2023 ■ Vol. 19 ■ no. 2

Analysis of frequency data: The ANOFA framework

Louis Laurencelle [a] and Denis Cousineau [b]

[a]Université du Québec à Trois-Rivières
[b]Université d'Ottawa

Abstract ■ Analyses of frequencies are commonly done using a chi-square test. This test, derived from a normal approximation, is deemed generally efficient (controlling type-I error rates fairly well and having good statistical power). However, in the case of factorial designs, it is difficult to decompose a total test statistic into additive interaction effects and main effects. Herein, we present an alternative test based on the G statistic. The test has similar type-I error rates and power as the former one. However, it is based on a total statistic that is naturally decomposed additively into interaction effects, main effects, simple effects, contrast effects, etc., mimicking precisely the logic of ANOVAs. We call this set of tools ANOFA (Analysis of Frequency data) to highlight its similarities with ANOVA. We also examine how to render plots of frequencies along with confidence intervals. Finally, quantifying effect sizes and planning statistical power are described under this framework. The ANOFA is a tool that assesses the significance of effects instead of the significance of parameters; as such, it is more intuitive to most researchers than alternative approaches based on generalized linear models.

Keywords ■ frequency; contingency table; analyses of frequencies; additive decomposition.

denis.cousineau@uottawa.ca

10.20982/tqmp.19.2.p173

Acting Editor ■ Sydney Lambert (Université d'Ottawa)

Reviewers
■ Four anonymous reviewers

Introduction

Analyzing frequencies is a complicated task with no agreed-upon solutions. In the psychological sciences, it is typically done using a chi-square test in which the differences between the observed and the predicted frequencies are transformed into approximate z scores. As the sum of squared z scores follows a χ^2 distribution, this results in an easy-to-perform test. However, the chi-square test is limited when the data are classified using two or more dimensions (2-way frequency tables or beyond): indeed, orthogonal decomposition of the total test score is difficult. Castellan (1965) and others presented convoluted techniques to analyze portions of two-way tables (e.g., Bresnahan & Shapiro, 1966; Fagen & Mankovich, 1980; ; see the review by Sharpe, 2015). As Shaffer (1973a) puts it: "when relationships among variables are defined in an intuitively acceptable manner, the partitioned chi-square values do not correspond to tests of these relationships" (p. 127). In the linguistic sciences, one approach uses the generalized linear model (GLM, initially developed by McCullagh & Nelder, 1989; ; also see Venables & Ripley, 2002). In education, we find other approaches including the one proposed in Light and Margolin (1971) called CATANOVA (also see D'Ambra et al., 2005). However, this last technique is limited to two-factor designs and is rigid as effects cannot be decomposed into simple effects or contrast effects; it will therefore not be discussed further.

Yet an exact approach has been around since the early days of statistical testing. This approach is not based on approximations as it uses the true underlying multinomial distribution, and tests of hypotheses use the likelihood ratio test. Cochran (1936) mentions that Fisher (1922) and Neyman and Pearson (1928) examined this alternative. It results in procedures akin to log-linear modeling (Shaffer, 1973a, 1973b). See Mood et al. (1974), Agresti (2013), and Fienberg (2007) for formal derivations, and Hoeffding (1965) for an examination of statistical power.

Herein, we expand this approach and show how main effects and interaction effects can be tested in frequency tables with any number of dimensions. We also show that simple effects can easily be obtained, owing to the decom-

position of the test statistic. Finally, we also show how orthogonal contrasts can be performed on frequency data, a novel application.

This approach to frequency analysis is versatile and covers almost the entire range of applications that classical ANOVA allows. Hence, we propose to name the present framework the Analysis of Frequency data (ANOFA) to emphasize the similarities. We also present a method to plot the frequencies along with confidence intervals. Finally, effect sizes and statistical power planning can be performed within the ANOFA framework using the same tools and concepts that are already familiar to ANOVA users.

