* **76**, 267-278.

Holmes, A.P., Blair, R.C., Watson, J.D.G., and Ford, I. (1996); Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood-flow Metabolism,* **16**, 7-22.

Hughes, J.R., and John, E.R. (1999); Conventional and Quantitative Electroencephalography in Psychiatry. *Journal of Neuropsychiatry and Clinical Neuroscience,* **11**, 190-208.

John, E. R., Ahn, H., Prichep, L. S., Trepetin, M., Brown, D., Kaye, H. (1980); Developmental Equations for the Electroencephalogram *Science,* **210**, 1255-1258.

John, E. R., Karmel, B.Z., Corning, W.C., Easton, P., Brown, D., Ahn, H., John, M., Harmony, T., Prichep, L. S., Toro, A., Gerson, I., Bartlett, F., Thatcher, R., Kaye, H., Valdes, P., and Schwartz, E. (1980); Neurometrics *Science,* **196**, 1393-1409.

John, E. R., Prichep, L. S., and Easton, P. (1987); Normative Data Banks and Neurometrics. Basic Concepts, Method and Results of Norm Constructions.in *Method of Analysis of brain Electrical and Magnetic Signals. EEG Handbook (revised series. Vol. 1).* (Gevins, A. S., and Remond, A. Ed.). Elsevier Science Publishers B.V. (Biomedical Division).

John, E. R., Prichep, L. S., Fridman, J., and Easton, P., (1988); Neurometrics: Computer Assisted Differential Diagnosis of Brain Dysfunctions. *Science,* **239**, 162-169.

Lipschutz, S., and Lipson, M.L. (2000); Probability. 2nd Ed. Schaum's Outline Series, *McGraw-Hill,* New York.

Lynn, P. A., and Fuerst, W., (1989); Introductory Digital Signal Processing with Computer Applications. *John Wiley & Sons,* New York.

Nunez, P. L., Srinivasan, R., Westdorp, A. F.,Wijesinghe, R. S., Tucker, D. M., Silberstein, and R. B., Cadusch, P. J., (1997); EEG Coherency I: statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales. *Electroencephalography and clinical Neurophysiology,* **103**, 499-515.

Nuwer, M.R., (1988); Quantitative #EEG: II. Frequency Analysis and Topographic Mapping in Clinical Settings. *Journal of Clinical Neurophysiology,* **5**, 45-85.

Oken, B.S., and Chiappa, K.H. (1988); Short-term variability in EEG frequency analysis. *Electroencephalography and Clinical Neurophysiology,* **69**, 191-198.

Pascual-Marqui, R. D., (1995); Reply to comments by Hämäläinen, Ilmoniemi and Nunez. In Source Localization: Continuing Discussion of the Inverse Problem (W. Skrandies Ed.). *ISBET Newsletter,* **6**, 16-28.

Pascual-Marqui, R. D., (1999); Review of Methods for Solving the EEG Inverse Problem. *International Journal of Bioelectromagnetism,* **1**, 75-86.

Pascual-Marqui, R. D., Michel, C. M., Lehmann, D. (1994); Low Resolution Electromagnetic Tomography: a new Method for Localizing Electrical Activity in the Brain. *International Journal of Psychophysiology,* **18**, 49-65.

Pollock, V.E., Schneider, L.S., and Lyness, S.A. (1990) EEG Amplitude in healthy, late-middle-aged and elderly adults: normality of the distributions and correlation with age. *Electroencephalography and Clinical Neurophysiology,* **75**, 276-288.

Prichep, L.S., John, E.R., and Tom, M. (2001); Localization of Deep White Matter Lymphoma using VARETA: A Case Study. *Clinical Electroencephalography.* 32(2), 62-65.

Swets, J.A. (1988). Measuring the Accuracy of Diagnostic Systems. *Science,* **240**, 1285-1293.

Swets, J.A., and Pickett, R.M. (1982); Evaluation of Diagnostic Systems. Methods from Signal Detection Theory. *Academic Press,* New York.

Szava, S., Valdes, P., Biscay, R., Galan, L., Bosch, J., Clark, I., and Jimenez, J.C. (1994); High Resolution Quantitative EEG Analysis. *Brain Topography,* **6**, 211, 219.

Talairach, J, and Tournoux, P. (1988); Co-planar Stereotaxic atlas of the Human Brain. *Thieme Medical Publishers,* New York.

Thatcher, R.W. (1999). EEG Database-Guided Neurotherapy. In: Quantitative EEG and Neurofeedback. Ed: Evans, J.R., and Abarbanel, A., Academic Press. San Diego, London.

Veldhuizen, R.J., Jonkman, E.J., and Poortvliet, D.C.J. (1993); Sex Differences in age regression parameters of healthy adults-normative data and practical implications. *Electroencephalography and Clinical Neurophysiology,* **86**, 377-384.

