Tutorials in Quantitative Methods for Psychology
2007, Vol.3 (2), p. 63-69.

DOI: 10.20982/tqmp.03.2.p063

Understanding statistical power using noncentral probability distributions: Chi-squared, G-squared, and ANOVA

Sébastien Hélie
Rensselaer Polytechnic Institute

This paper presents a graphical way of interpreting effect sizes when more than two groups are involved in a statistical analysis. This method uses noncentral distributions to specify the alternative hypothesis, and the statistical power can thus be directly computed. This principle is illustrated using the chi-squared distribution and the *F* distribution. Examples of chi-squared and ANOVA statistical tests are provided to further illustrate the point. It is concluded that power analyses are an essential part of statistical analysis, and that using noncentral distributions provides an argument in favour of using a factorial ANOVA over multiple *t* tests.

Statistics occupy a large portion of psychological papers. While this increase in the space allowed to describe the analyses used in psychological experiments allows for a more thorough understanding of the data, it was not accompanied by a corresponding boost in the formation of psychologists (Giguère, Hélie, & Cousineau, 2004). In particular, a substantial amount of class time is usually devoted to the explanation of type I error (wrongfully rejecting the null hypothesis), but type II error is usually not worth more than a mere mention in undergraduate classes (failing to reject the null hypothesis when it is false). This latter type of error is very important, because it complements the notion of statistical power. Statistical power, also referred to as *sensitivity*, is the probability of correctly rejecting the null hypothesis. Figure 1 illustrates statistical power in the case of a chi-squared distribution. As in all statistical tests, the aim is to choose which of two hypotheses (distributions) is correct. When the test statistic is higher than a predefined threshold $f(\alpha)$, the alternative hypothesis is chosen (full line); otherwise, the null hypothesis is chosen (dashed line). Because statistical power is the probability to correctly choose the alternative hypothesis, it is illustrated by the portion of the former distribution to the right of the threshold (gray part).

Statistical power is a function of three parameters: the probability of committing type I error, the reliability of the sample, and the effect size (Cohen, 1988). The first parameter, probability of type I error, is positively related to statistical power: when the probability of wrongfully rejecting the null hypothesis is increased, statistical power is also increased. This can be easily seen in Figure 1: when the probability of type I error is increased (by moving the threshold to the left), the area of the grey portion is bigger. Likewise, the statistical power can be decreased by moving the threshold to the right (diminishing the probability of type I error).

Sample reliability is usually controlled in psychology by randomly selecting participants. Hence, by the law of large numbers, bigger samples are more reliable than smaller ones. Operationally, the reliability of a sample is usually

This research was supported by a postdoctoral fellowship from *Le fonds québecois de la recherche sur la nature et les technologies*. The author would like to thank Dr. Jean Descôteaux and an anonymous reviewer for their useful comments on a previous version of this manuscript. Requests for reprints should be addressed to Sébastien Hélie, Rensselaer Polytechnic Institute, Cognitive Science Department, 110 Eighth Street, Carnegie 108, Troy (NY), 12180-3590 (USA) , or using e-mail at helies@rpi.edu

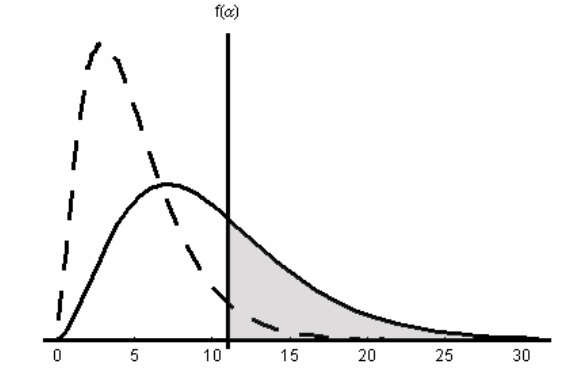

Figure 1. Chi-squared and noncentral chi-squared distributions. The former represents the null hypothesis and is illustrated using the dashed line (υ = 5), while the latter represents the alternative hypothesis and is illustrated using the full line (υ = 5, λ = 5). The grey portion represents statistical power.