The most salient strength of the ANOFA is that it focuses on effects. Other approaches such as GLM are efficient at estimating parameters, however, when asking simple questions such as "Is there an interaction?", they are not very potent because the interaction effect is spread across multiple parameters estimated by the GLM procedure. We will return to this key distinction when we present the third illustration.

The G statistic

The ANOFA takes its root in the multinomial probability distribution whereby a total of N observations are distributed across C classes. The multinomial model is an extension of the binomial model to more than two outcomes. The probability of belonging to a certain class i, say π_i, may be different across classes as long as $0 < \pi_i < 1$ and $\sum_{i=1}^{C} \pi_i = 1$. Under a multinomial model, the probability of observing the given cell counts $(n_1, n_2, \ldots, n_C)$ (where $\sum_{i=1}^{C} n_i = N$) is given by

$$Pr\{n_1, n_2, \ldots, n_C \mid \pi_1, \ldots, \pi_C\} = \frac{N!}{n_1! n_2! \ldots n_C!} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_C^{n_C}.$$

In cases where there are only two classes, this reduces to the binomial probability model. The π_i parameters are unknown, but they can be estimated by $\hat{\pi}_i = n_i/N$; these proportions $\hat{\pi}_i$ are the maximum likelihood estimators of the corresponding π_i (Mood et al., 1974). Using these estimates, the likelihood (or conditional probability) of the set of estimated parameters given the observed counts is maximal and equal to

$$\ell\{\hat{\pi}_1, \ldots, \hat{\pi}_C \mid n_1, n_2, \ldots, n_C\} = \frac{N!}{n_1! n_2! \cdots n_C!} \hat{\pi}_1^{n_1} \hat{\pi}_2^{n_2} \cdots \hat{\pi}_C^{n_C}$$

which is consequently the likelihood of the best-fitting model.

In the presence of a hypothesis $\mathcal{H}_0$ specifying the expected cell probabilities under the “null model” $\mathcal{H}_0 : \pi_1 = \pi_{01}, \pi_2 = \pi_{02}, \ldots \pi_C = \pi_{0C}$, and in particular the hypothesis of no difference across categories, $\mathcal{H}_0 : \pi_{0i} = 1/C$, it is possible to contrast the best-fitting model to this null model using the ratio of their likelihoods, sometimes noted LR, with

$$\begin{aligned} LR &= \frac{\ell\{\pi_{01}, \ldots, \pi_{0C} \mid n_1, \ldots, n_C\}}{\ell\{\hat{\pi}_1, \ldots, \hat{\pi}_C \mid n_1, \ldots, n_C\}} \\ &= \frac{\frac{N!}{n_1! n_2! \cdots n_C!} \pi_{01}^{n_1} \pi_{02}^{n_2} \cdots \pi_{0C}^{n_C}}{\frac{N!}{n_1! n_2! \cdots n_C!} \hat{\pi}_1^{n_1} \hat{\pi}_2^{n_2} \cdots \hat{\pi}_C^{n_C}} \\ &= \left(\frac{\pi_{01}}{\hat{\pi}_1}\right)^{n_1} \left(\frac{\pi_{02}}{\hat{\pi}_2}\right)^{n_2} \cdots \left(\frac{\pi_{0C}}{\hat{\pi}_C}\right)^{n_C} \\ &= \prod_{i=1}^{C} \left(\frac{\pi_{0i}}{\hat{\pi}_i}\right)^{n_i} \end{aligned}$$

and twice the negative of the log likelihood ratio, hereafter named G, can compactly be written as

$$G = -2 \sum_{i=1}^{C} n_i \left(\log \pi_{0i} - \log \hat{\pi}_i\right). \tag{1}$$

This G statistic is an exact and sufficient test statistic. Its distribution follows asymptotically a χ^2 distribution with $C - 1$ degrees of freedom (df; Black & Laurencelle, 1987; Laurencelle, 2022; Wilks, 1938, this measure is sometimes noted G^2, L or $-2 \log LR$). Williams (1976) examined this formula and found that, for small cell counts, the chi-square critical values are biased downward. Williams suggested a correction factor c_W, whose generic form is

$$c_W(n_{\text{cells}}, \nu, N) = 1 + \frac{n_{\text{cells}}^2 - 1}{6 \nu N}, \tag{2}$$

in which n_{cells} is the number of cells involved in the G statistic to be corrected, ν are the associated df (equal to $C-1$ in one-way frequency tables), and N is the total sample size. The Williams correction factor is used to divide the G statistic; it yields a number larger than 1 but tends to 1 as N gets large.

