# A Priori Sample Size Determination for the Number of Subjects in an EEG Experiment

E. Guttmann-Flury, X. Sheng, D. Zhang, and X. Zhu, *Member, IEEE*

***Abstract*— This paper represents a first attempt to perform a priori sample size determination from a "historic" Electroencephalography (EEG) dataset. The importance of adequate sample size is firstly highlighted, and evidence is given against the use of normal distribution for such computations, when the data cannot be assumed to be Gaussian. The "historic" dataset is then thoroughly examined to determine the least less likely underlying distribution for the desired phenomenon, in this case the spontaneous blinks potential. Two Monte Carlo simulations, using different distribution assumptions, are subsequently computed to estimate the a priori minimum sample size. Finally, these choices are discussed considering practical limitations, as well as the computational differences for other phenomena to study.**

## I. Introduction

Brain research usually studies the structure and function of the brain by investigating broad questions such as "what brain regions are associated with the state of depression and stress?" [1] or "can waveforms associated with performance monitoring (Error-related negativity, …) be used as a diagnosis for anxiety disorder or major depressive disorder?" [2]. One of the main pillars for scientific rigor in biomedical research is the redundancy in experimental design, which enhances the likelihood of results reproducibility [3]. To this end, sample size calculations are the first step to ensure non-biased and reproducible effects [4]. Moreover, a priori sample size calculation can be conducted to reduce the risk of an underpowered (false-negative) result [5].

Despite the importance of adequate sample sizes, a literature review focusing on human Electroencephalography (EEG) and Event-related potential (ERP) showed that 0% of the selected studies reported sample size calculations [4]. The authors conjecture that one reason could be the uncertainty on how to conduct such calculations and the difficulty of finding the necessary information to report the computations. Furthermore, in the scarce non-EEG medical studies performing a sample size calculation, the underlying hypothesis considers the distribution as Gaussian for this computation, even when the data are to be analyzed by generalized linear models afterwards [6]. Hence, this approach is "logically inconsistent and might yield a large magnitude of error, considering the discrepancies between the cumulative distribution functions (CDF) between the normal approximation and the exact distributions" [6].

This paper is an attempt to estimate the minimum number of subjects for an EEG experiment, in which the observed phenomenon is the effect of involuntary eyelid movements (i.e. spontaneous blinks) in the signal. Consequently, the observed variable is defined as the maximum potential (in $\mu V$) on a reference channel (FP1) during an involuntary eyelid movement. From now on, let's define $FP1\_Max$ as the observed variable. A priori sample size computation is conducted on a historical dataset to estimate the minimum number of subjects for a similar EEG experiment.

## II. Dataset Description and Descriptive Statistics

### A. Dataset

The online-available dataset from [7] provides brain activity from 26 healthy subjects using the P300-speller as a Brain-Computer Interface (BCI) paradigm. With the goal of restoring communication in locked-in patients, spelling is achieved by only paying attention to rare visual stimuli, eliciting the P300 response. The subjects had to go through five P300 spelling sessions. Each session consisted of twelve 5-letter words, except the fifth which consisted of twenty 5-letter words. Thus, through all sessions the BCI task remained identical. The recording time was similar for the first four sessions, and longer for the fifth one, reaching a total recording time of approximately 1 hour per subject.

### B. Blinks Detection

The blinks detection algorithm used to extract $FP1\_Max$ is not the focus of this paper. Very briefly, if the data satisfies three criteria (amplitude, pre-&post- amplitude, and propagation), a blink is extracted (refer to [8] for more details). For each blink, $FP1\_Max$ is then computed as the maximum potential on FP1.

### C. Descriptive Statistics

Contrarily to [7], the variable of interest is $FP1\_Max$. Hence all following statistical description will focus on the blinks distribution. This does not affect the generality of this study, since participants would still spontaneously blink throughout the sessions whatever BCI paradigm is chosen.

Heuristically [8], all blinks are included in a $[40\ 450]\ \mu V$ interval. These choices have been made considering that a $FP1\_Max$ lower than $40\ \mu V$ is likely to be caused by another phenomenon than a blink, while a $FP1\_Max$ higher than $360\ \mu V$ is probably caused by a mixed of events, for example two successive blinks or a blink accompanied by an eye movement. Though blinks between $[360\ 450]\ \mu V$ are not "lone" blink per se, they are still kept in this analysis as they can occur frequently for some subject.

* This work was supported by in part the National Science Foundation of China (No.51620105002), and the Science and Technology Commission of Shanghai Municipality (No.17JC1402700).