Appendix

Table 1: Two-by-two contingency table. Schematic results summarizing the outcome of an experiment testing the accuracy of a diagnostic system.

		Event (E)	[input]
		Positive	Negative
Diagnosis (D)	Positive	TRUE POSITIVE	FALSE POSITIVE
[Output]	Negative	FALSE NEGATIVE	TRUE NEGATIVE

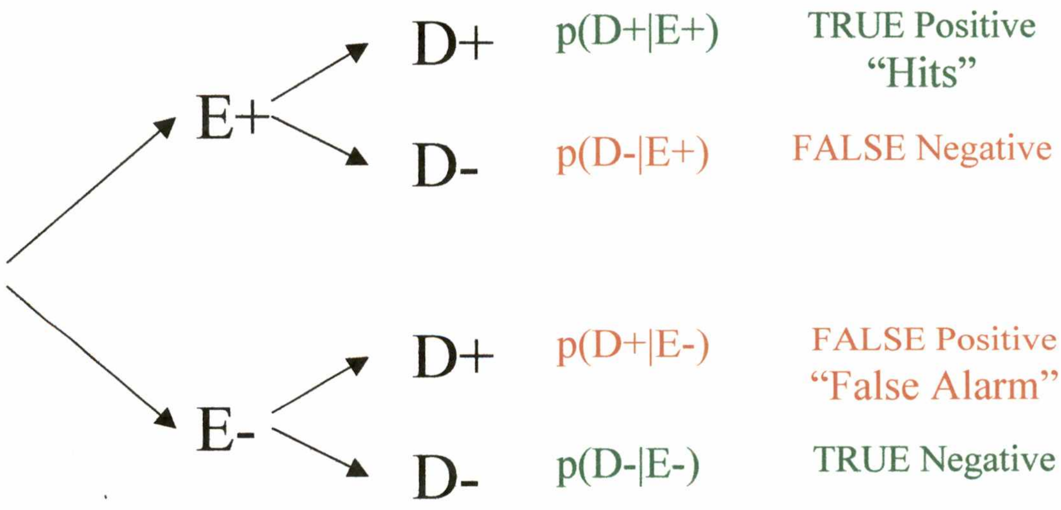


Figure 1: Probability Tree. The same data summarized in table 1 can be arranged, after normalization, in a probability tree. The tree shows the resulting conditional probabilities.

Table2: Table of Normative Database Steps. Description of steps required in order to build a Normative Database in the case of the parametric method and of the non-parametric method.

Steps	Parametric Method	Non-Parametric Method
1	A reference Population (usually normal) is defined and a sample of **N** subjects is selected. Each subject is screened in order to match inclusion criteria previously chosen. The **N** subjects constitute the database.	A reference Population (usually normal) is defined and a sample of **N** subjects is selected. Each subject is screened in order to match inclusion criteria previously chosen. The **N** subjects constitute the database.
2	The set of **F** features is defined. Each feature refers to a quantitative measure for a particular frequency range. For example, a feature could be "Delta Relative Power" or " Alpha Coherence".	The set of **F** features is defined. Each feature refers to a quantitative measure for a particular frequency range. For example, a feature could be "Delta Relative Power" or " Alpha Coherence".
3	For each of the **N** subject constituting the database, for each location (electrode or voxel) or pair of locations (electrodes or voxels), **L** measures for each of the chosen set of **F** features are derived. Each combination of measure and Feature is called **Descriptor**.	For each of the **N** subject constituting the database, for each location (electrode or voxel) or pair of locations (electrodes or voxels), **L** measures for each of the chosen set of **F** features are derived. Each combination of measure and Feature is called **Descriptor**.
4	Database Data form a **L** x **F** x **N** matrix	Database Data form a **L** x **F** x **N** matrix
5	For each feature, an appropriate transformation (based on log) is applied to all locations and subjects in order to approximate gaussianity.	For each feature and location the **N** data of the database subjects is sorted.
6	For each new individual to be compared to the database, a corresponding data matrix of descriptors (**LxF**) is derived.	For each new individual to be compared to the database, a corresponding data matrix of descriptors (**LxF**) is derived.
7	For each location (**L**) and feature (**F**), i.e., for each descriptor, the deviation from normality is expressed in terms of z-scores, using the mean and standard deviation of the descriptor computed for all database subjects.	For each location (**L**) and feature (**F**), i.e, for each descriptor, the deviation from normality is expressed in terms of discrete random variable sp (sample proportion) expressing the proportion of the subjects in the database falling above (right-handed test) or below (left-handed test) the new individual.
8	Additional statistics are performed in order to correct for multiple comparisons.	Additional statistics are performed in order to correct for multiple comparisons.