Eq. (1) is analogous to an ANOVA equation. It applies in cases of a single "factor", i. e., where the observations are classified along a single dimension. However – and we will show examples in the subsequent section – this formula can be generalized to any number of classification dimensions. A table is a case where there are two classification dimensions (e.g., in general terms, the rows and the columns), but there can also be three dimensions (a classification within a cube having multiple layers), etc. More importantly, in situation presenting two or more dimensions, the formula can be generalized for analyzing, for instance, main effects, interaction effects, and components of these in the form of simple or interaction effects.

In what follows, we present examples of the use of the ANOFA for a two-way design. The appendix provides the

Figure 1 ■ The Landis et al. (2013) observed frequencies as a function of the financing source for three family medicine residency programs. Error bars shows the difference-adjusted 95% confidence intervals of the observed frequencies.

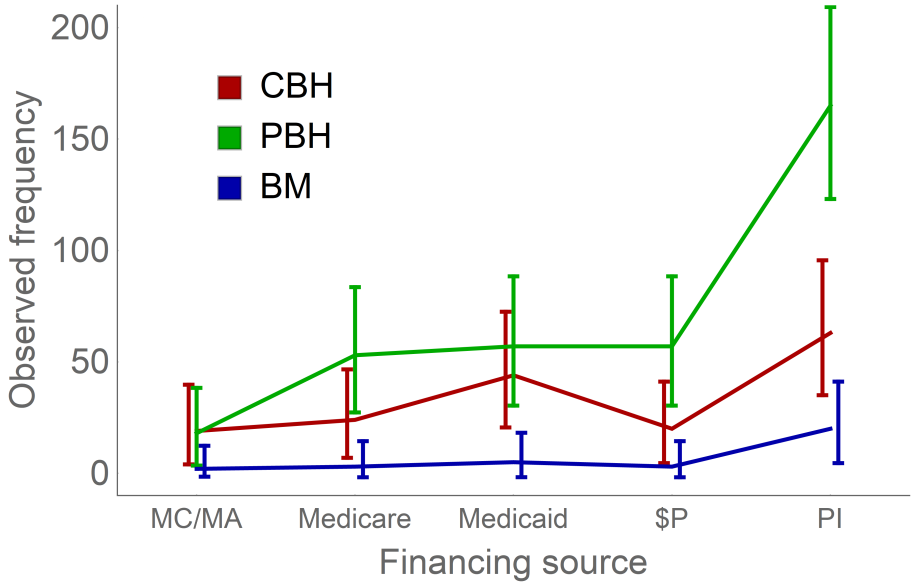


Table 1 ■ Counts for the participants based on the *Source of their financing* and *type of health service* factors.

	Collocated behavioral health(CBH)	Primary-care behavioral health (PBH)	Blended model (BM)
Medicare / Medicaid (MC/MA)	19	18	2
Medicare (MC)	24	53	3
Medicaid (MA)	44	57	5
Self-paid ($P)	20	57	3
Personal insurance (PI)	63	165	20

Note. The total sample size is 553, an important sample size.

equations for designs with a single, two, three, or four factors; they can be generalized to any number of factors. We conclude by briefly reviewing Monte Carlo simulations that examined type-I error rates, specificity, and statistical power of the ANOFA. In a nutshell, type-I error rates match the decision threshold when the correction factor (Eq. 2) is used. Without the correction factor, they can reach up to .062 for a .05 decision threshold (similar excess of type-I error rates are found for the chi-square test; e. g., Laurencelle, 2022). As for statistical power, tests of the G statistic equal or surpass slightly the Pearson's chi-square tests on the tests that the chi-square test can perform (as demonstrated in Hoeffding, 1965).