E. Guttmann-Flury, X. Sheng, D. Zhang and X. Zhu are with the State Key Laboratory of Mechanical System and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Minhang District, Shanghai, 200240, P. R. China. (Corresponding e-mail: mexyzhu@sjtu.edu.cn).

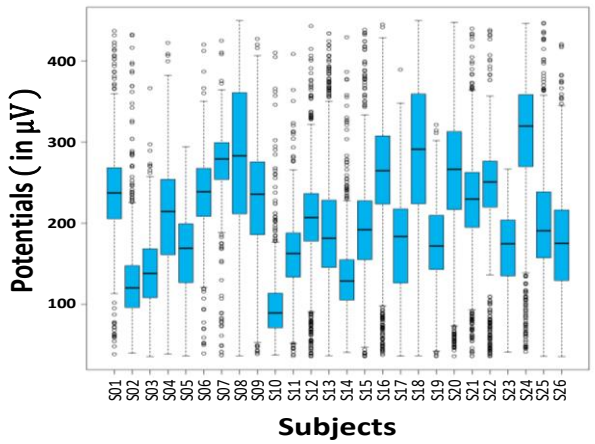


Figure 1. Boxplot blinks distribution for 26 subjects

As Fig. 1 shows, the blinks distribution varies greatly from one subject to another. Across subjects, the average of $FP1\_Max$ can be as low as $\sim 100\,\mu V$ and as high as $\sim 300\,\mu V$. Similarly, the total number of blinks is also highly subject-dependent. On the other hand, for one specific subject, the blinks distribution is quite similar from one session to another. The average of $FP1\_Max$ varies slightly between sessions ($\sim 240\,\mu V \pm 20\,\mu V$ for Subject 1), as shown in Fig. 2.


Figure 2. Blinks Distribution differences in 5 sessions for Subject 1

### *D. Fitting Blinks Distribution to Common Distribution*

To simplify, blinks distribution is hence assumed to be similar across sessions for each subject. With this hypothesis, blinks distribution per subject are fitted to existing distributions. Seven common distributions are chosen for this analysis: Normal, Logistic, Lognormal, Gamma, Weibull, Cauchy and Gumbel.

For each subject, the Kolmogorov-Smirnov (KS), Cramer von Mises (CVM) and Anderson-Darling (AD) statistics are computed. The KS distance is defined to be the largest absolute difference between the blinks' distribution CDF and the hypothesized underlying distribution CDF evaluated at any point. The CM distance is obtained by taking the square root of the sum of the squared difference between the two distributions. Finally, the AD test uses the fact that if the data arises from the hypothesized distribution, the CDF can be assumed to follow a uniform distribution. These three tests give different results as they exhibit varying degrees of test bias depending on the situation. For example, AD test has better power against fatter tails, while KS has more power against deviations to central values.

To complete this analysis, Akaike information criterion (AIC) and Bayesian information criterion (BIC) are also computed. AIC estimates the relative amount of information lost by a given model and thus deals with the trade-off between the goodness of fit and the simplicity of the model. BIC is based, in part, on the likelihood function, and is closely related to AIC. Both AIC and BIC try to avoid overfitting, but do not provide a test of a model in the sense of testing a null hypothesis. They merely give intel on the relative quality to other models.

All these tests cannot find the "best" distribution. On the contrary, by considering all these statistics, one can infer which distribution is the least less likely. Visual inspection of the Q-Q and P-P plots complete this analysis and give an informed choice on the fitted distribution. Table I summarizes the calculated and chosen distributions per subject.