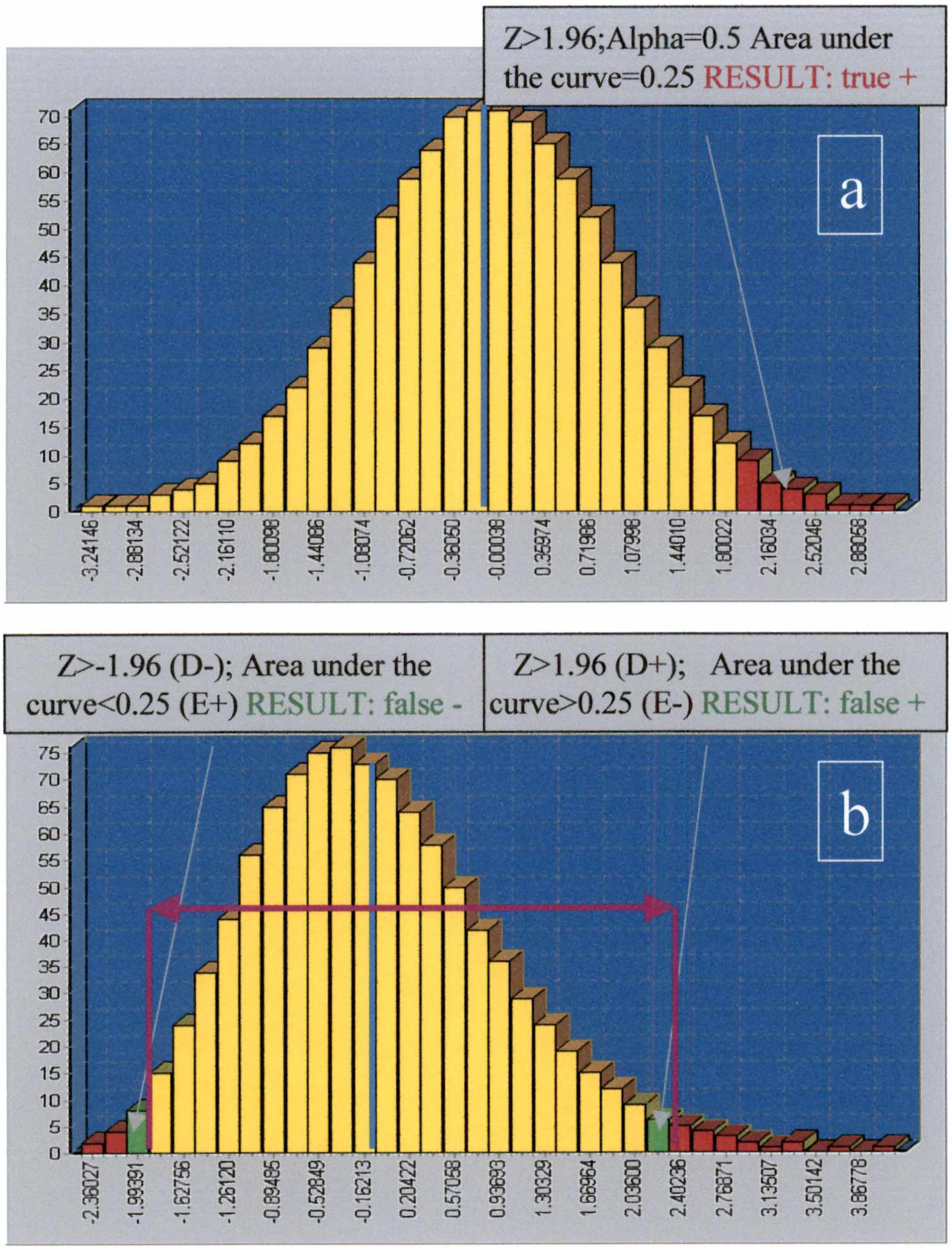


Figure 2: Depiction of gaussianity and non-gaussianity. 2a: The normality case. If the sample is truly gaussian then the outcome of the normative database leads to true positives and true negatives only. 2b: The non-normality case. The sample distribution is right skewed. On the right side of the distribution we have false positives, while on the left-side of the distribution we have false negative. See text for details.