We preserve four significant digits in the calculations exemplified herein, although this is unrealistic in most practical applications (Cousineau, 2020): this is to allow interested readers to check computations if they wish to reproduce them.

Three illustrations

In this section, we go through three examples. The first is a 5×3 design examining modality of care (Landis et al., 2013). A two-way omnibus ANOFA is performed, and the significant interaction is decomposed using two different strategies. The second is a 2×5 example examining education vocation of boys and girls in the 1960s (first presented in Light & Margolin, 1971). Following an omnibus ANOFA, simple effects are examined. The last example is a $3 \times 2 \times 2 \times 2$ example of detergent use (Ries & Smith, 1963). This last example is more complex, with four factors resulting in 24 cells. Yet, ANOFA readily and easily locates the significant effects, making the data interpretable.

Table 2 ■ ANOFA results of data from Table 1.

Omnibus ANOVA	G	df	$G_{\text{corrected}}$	p-value
Total	533.187	14		
Source of financing (A)	206.568	4	206.196	$< .00001$
Residency program (B)	307.773	2	307.403	$< .00001$
$A \times B$	18.8455	8	18.6878	.01662

Simple effects	G	df	$G_{\text{corrected}}$	p-value
Total	533.187	14		
Source of financing (A)	206.568	4	206.196	$< .00001$
Within Medicare/Medicaid ($B\|a_1$)	18.6486	2	18.6261	.00009
Within Medicare ($B\|a_2$)	54.6429	2	54.5777	$< .00001$
Within Medicaid ($B\|a_3$)	54.2676	2	54.2023	$< .00001$
Within Self-paid ($B\|a_4$)	61.9825	2	61.9079	$< .00001$
Within Personal insurance ($B\|a_5$)	137.077	2	136.912	$< .00001$

Note. The top part shows the omnibus analysis with main effects and their interaction; The correction factors c_W are 1.0018, 1.0012 and 1.0084 for the factor *Source of financing*, the factor *Residency program*, and for the interaction respectively. The bottom part shows the analysis with simple effect for each source of financing. The correction factor for the simple effects are all identical, $c_W = 1.0012$.

An example from Landis et al. (2013)

Consider the Landis et al. (2013) data examined in Sharpe (2015) and based on a 5×3 frequency table. Landis et al. were interested in different modalities of care in a family medicine residency program. They compared a Collocated Behavioral Health service (CBH) with a Primary-Care Behavioral Health service (PBH) and a Blended Model (BM). They also considered how a patient's care was financed: Medicare (MC), Medicaid (MA), a mix of Medicare/Medicaid (MC/MA), personal insurance (PI), or self-paid ($P). The data are presented in Table 1. We also illustrate the counts in Figure 1, after reordering the financing sources to better see the trends (how the error bars were obtained will be discussed in the next section).

As hinted at by Figure 1, the counts are generally increasing from Medicare/Medicaid to private insurance. Also, CBH (red line) seems to have intermediate counts in most situations. There might be two exceptions to this general finding: Recipients of Medicare/Medicaid are equally numerous to opt for PBH than they are for CBH services. A similar exception seems to occur for Medicaid.

Following the indications given in the Appendix, we conducted a standard, omnibus, ANOFA (main effects and interaction); it was suspected from the Figure that the interaction would be statistically significant. In Supplementary materials on OSF at https://osf.io/q3yem/, we provide the computations made using R. We first report the results as if this was a typical research report, and then we highlight a few key properties of the analyses.

As seen in Table 2 (top part), the interaction is statistically significant ($G(8) = 18.85$, $G_{\text{corrected}}(8) = 18.69$, $p = .0166$). To untangle the interaction, two options are possible: decompose the interaction effects into interaction components (when there are 2 or more df at play), or run simple effects. We present these two options in turn as illustrations of what ANOFA can perform.