TABLE I. FITTED DISTRIBUTION PER SUBJECT

| Subject | CVM | AD | KS | AIC | BIC | Choice |
|---|---|---|---|---|---|---|
| S01 | No distribution | No distribution | Logistic | Logistic | Logistic | Logistic |
| S02 | No distribution | No distribution | No distribution | Lognormal | Lognormal | Logistic |
| S03 | Logistic | Logistic | Normal, Logistic, Gamma, Weibull | Logistic | Logistic | Logistic |
| S04 | No distribution | No distribution | No distribution | Weibull | Weibull | Logistic |
| S05 | No distribution | No distribution | No distribution | Weibull | Weibull | Normal |
| S06 | No distribution | No distribution | Logistic, Cauchy | Logistic | Logistic | Logistic |
| S07 | No distribution | No distribution | No distribution | Logistic | Logistic | Logistic |
| S08 | No distribution | No distribution | No distribution | Normal | Normal | Logistic |
| S09 | Logistic | No distribution | Normal, Logistic | Logistic | Logistic | Logistic |
| S10 | No distribution | No distribution | No distribution | Lognormal | Lognormal | Gumbel |
| S11 | No distribution | No distribution | Logistic | Logistic | Logistic | Logistic |
| S12 | No distribution | No distribution | No distribution | Logistic | Logistic | Logistic |
| S13 | No distribution | No distribution | Lognormal, Gumbel | Gumbel | Gumbel | Gamma |
| S14 | No distribution | No distribution | Logistic | Gumbel | Gumbel | Gumbel |
| S15 | No distribution | No distribution | No distribution | Gumbel | Gumbel | Logistic |
| S16 | No distribution | No distribution | No distribution | Logistic | Logistic | Logistic |
| S17 | No distribution | No distribution | No distribution | Normal | Normal | Logistic |
| S18 | No distribution | No distribution | No distribution | Normal | Normal | Normal |
| S19 | No distribution | No distribution | Normal, Logistic, Weibull | Weibull | Weibull | Normal |
| S20 | No distribution | No distribution | No distribution | Weibull | Weibull | Weibull |
| S21 | No distribution | No distribution | Logistic | Logistic | Logistic | Logistic |
| S22 | No distribution | No distribution | No distribution | Cauchy | Cauchy | Logistic |
| S23 | No distribution | No distribution | Logistic, Weibull | Weibull | Weibull | Logistic |
| S24 | No distribution | No distribution | No distribution | Logistic | Logistic | Logistic |
| S25 | No distribution | No distribution | No distribution | Gumbel | Gumbel | Logistic |
| S26 | Logistic | No distribution | Normal, Logistic, Weibull | Weibull | Weibull | Logistic |

By definition, AIC and BIC always give a result, since they estimate the relative quality of statistical models. The preferred model is merely the one with the minimum value. On the other hand, CVM, AD and KS rely on testing a null hypothesis. For a candidate model, if the P-value is "small", the null hypothesis will be rejected, meaning that the test failed to show that the considered distribution follows the candidate distribution. This the reason why different tests are computed along visual inspection to provide an informed choice.

Fig. 3 shows the logistic and empirical distribution for subject S01, as well as the CDFs, Q-Q plot and P-P plot, using the fitdistrplus library available in the Comprehensive R Archive Network [9].

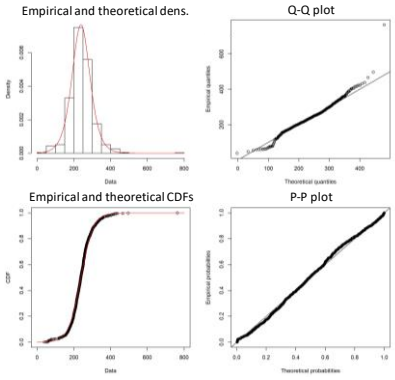


Figure 3. Visual inspection for S01 with the choice of the logistic distribution

The observed variable $FP1\_Max$ is the result of spontaneous blinks on the EEG. Considering the heuristic choices stated previously, $FP1\_Max$ cannot be smaller than $40\ \mu V$, which can explain the left tail shape of the distribution. Furthermore, let's assume hereafter that this variable value is related to the total extent of travel of the upper eyelid. [10] demonstrated that most spontaneous blinks in most subjects are not complete, in order to minimize the time that vision is obscured. Consequently, the highest values correspond to the complete blinks. This might explain why the logistic distribution is the most common one, since it displays heavier tails compared to a normal distribution.

However, the sum of logistic random variables does not have a logistic distribution. If a logistic distribution is approximated as a normal distribution, then since the sum of normal random variables is normal, the sum can be considered approximately logistic. Thus, the overall blinks distribution, i.e. the blinks distribution of all subjects, might be approximated as a logistic distribution. Another choice would be to directly fit the overall blinks distribution to the common distributions stated beforehand, which yields a Weibull distribution.

## III. Sample Size Determination

The previous analyses have shown that neither the blinks distribution per subject nor the overall blinks distribution can be assumed to be Gaussian. Hence, it would be incorrect to directly use the Central Limit Theorem. Instead, Monte Carlo (MC) simulation can generate samples from the empirical historical blinks distribution and calculate the sample size required to reach a given power for a new EEG experiment.