measured using the standard error of the sample mean:

$$SE_{\bar{X}} = \sqrt{\frac{s^2}{n}} \quad (1)$$

where s^2 is the unbiased estimate of the population's variance, and n is the sample size. It is easy to see that Eq. 1 diminishes when the sample size is increased. Because test statistics are usually scaled using standard errors, decreasing this measure increases the test statistics, and thus the statistical power of the test.

The last parameter that affects statistical power is effect size, which directly refers to the proportion of the change in the dependant variable that can be attributed to the controlled factor. While this parameter is easy to interpret in the context of tests that compare two means (e.g., the scaled difference between the means in a *t* test), it is more difficult to understand in cases involving more than two groups (e.g., ANOVAs) or entire distributions (e.g., chi-squared, G-squared). For instance, the null hypothesis when performing an ANOVA is that the variance of the groups' mean is zero. Likewise, the null hypothesis when performing a chi-squared or G-squared test is that the variance of the proportions is zero. In both cases, this can only be achieved if all the means or all the proportions are the same, and the effect size can be best understood using the noncentral chi-squared (chi-squared, G-squared) or noncentral *F* distributions (ANOVA). These two cases are now illustrated.

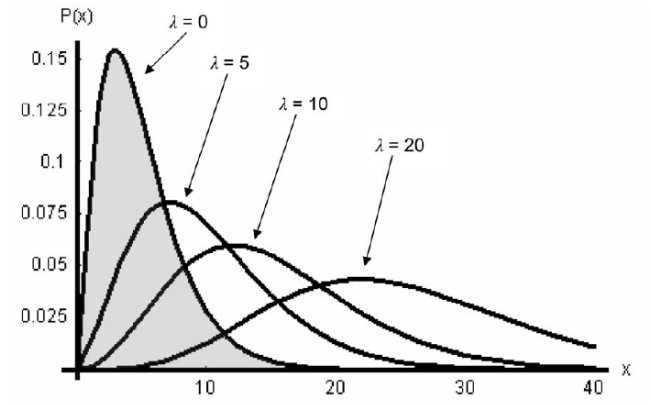

Figure 2. Noncentral chi-squared distributions (υ = 5) with varying λs. The filled distribution represents a null hypothesis. As λ increases, the overlap of the alternative hypotheses with the null hypothesis (grey region) diminishes.

Chi-squared, G-squared, and the noncentral chi-squared distribution

As argued earlier and shown in Figure 1, hypotheses in statistical tests are usually represented using distributions: the test statistics measures how much they differ. When performing statistical tests using the chi-squared distribution, the alternative hypothesis is not usually represented; only the null hypothesis is (the usual chi-squared distribution), and the threshold found in a statistical table is used to either accept or reject the null hypothesis. However, when performing a power analysis, the alternative hypothesis has to be specified. In the case of the chi-squared and the G-squared statistical tests, the alternative hypothesis is a noncentral chi-squared distribution (as shown in Figure 1). This distribution is described by:

$$p(x \mid \upsilon, \lambda) = \frac{e^{-\frac{(x+\lambda)}{2}} x^{\frac{\upsilon}{2}-1}}{2^{\frac{\upsilon}{2}}} \sum_{k=0}^{\infty} \frac{(\lambda x)^k}{2^{2k} k! \Gamma\left(k + \frac{\upsilon}{2}\right)} \quad (2)$$

with $\upsilon > 0$ degrees of freedom, $\lambda \geq 0$ is the noncentrality parameter, and $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ is the Gamma function. When $\lambda = 0$, Eq. 2 can be simplified to:

$$p(x \mid \upsilon, 0) = p(x \mid \upsilon) = \frac{x^{\frac{\upsilon}{2}-1} e^{-\frac{x}{2}}}{\Gamma\left(\frac{\upsilon}{2}\right) 2^{\frac{\upsilon}{2}}} \quad (3)$$

which is the usual chi-squared distribution with $\upsilon > 0$ degrees of freedom. The smaller the value given to λ, the bigger the overlap between the null and alternative hypotheses. On the other hand, the bigger the value given to λ , the smaller the overlap and, as a result, the higher the statistical power. Figure 2 illustrates several examples of noncentral chi-squared distributions.