First option: Decomposition of the interaction effect. The data can be examined by decomposing the 8 df into two orthogonal interaction effects, one comparing CBH to PBH along all the levels of the financing sources (hence, a 2×5 interaction with $df = (2-1) \times (5-1) = 4$), and another one comparing jointly CBH and PBH to BM along all the levels of financing sources (here again a 2×5 interaction with $df = 4$).

Summing the two CBH and PBH program frequencies vs. BM and then running an ANOFA, we get for this interaction, $G(4) = 3.7340$, $G_{\text{corrected}} = 3.7063$, $p = 0.4472$. As for the analysis of CBH vs. PBH, we find $G(4) = 15.1112$, $G_{\text{corrected}} = 14.9997$, $p = 0.0047$. This second comparison is obviously the one responsible for the global interaction significance. The uncorrected Gs are additive, the two G (15.1112 and 3.7340) yielding 18.8452, the G of the global interaction. This additivity shows that the G-test complies numerically with the independence (or formal orthogonality) of the calculations, contrarily to the chi-square test.

Second option: Examination of the simple effects. As an alternative approach, we examine the simple main effects within each source of financing (see Jaccard & Guilamo-Ramos, 2002, for the nomenclature). The results

Table 3 ■ ANOFA results of the data from Table 1 with the main effects broken down into contrasts.

Source of variation	G	df	$G_{\text{corrected}}$	p-value	
Total	533.187	14			
Source of financing (A)	206.568	4	206.196	$<$ .00001	
Service for MC/MA ($B\|a_1$)	18.6486	2	18.6261	.00009	
(PBH & CBH) vs. BM	18.6216	1	18.5992	.00002	
PBH vs. CBH	0.02703	1	0.02700	.86953	n.s.
Service for Medicare ($B\|a_2$)	54.6429	2	54.5771	$<$ .00001	
(PBH & CBH) vs. BM	43.4467	1	43.3944	$<$ .00001	
PBH vs. CBH	11.1962	1	11.1827	.00083	
Service for Medicaid ($B\|a_3$)	54.2676	2	54.2023	$<$ .00001	
(PBH & CBH) vs. BM	52.5897	1	52.5264	$<$ .00001	
PBH vs. CBH	1.67792	1	1.67590	.19561	n.s.
Service for \$P ($B\|a_4$)	61.9825	2	61.9079	$<$ .00001	
(PBH & CBH) vs. BM	43.4467	1	43.3944	.00002	
PBH vs. CBH	18.5358	1	18.5134	.00002	
Service for PI ($B\|a_5$)	137.077	2	136.912	$<$ .00001	
(PBH & CBH) vs. BM	89.7868	1	89.6787	$<$ .00001	
PBH vs. CBH	47.2905	1	47.2335	$<$ .00001	

are listed in the second part of Table 2. As seen, the services are significantly different and that for (1) Medicare/Medicaid condition where the frequencies 19, 18 and 2 are compared: $G(2) = 18.6$; (2) Medicare conditino where frequencies 24, 53 and 3 are compared: $G(2) = 54.6$; (3) Medicaid condition where frequencies 44, 57 and 5 are compared: $G(2) = 54.3$; (4) self-paid condition where frequencies 20, 57, and 3 are compared: $G(2) = 62.0$; and finally, (5) personal insurance condition where frequencies 63, 165 and 20 are compared: $G(2) = 137.1$ (all $p < .00001$).

To better characterize the interaction noted previously, we further decomposed each simple main effect into two simple main effect contrasts having 1 degree of freedom each. Note that we are multiplying the analyses in this example for illustrative purposes only. Within every source of financing, we compared CBH to PBH in a first contrast, and these two merged to BM in a second contrast. For example, in the Medicare/Medicaid class, the frequencies 19 and 18 are compared on the first contrast, whereas 18.5 (the combined frequency for two classes containing 19 and 18 observations) is compared to 2 on the second contrast. The results are given in Table 3. All are highly significant except two: The difference between PBH and CBH is not significant at the .05 level for Medicare/Medicaid and for Medicaid. These two exceptions pinpoint the locus of the interaction.