### *A. Monte Carlo simulation for Power Analysis*

With a sufficient number of subjects, the overall blinks distribution for a new EEG experiment should be similar to the one described previously. In other words, no difference should be found between the historic experiment and the new one. Hence, the hypothesis to be tested is:

$$\begin{cases} null\ hypothesis\ (H_0): & \mu = \hat{\mu} \\ alternative\ hypothesis\ (H_a): & \mu \neq \hat{\mu} \end{cases} \quad (1)$$

The Type I error rate corresponds to the situation where the null hypothesis ($H_0$) is true but rejected. Hereafter, the Type I error rate (or significance level) is fixed at $\alpha = 0.05$. The Type II error rate (null hypothesis is false, but erroneously fails to be rejected) is fixed at $\beta = 0.2$, leading to a power of 0.80. This power would indicate that one in five times a difference that is present in the data is not found. Furthermore, the relative seriousness of type I to type II error is 4 to 1. Thus, mistaken rejection of the null hypothesis is considered 4 times as serious as mistaken acceptance. Finally, Cohen's effect size, or degree of departure from the null hypothesis, is defined as [10]:

$$d = \frac{|\, m_A - m_B \,|}{\sigma} \quad (2)$$

Where $m_A$ and $m_B$ are the population means expressed in raw unit (μV) and $\sigma$ is the standard deviation of either population (since they are assumed equal). At this point, one can either use the effect size suggested by Cohen [10] (Small: $d = 0.20$ ; Medium: $d = 0.50$ ; Large: $d = 0.80$) or run the new EEG experiment on a few subjects to calculate the approximate effect size.

### *B. Method 1: Using the Normal Distribution*

In this paragraph, the overall blinks distribution is assumed to be Gaussian. As discussed previously, this is an incorrect assumption. However, since the Gaussian distribution is the most well know one, the sample size calculation is easy and can provide some insight. Fig. 4 shows the Monte Carlo power analysis for different effect sizes:

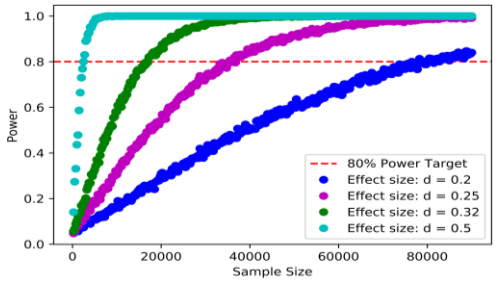


Figure 4. Monte Carlo Power analysis under normal hypothesis

The sample data is generated from a normal distribution with mean $\mu \approx 180\ \mu V$ and standard deviation $\sigma \approx 90\ \mu V$. These values correspond to the mean and standard deviation of the overall blinks distribution. Since the effect size is not known, several ones are tested to compute the minimum sample size required to achieve an 80% power target. Moreover, the new EEG experiment is first carried out on a few subjects to approximate the effect size. Depending on the number of subjects tested (1 to 2 subjects), this effect size corresponds to either a small or medium effect. Since EEG recordings are usually influenced by unwanted noise, a small effect size is understandable. Additionally, since the phenomena under investigation, i.e. the blinks, is highly subject-dependent, the average difference is "large enough to be visible to the naked eye", which is the definition of a medium effect size [11].

The corresponding minimum sample sizes are then the minimum number of blinks required, as shown in Table II.

Furthermore, adults usually blink 15 times per minute. From the number of blinks, one can then deduce the minimum hours of recording (rounded up) necessary depending on the effect size.

TABLE II. SAMPLE SIZE DETERMINATION UNDER NORMAL ASSUMPTION

| **Effect Size ($d$)** | 0.2 | 0.25 | 0.32 | 0.5 |
|---|---|---|---|---|
| **Number of blinks** | 80900 | 37000 | 17500 | 2750 |
| **Hours of recording** | 90 | 42 | 20 | 4 |

Consequently, for an effect size $d = 0.2$, the new EEG experiment should either be carried out on 90 subjects for one hour each, or several sessions of each one hour could be carried out to reduce the number of subjects. Obviously, it is not always easy to convince volunteers to come back for many sessions, since the previous calculations only take into account the number of hours of recording and not the whole preparation time. A trade-off could be to ask some volunteers to carry out 3 sessions of each one hour of recording. In this case, the number of subjects required drops down to $N_{subjects} = 30$ for a small effect size.

### C. Method 2: Using the Fitted Distribution(s)

From the considerations of the section II.D., the overall blinks distribution is the least less likely to follow either a Weibull or a Logistic distribution. A similar Monte Carlo power analysis is carried out with both these assumptions. The parameters for the Weibull distribution are calculated from the estimated overall distribution, that yielded $location \approx 2$ and $scale \approx 204$. Similarly, the parameters for the Logistic distribution are $location \approx 215$ and $scale \approx 37$.