Because the alternative hypothesis necessarily has the same number of degrees of freedom as the null hypothesis,

all that is needed to specify the alternative hypothesis is an estimation of the noncentrality parameter (λ). As hinted earlier, λ is a function of the effect size:

$$\lambda = n \times \omega^2 \tag{4}$$

where n is the sample size, and ω is the effect size. In the case of a chi-squared or a G-squared statistical test, the effect size can be estimated by:

$$\omega = \sqrt{\sum_{i=1}^{k} \frac{(H_{1i} - H_{0i})^2}{H_{0i}}} \tag{5}$$

where k is the number of cells, H_{0i} is the expected proportion of data in cell i, and H_{1i} is the actual proportion of data in cell i. It is noteworthy that Eq. 5 is similar to the formula used to compute the chi-squared test statistic, except that proportions are used instead of frequencies.

Once the effect size estimated, the statistical power of the test is easily computed:

$$power = 1 - \left(\int_0^x p(x \mid \upsilon, \lambda)\, dx \right)\Bigg|_{f(\alpha)} \tag{6}$$

where the term in parenthesis is the cumulative density function of the noncentral chi-squared distribution (Eq. 2) evaluated at the threshold found in a chi-squared table ($f(\alpha)$). This value can be given by any scientific computation software (e.g., Mathematica, Matlab). Listing 1 presents the Mathematica code to compute the power of a chi-squared or G-squared statistical test.

Example

The owner of a small gift shop wanted to know if people were buying their Christmas gifts at the last minute or if they were gradually buying them throughout the entire month of December. To do that, he calculated the number of gifts sold each week with the following result: H_1 = {37, 21, 21, 21}. These results were to be compared with a regular month, in which the gift sells are uniformly distributed across the weeks: H_0 = {25, 25, 25, 25}. In this particular case, k = 4, and H_0 ~ chi-squared(3), while H_1 ~ noncentral chi-squared(3, 7.68). If the probability of committing type I error is set to 0.05, applying the chi-squared statistical test lead to the rejection of H_1 ($\chi^2(3) = 2.77$, $p > .05$). Hence, the consumers' habits in December do not differ from their usual. This conclusion can be made with confidence, because the statistical power of the test is 0.63 (see Listing 1). Hence, if the consumers' habits were in fact different in the month of December, the probability of detecting this difference would be 0.63, which is sufficient to conclude with relative confidence.

The ANOVA and the noncentral F distribution

As in the chi-squared and G-squared statistical tests, only the null hypothesis is usually represented when performing an ANOVA (a F distribution). This reluctance to illustrate the alternative hypothesis follows from a difficulty to interpret the effect size, which is related to the noncentrality parameter of a noncentral F distribution. This latter distribution is described by:

$$p(x \mid u, v, \lambda) = e^{-\frac{\lambda}{2} + \frac{\lambda u x}{2(v+ux)}} u^{\frac{u}{2}} v^{\frac{v}{2}} x^{\frac{u}{2}-1} \times (v+ux)^{-\frac{(u+v)}{2}} \frac{\Gamma\left(\frac{u}{2}\right)\Gamma\left(1+\frac{v}{2}\right) L_{\frac{v}{2}}^{\frac{u}{2}-1}\left(-\frac{\lambda u x}{2(v+ux)}\right)}{B\left(\frac{u}{2}, \frac{v}{2}\right)\Gamma\left(\frac{u+v}{2}\right)} \tag{7}$$