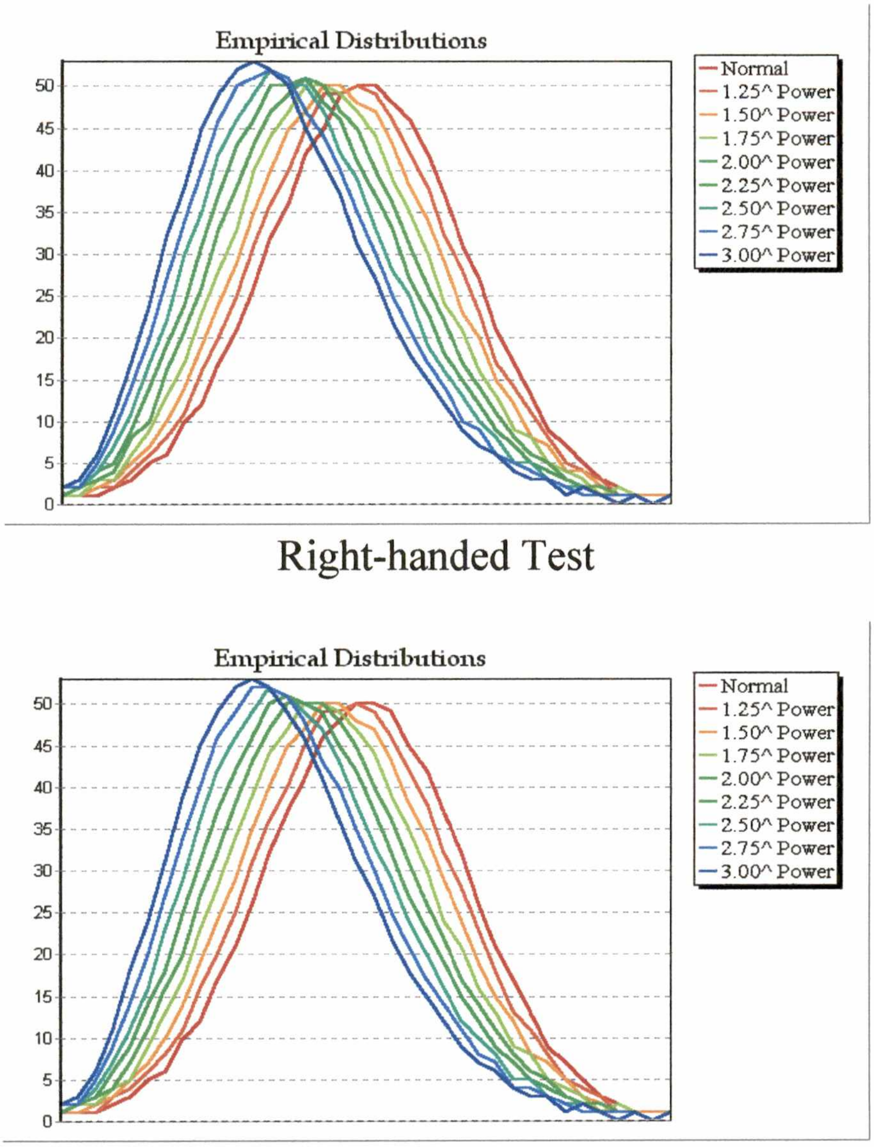
Table 3: Skewness and kurtosis table. Means and standard deviations (across sample size) of the skewness and kurtosis of the empirical distributions employed in the simulations for both the right-handed test and the left-handed test.

Distribution	Mean Sk	Sd Sk	Mean Kt	Sd Kt
Pw of1.00	0.000	0.002	2.869	0.085
Pw of1.25	0.070	0.004	2.870	0.082
Pw of1.50	0.140	0.006	2.884	0.084
Pw of1.75	0.209	0.009	2.916	0.091
Pw of2.00	0.279	0.010	2.962	0.093
Pw of2.25	0.347	0.015	3.022	0.103
Pw of2.50	0.417	0.017	3.100	0.113
Pw of2.75	0.486	0.021	3.192	0.124
Pw of3.00	0.556	0.024	3.299	0.137

Right-handed test

Distribution	Mean Sk	Sd Sk	Mean Kt	Sd Kt
Pw of1.00	0.000	0.001	2.869	0.085
Pw of1.25	0.070	0.004	2.870	0.083
Pw of1.50	0.140	0.006	2.886	0.086
Pw of1.75	0.210	0.008	2.916	0.089
Pw of2.00	0.279	0.012	2.962	0.095
Pw of2.25	0.348	0.015	3.026	0.102
Pw of2.50	0.418	0.017	3.104	0.111
Pw of2.75	0.486	0.021	3.193	0.125
Pw of3.00	0.555	0.025	3.298	0.141