Consequently, to achieve an effect size of $d = 0.2$ with the Weibull assumption, there should be a minimum of 76000 blinks, which correspond to 85 hours of recording. Thus, the new EEG experiment should record 3 sessions of 1-hour for 29 subjects. With the Logistic assumption, the minimum number of blinks decreases to 45000, giving a minimum of 45 hours of recording. The sample requirement decreased by approximately 6% with the former assumption, and 44% with the latter.

## IV. DISCUSSION

Monte Carlo simulations can be used to compute the minimum number of subjects for a new EEG experiment. Different underlying distribution hypotheses yield different requirements for the minimum sample size. However, even though the least less likely hypothesis should return a better approximation, one should keep in mind the physical limitations of the new EEG experiment to be designed. A session longer than one hour of recording would be difficult to bear for the subject and might yield invalid data. Similarly, asking volunteers to come for more than 3 sessions is likely to be difficult.

As far as we know, this paper seems to be the first attempt to calculate the minimum number of subjects required in an EEG experiment. As such, there is a still a lot of improvements to ensure the correct computations as well as the easiness of use for new a priori calculations. Particularly, one should very carefully considerate the choice of a fitting distribution, as it influences greatly the sample requirement, and ponder that this result is, in any case, a "best case scenario" estimate.

Furthermore, several hypotheses have been made throughout this paper and should be highlighted. The first one is that these computations can only be calculated for similar EEG experiment/phenomenon. In this paper, we focused on the spontaneous blinks' distribution, which is a prominent and easily detectable EEG feature. Thus, simply extracting $FP1\_Max$ should already yield satisfactory results. Results would have undoubtedly been different for voluntary blinks, since conscious blinks tend to be complete and thus should yield a more uniform distribution.

The observed variable to be chosen is highly classification dependent. In our work, we have chosen to classify blinks using the maximum amplitude on a frontopolar channel as a parameter. For other EEG/ERP features, the maximum potential might not be adequate. Instead, for ERN or P300, one might consider the area under specified electrodes and for a fixed time as a suitable measure. Moreover, if Riemannian geometry classifiers were to be used for Motor Imagery classification, one might consider the minimum distance to the Riemannian mean as a parameter.

Finally, the tasks performed by the subjects should be similar in both experiments (historical data and the experiment to be done). Depending on the phenomenon to be investigated, diverse tasks might very well influence differently the measure.

## REFERENCES


[1] A. S. Malik, H. U. Amin, *Designing EEG Experiments for Studying the Brain*, Academic Press, 2017.

[2] S. A. Baldwin, M. J. Larson, P.E. Clayson, "The dependability of electrophysiological measurements of performance monitoring in a clinical sample: A generalizability and decision analysis of the ERN and Pe", *Psychophysiology*, Vol 52, Issue 6, pp. 790 – 800, 2015.

[3] A. Casadevall, F. C. Fang, "Rigorous Science: a How-To Guide", *mBio*, 7 (6) e01902-16, Nov 2016.

[4] M. J. Larson, K. A. Carbine, "Sample size calculations in human electrophysiology (EEG and ERP) studies: A systematic review and recommendations for increased rigor", *International Journal of Psychophysiology*, 111, pp. 33 – 41, 2016.

[5] F. Farrokhyar, D. Reddy, R. W. Poolman, M. Bhandari, "Why perform a priori sample size calculation?", *Can J Surg.*, Vol 56, Issue 3, pp. 207 – 13, Jun 2013.

[6] B. Cundill, N. D. E. Alexander, "Sample size calculations for skewed distributions", *BMC Medical Research Methodology*, 15, 28, 2015.

[7] M. Perrin, E. Maby, S. Daligault, O. Bertrand, J. Mattout, "Objective and subjective evaluation of online error correction during P300-based spelling", *Advances in Human-Computer Interaction*, Vol 4, 2012.

[8] E. Guttmann-Flury, X. Sheng, D. Zhang, X. Zhu, "Preliminary results on a new algorithm for blink correction adaptive to inter- and intra-subject variability", *9th International IEEE EMBS Conference on Neural Engineering* (Accepted), 2019.

[9] M. L. Delignette-Muller, C. Dutang, "fitdistrplus: An R package for Fitting Distributions", *Journal of Statistical Software*, Vol 64, Issue 4, pp. 1 – 34, 2015.

[10] M. G. Doane, "Interaction of Eyelids and Tears in Corneal Wetting and the Dynamics of the Normal Human Eyeblink", *American Journal of Ophthalmology*, Vol 59, pp. 507 – 516, 1980.

[11] J. Cohen, "Statistical Power Analysis for the Behavioral Sciences", 2nd Edition, Academic Press, 1988.