where u > 0 is the number of degrees of freedom of the tested effect (numerator), v > 0 is the number of degrees of freedom of the error term (denominator), $\lambda \geq 0$ is the noncentrality parameter, $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ is the usual Beta function, and

$$L_n^k(x) = \frac{e^x x^{-k}}{n!} \frac{d^n}{dx^n}(e^{-x} x^{n+k})$$

is a generalized Laguerre polynomial. Fortunately, when λ = 0, Eq. 7 is simplified to the usual F distribution:

$$p(x \mid u, v, 0) = p(x \mid u, v) = \frac{u^{\frac{u}{2}} v^{\frac{v}{2}} x^{\frac{u}{2}-1} (v+ux)^{-\frac{u+v}{2}}}{B\left(\frac{u}{2}, \frac{v}{2}\right)} \tag{8}$$

Figure 3 shows noncentral F distributions with different λs. Like the noncentral chi-squared distribution, an increase in the value of the noncentrality parameter diminishes the overlap between the null and alternative hypotheses, which in turn increases the statistical power.

Because the alternative hypothesis has the same number of degrees of freedom as the null hypothesis, only the noncentrality parameter (λ) remains to be specified. The latter is expressed as:

$$\lambda = \omega^2 \frac{(u+1)}{k} \sum_{i=1}^{k} n_i \tag{9}$$

where ω is the effect size, n_i is the number of participants in group i, and k is the number of groups. It is noteworthy that this Equation is the same as Eq. 4 in the case of one-way ANOVAs. The effect size can be estimated using:

$$\omega = \frac{\sigma_m}{\sigma} = \sqrt{\frac{\eta^2}{1-\eta^2}} \tag{10}$$

where η^2 is the amount of explained variance, σ is the

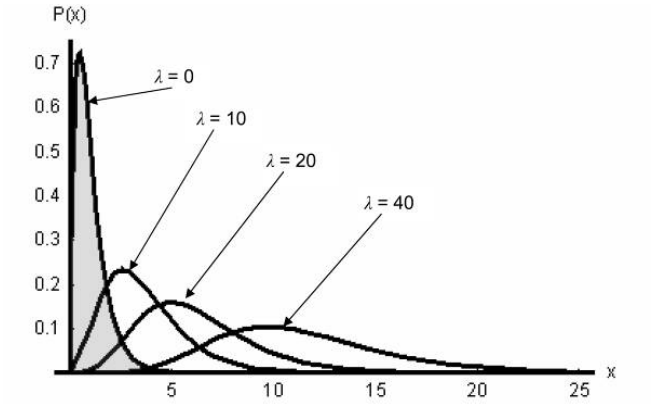

Figure 3. Noncentral F distributions (u = 4, v = 45) with varying λs. The filled distribution represents a null hypothesis. As λ increases, the overlap of the alternative hypotheses with the null hypothesis (grey region) diminishes.

common standard deviation of the sample (without considering group membership):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - m\right)^2}{\sum_{i=1}^{k} n_i - 1}} \qquad (11)$$

where x_{ij} is data j in group i, and m is the common mean of the sample (without considering group membership):

$$m = \frac{\sum_{i=1}^{k} n_i m_i}{\sum_{i=1}^{k} n_i} \qquad (12)$$

where m_i is mean of group i. In Eq. 10, the numerator (σ_m) is the standard deviation of the k means. It can be computed as:

$$\sigma_m = \sqrt{\frac{\sum_{i=1}^{k} n_i (m_i - m)^2}{\sum_{i=1}^{k} n_i}} \qquad (13)$$

Of course, if the ANOVA was performed using a statistical software (e.g., SPSS), the amount of explained variance is directly provided, and the effect size is quickly computed (Eq. 10). However, when computed by hand, all the preceding steps must be done to accurately estimate the effect size (Eq. 10 – Eq. 13).