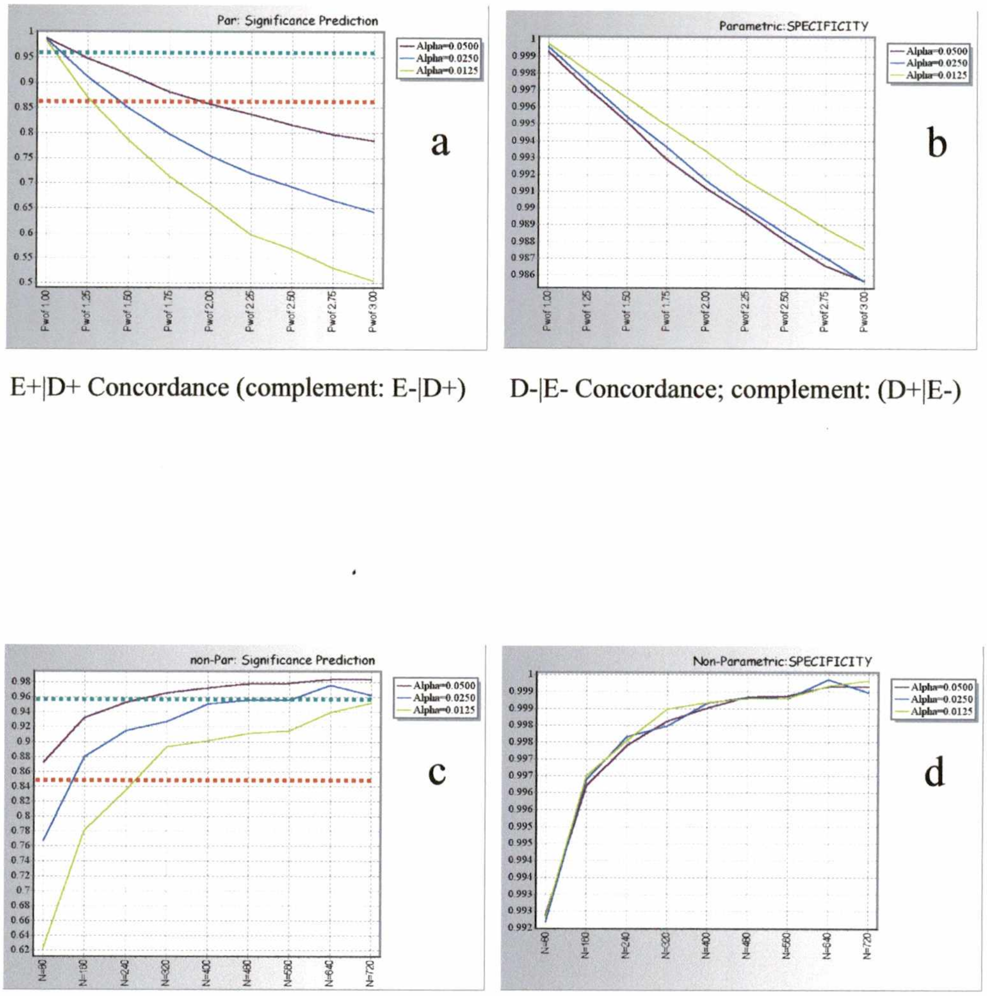
Left-handed test



Right-handed Test

Left-handed Test

Figure 3: Empirical distributions. Frequency polygons of the empirical distributions employed in the right-handed and left-handed test simulations for N=720. Shown are the nine distributions corresponding to the nine kinds of power transformation used to vary gaussianity. The displacement along the x-axis due to the power transformation has been removed in these pictures.



E+|D+ Concordance (complement: E-|D+)   D-|E- Concordance; complement: (D+|E-)

E+|D+ Concordance (complement: E-|D+)   D-|E- Concordance; complement: (D+|E-)

Figure 4: Results of the simulations for the RIGHT-HANDED test. Reported are the proportion of true acceptance (a) and specificity (b) for the parametric method, and the proportion of true acceptance (c) and specificity (d) for the non-parametric method. For the parametric method results are shown as a function of non-gaussianity of the normative reference distribution and alpha level. For the non-parametric method results are shown as a function of sample size and alpha level. The green line indicates where the measure of accuracy is equal to 0.95 (very good level of accuracy). The red line indicates where the measure of accuracy is equal to 0.85 (acceptable level of accuracy).