Once the effect size is estimated, the alternative hypothesis is fully specified and the power can be computed:

$$power = 1 - \left(\int_0^x p(x \mid u, v, \lambda)\, dx \right)\Bigg|_{f(\alpha)} \qquad (14)$$

Table 1. Intellectual Quotients used in the example

High school	Undergraduate	Graduate
91	95	85
85	113	119
95	80	102
75	93	124
111	102	97
88	97	87
87	99	87
111	91	87
114	82	92
84	88	111

where the term in parenthesis is the cumulative density function of the noncentral F distribution (Eq. 7) evaluated at the threshold found in a F table ($f(\alpha)$). This value can be given by any scientific computation software (e.g., Mathematica, Matlab). Listing 2 presents the Mathematica code to compute the power of an ANOVA.

Example

A psychologist wanted to find out if the Intelligence Quotient (IQ) was related to education. Hence, he measured the IQ of thirty 35 years-old workers, ten of which had a high-school diploma, another ten had undergraduate college degrees, and the remaining had graduate university degrees. The obtained measures are shown in Table 1. A one-way ANOVA was performed on the scores, and no group effect was found (F(2, 27) = 0.53, p > .05). However, the researcher should be careful before making strong conclusions about the absence of link between IQ and education. Because the effect size is small (~ 0.20) and the chosen probability of committing type I error was set to 0.05, the power of this statistical test is only 0.13. Hence, if there is a difference between the groups, the probability of detecting it is only 0.13. These results should thus be interpreted as inconclusive. The experiment should be redone with a bigger sample.

Discussion

This short paper presented a simple way to interpret the effect size in statistical tests involving more than two groups: it specifies the noncentrality parameter of the distribution used to represent the alternative hypothesis. With a complete specification of the alternative hypothesis, the notion of statistical power can be intuitively grasped by plotting the distributions representing the hypotheses. This was shown using the noncentral chi-squared distribution and noncentral F distribution. The former is used to analyze categorical data using the chi-squared or G-squared

statistical test. While the example provided in Listing 1 shows a simple association between two measures, the same method applies to multi-way contingency tables involving the G-squared statistics (Milligan, 1980): only the computation of the number of degrees of freedom changes (for a detailed presentation of multi-way table analyses, see Agresti, 1996). Moreover, this rationale also applies to other linear models, such as the ANOVA (which uses the latter distribution). While the example in Listing 2 shows a one-way ANOVA, the power of factorial ANOVAs can be computed in the exact same way: all that changes is the number of degrees of freedom and the grouping. For instance, in a 2 × 3 factorial design, the power of the first main effect is computed using two groups (without considering the second factor). Likewise, the power of the second main effect is computed using three groups (without considering the first factor), and the power of the interaction is computed using six groups (all factors are now considered). The degrees of freedom used to compute the power of each effect follows the standard decomposition presented in any introductory statistical text (e.g., Hays, 1981).

The computation of the statistical power of an interaction brings forward the importance of correctly performing a factorial ANOVA when several factors are involved (instead of using multiple *t* tests). By examining Eq. 9, it is easy to see that the noncentrality parameter (and thus the power) diminishes as the number of groups increases. This explains why it is more difficult to obtain a statistically significant interaction than several statistically significant *t* tests. However, the loss of power when analyzing a factorial design is statistically sound and should not be avoided.

To summarize, statistical power is a very important notion that relies for a large part on the notion of effect size. While this notion is easy to visualize when only two groups are included in the analysis, it is more difficult to interpret when several groups are involved. This paper presented an intuitive way to interpret effect sizes by estimating a noncentrality parameter to fully define the alternative hypothesis. It is our hope that a better understanding of these issues will result in more power analyses.