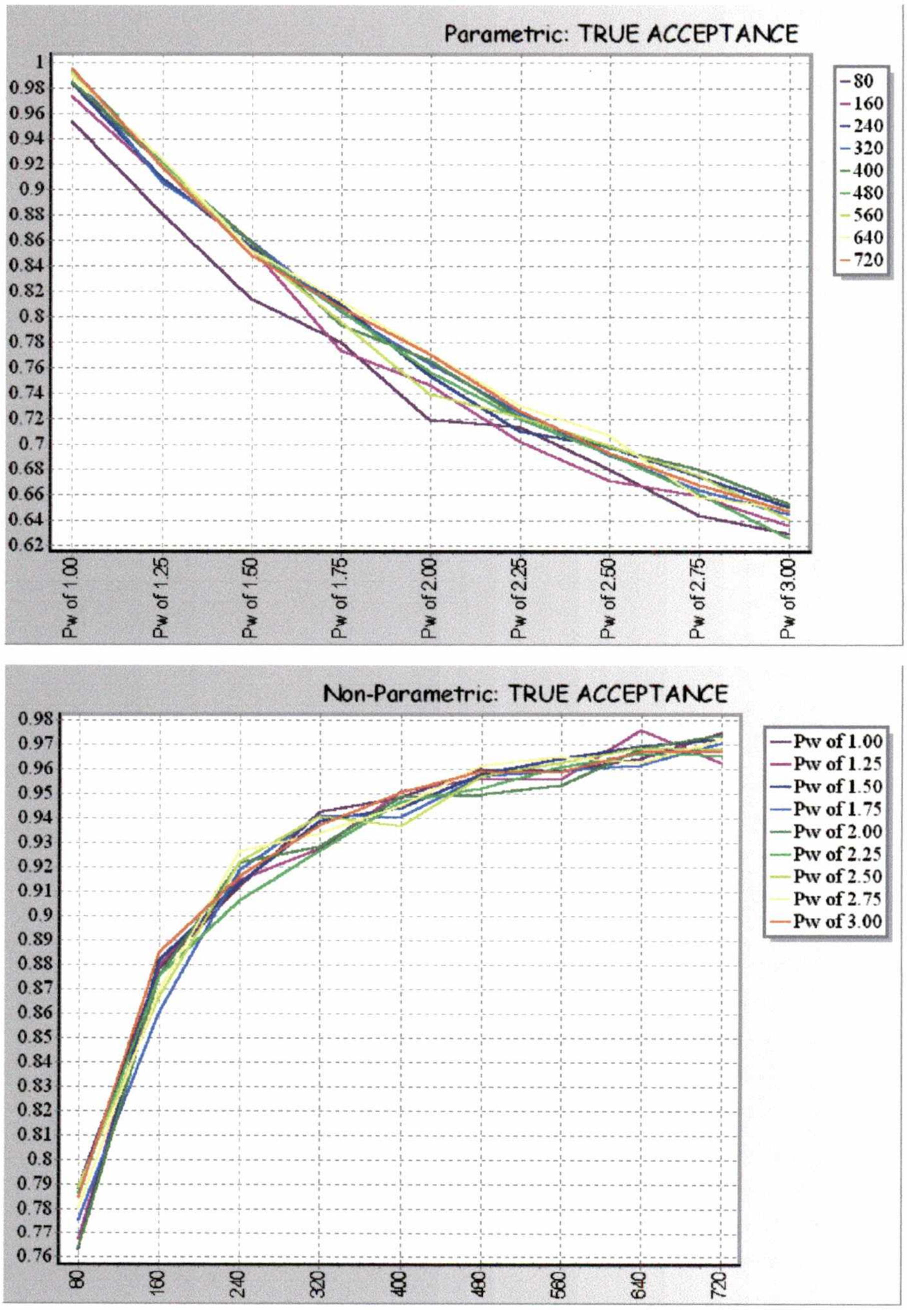
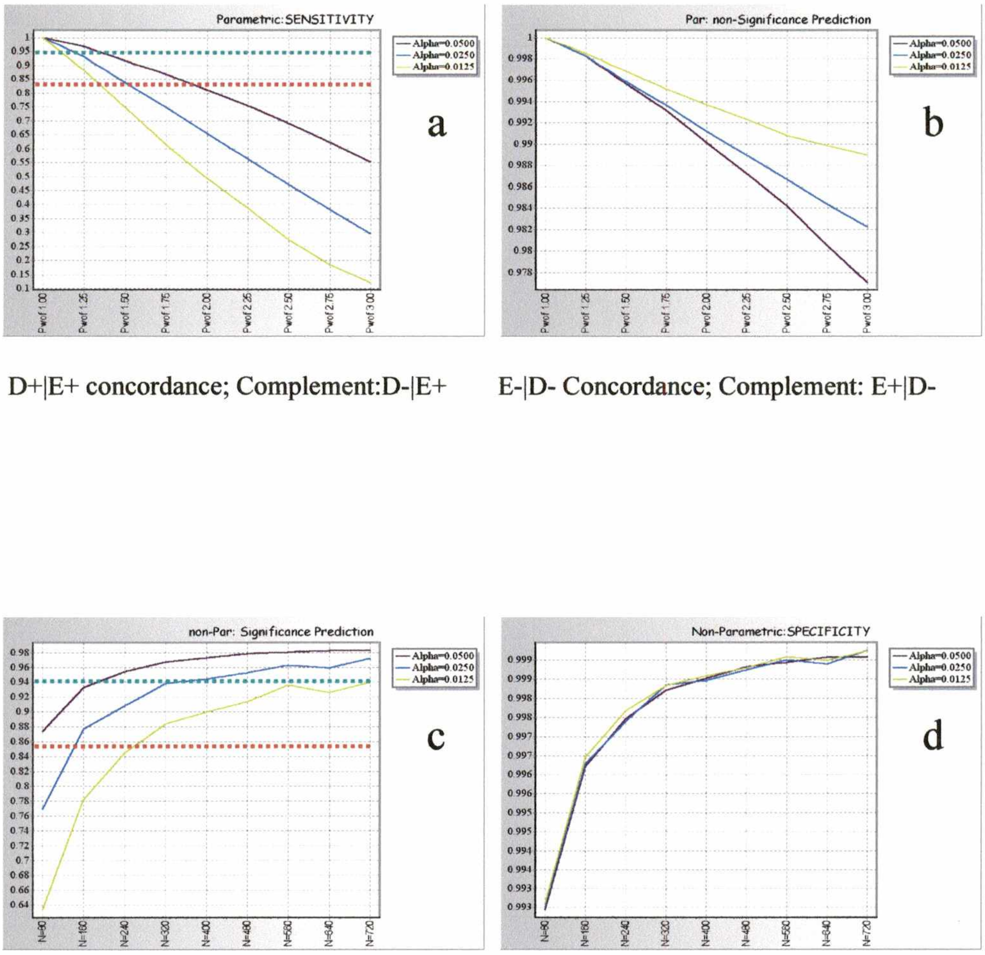


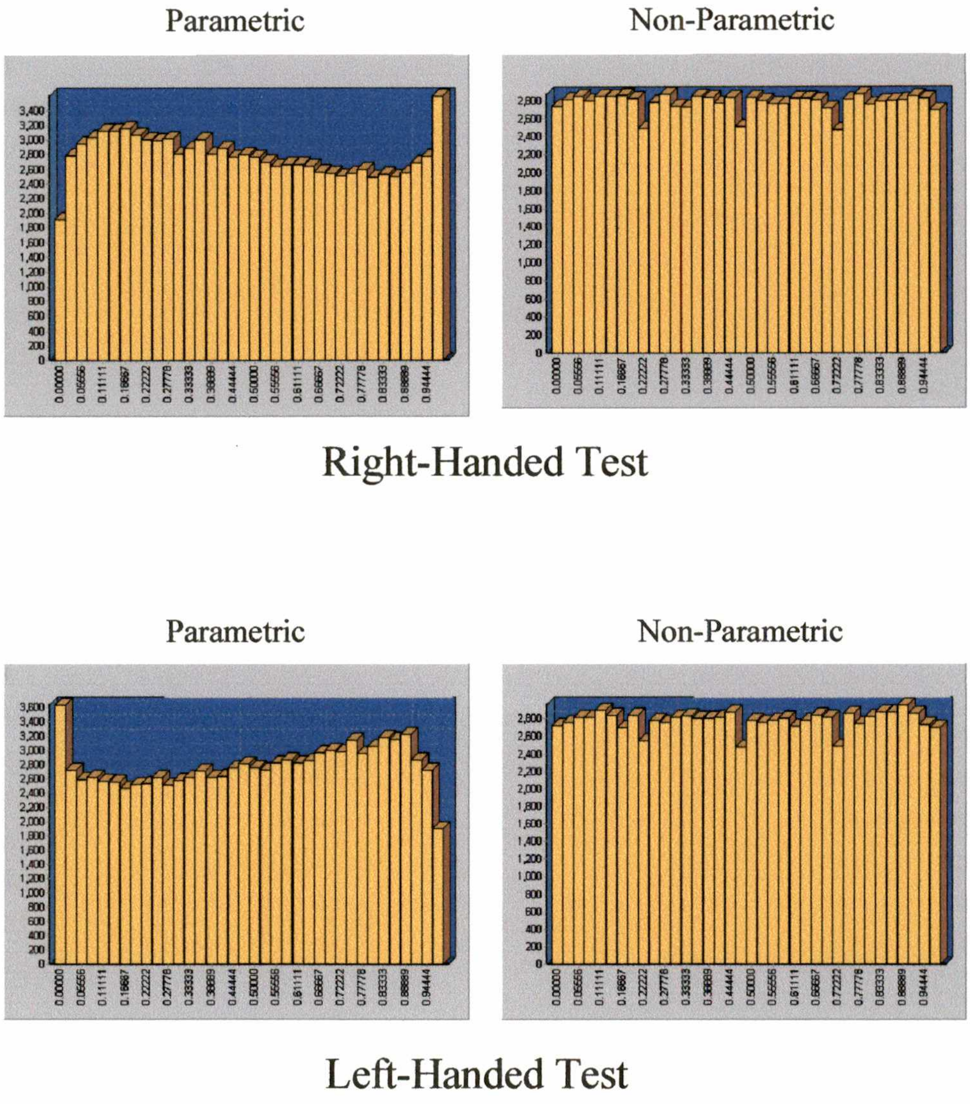
Figure 5: Uncollapsed data for the RIGHT-HANDED test with alpha=0.025. Top: critical test for the parametric method. Bottom: critical test for the non-parametric method. These graphs show the rationale for collapsing data across sample size levels in the case of the parametric method, and across gaussianity levels for the non-parametric data.