References

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York: John Wiley & Sons.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. *2nd Edition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Giguère, G., Hélie. S., & Cousineau, D. (2004). Manifeste pour le retour des sciences en psychologie. *Revue Québécoise de Psychologie*, *25*, 117-130.

Hays, W.L. (1981). *Statistics. 3rd Edition*. New York: CBS College Publishing.

Milligan, G.W. (1980). Factors that affect type I and type II error rates in the analysis of multidimensional contingency tables. *Psychological Bulletin*, *87*, 238-244.

Manuscript received November 14th, 2006
Manuscript accepted September 17th, 2007

Listings follows

Computing the power of a chi-squared or G-squared statistical test

```
In[12]:= << Statistics`ContinuousDistributions`
         Off[General::"spell"]
         Off[General::"spell1"]
         Off[Solve::"ifun"]
```

Computing the effect size (ω)

```
In[16]:= k = 4; (* number of cells *)
```

$$\omega[H0\_, H1\_] := \sqrt{\sum_{i=1}^{k} \frac{(H1[\![i]\!] - H0[\![i]\!])^2}{H0[\![i]\!]}}$$

Finding the threshold without using a table [f(α)]

```
In[17]:= ν = 3; (* number of degrees of freedom *)
         α = .05; (* probability of committing type I error *)
         threshold = NSolve[CDF[ChiSquareDistribution[ν], x] == (1 - α), x][[1, 1, 2]]
```

```
Out[18]= 7.81473
```

Computing the noncentrality parameter (λ)

```
In[19]:= n = 100; (* sample size *)
         H0 = {.25, .25, .25, .25};
         (* expected proportion for each cell: the list must sum to unity *)
         H1 = {.37, .21, .21, .21};
         (* observed proportion for each cell: the list must sum to unity *)
         λ = (n ω[H0, H1]^2)
```

```
Out[21]= 7.68
```

Computing the power of the statistical test

```
In[22]:= power = 1 - CDF[NoncentralChiSquareDistribution[ν, λ], threshold]
```

```
Out[22]= 0.634259
```

Computing the power of an ANOVA

```
In[1]:= << Statistics`ContinuousDistributions`
        Off[Solve::"ifun"]
        Off[General::"spell1"]
```

Computing the effect size (ω) using the explained variance (η^2)

```
In[4]:= η2 = .038; (* the effect size was estimated with SPSS *)
        ω = Sqrt[η2 / (1 - η2)]
Out[4]= 0.198749
```

Computing the effect size (ω) without using the explained variance (η^2)

```
In[5]:= data = {{91, 85, 95, 75, 111, 88, 87, 111, 114, 84},
          {95, 113, 80, 93, 102, 97, 99, 91, 82, 88}, {85, 119, 102, 124, 97, 87, 87, 87, 92, 111}};
        (* each sublist is a different group *)
        k = Length[data];
        m = Mean[Flatten[data]];
        mi = Table[Mean[data[[i]]], {i, 1, k}];
        ni = Table[Length[data[[i]]], {i, 1, k}];
        σm = Sqrt[Sum[ni[[i]] (mi[[i]] - m)^2, {i, 1, k}] / Plus @@ ni];
        σ = StandardDeviation[Flatten[data]];
        ω = N[σm / σ]
Out[11]= 0.191351
```

Finding the threshold without using a table [f(α)]

```
In[12]:= u = k - 1; (* number of degrees of freedom *)
         v = (Plus @@ ni) - k;
         α = .05; (* probability of committing type I error *)
         threshold = Last[Solve[CDF[FRatioDistribution[u, v], x] == (1 - α), x]][[1, 2]]
Out[13]= 3.35413
```

Computing the noncentrality parameter (λ)

```
In[14]:= λ = ω^2 (Plus @@ ni) / k (u + 1)
Out[14]= 1.09846
```

Computing the power of the statistical test

```
In[15]:= power = N[1 - CDF[NoncentralFRatioDistribution[u, v, λ], threshold]]
Out[15]= 0.131313
```