D+|E+ concordance; Complement:D-|E+ E-|D- Concordance; Complement: E+|D-

E+|D+ Concordance (complement: E-|D+) D-|E- Concordance; complement: (D+|E-)

Figure 6: Results of the simulations for the LEFT-HANDED test. Reported are the proportion of sensitivity (a) and true rejection (b) for the parametric method, and the proportion of true acceptance (c) and specificity (d) for the non-parametric method. For the parametric method results are shown as a function of non-gaussianity of the normative reference distribution and alpha level. For the non-parametric method results are shown as a function of sample size and alpha level. The green line indicates where the measure of accuracy is equal to 0.95 (very good level of accuracy). The red line indicates where the measure of accuracy is equal to 0.85 (acceptable level of accuracy).



Parametric

Non-Parametric

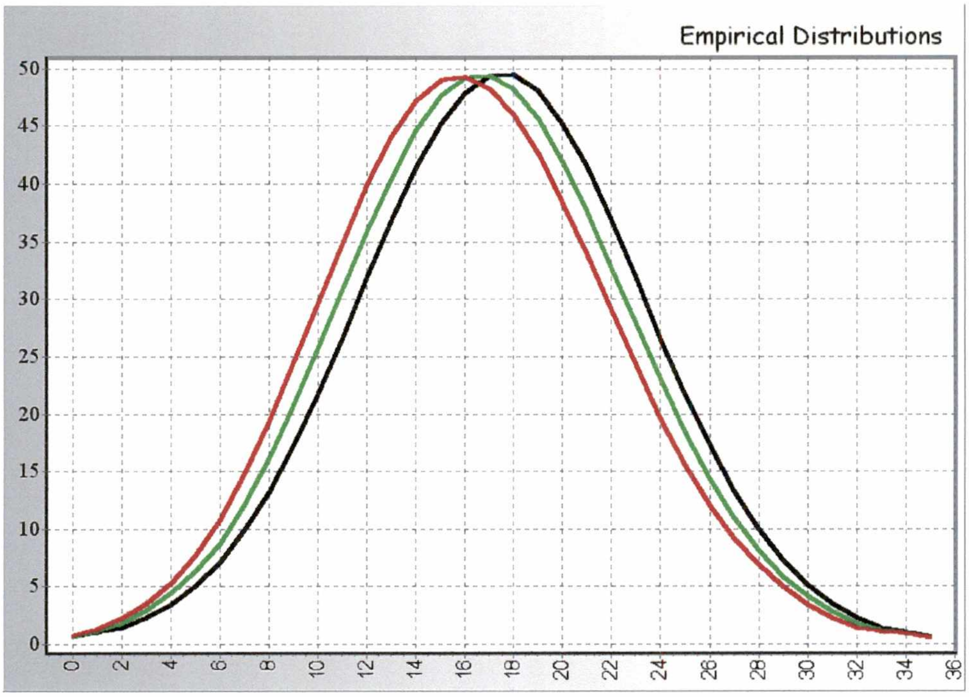
Right-Handed Test

Parametric

Non-Parametric

Left-Handed Test

Figure 7: Histograms of p-values . p-values as computed by the parametric method and the non-parametric method in the case of the right-handed and left-handed test. All histograms are based on 100.000 simulation entries, with N=320, power transformation=2, and alpha=0.0125. Since the normative reference sample is skewed, the histogram of the p-values for the parametric method is not uniform.



⬛ Empirical normal distribution smoothed with a Hanning routine. N=720, skewness=0, kurtosis=2.94

🟩 Empirical normal distribution power (1.25) transformed and smoothed with a Hanning routine. N=720, skewness=0.07, kurtosis=2.94

🟥 Empirical normal distribution power (1.5) transformed and smoothed with a Hanning routine. N=720, skewness=0.15, kurtosis=2.95

Figure 8: Smoothed distributions. Typical distributions leading to an acceptable error rate (<0.15) in the case of the parametric method. The black distribution is a true normal empirical distribution, while the green and the red distributions have been obtained elevating to the power each sample (the exponent was 1.25, and 1.5 respectively). Notice the low value of skewness for the red distribution. A value bigger than that leads to an unacceptable error rate (>0.15) for the parametric method.

Vita

Marco Congedo was born in Bari, Italy, on the 20th of October 2001. He attended the University of Padua (Padova, Italy), where in 1998 he received the B.A. degree in Psychology. In 1996-97 he studied at the University René Descartes, Paris, France. He speaks fluently in English and French, in addition to his mother-tongue (Italian). Marco Congedo is currently enrolled in a PhD program in Biological Psychology at the University of Tennessee, Knoxville.