This ANOFA is based on a total G statistic which is decomposed. As a check, note that, in Table 2 (top part), the main effects Gs + interaction G totalize G_{Total}. If instead the Modality of care effects and the interaction are decomposed into the associated main effects as done in Table 2 (bottom part), again the Gs of all these effects totalize G_{Total}. Finally, if a main effect contrast is further decomposed into orthogonal main effect contrasts, the sum of those contrasts' G equals the global main effect G; the same equality also applies to the df. Thus, all along, we ran tests obeying strict numerical additivity.

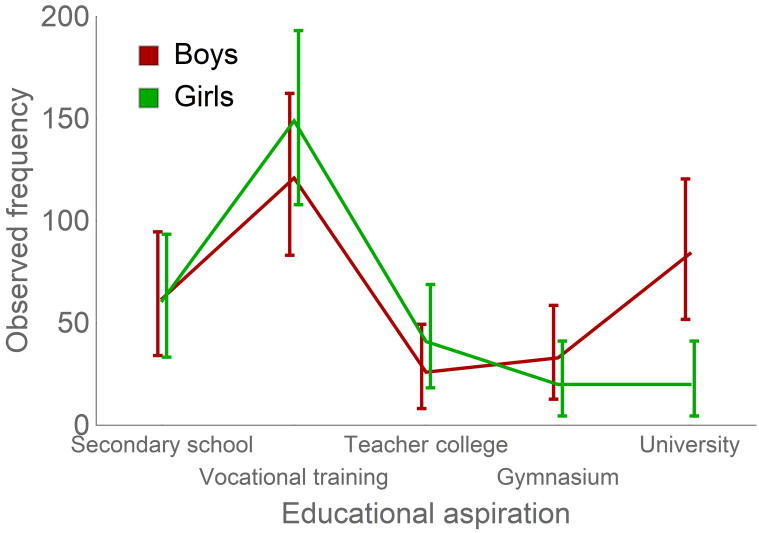
One needs to read Sharpe (2015) to realize that the ANOFA framework is far more potent to describe the pattern found in these data than any of the past attempts. Once more, the logic underlying ANOFA is the same as the one underlying ANOVA, a framework that is familiar to many readers.

An example from Light and Margolin (1971)

Light and Margolin (1971) reanalyzed unpublished data they attributed to Lesser and colleagues in which they examined educational aspiration of a large sample of $N = 617$ adolescents. The participants are classified by their gender (2 levels) and by their educational aspiration (complete secondary school, complete vocational training, become college teacher, complete gymnasium, or complete university; 5 levels). The data are therefore from a 2×5 design. The counts are for Boys: $62, 121, 26, 33, 84$; and for Girls: $61, 149, 41, 20, 20$. The frequency data are illustrated in Figure 2.

As seen in the Figure, the two groups of teenagers had nearly identical educational aspiration except for their aspiration to go to university. This suggests an interaction, confirmed by an ANOFA analysis. The omnibus ANOFA and the decomposition into five simple main effects along vocational aspiration are given in Table 4. The interaction is

Figure 2 ■ The Lesser et al. observed frequencies (unpublished) as a function of educational aspiration and gender. Error bars shows the difference-adjusted 95% confidence intervals of the observed frequencies.



important ($G_{\text{corrected}} = 49.55$, $p < .0001$). The largest discrepancy between boys and girls is for university aspiration ($G_{\text{corrected}} = 42.33$, $p < .0001$).

To characterize the magnitude of the effects, it is possible to compute effect sizes (how this is done will be described below). The eta square is a method to quantify the proportion of explained variance by each factor which is often found along ANOVA analyses. The eta square (η^2) for the gender effect is 0.02146 (2.146% of the variance is explained by this factor), a result also found in Light and Margolin (1971) using a different methodology. The vocational aspiration factor explains 14.86% of the variance, but the interaction explains more, i.e. 22.83%. This last effect size is qualified as a very large effect by some researchers (e.g., Cohen, 1992).

An example from Ries and Smith (1963)

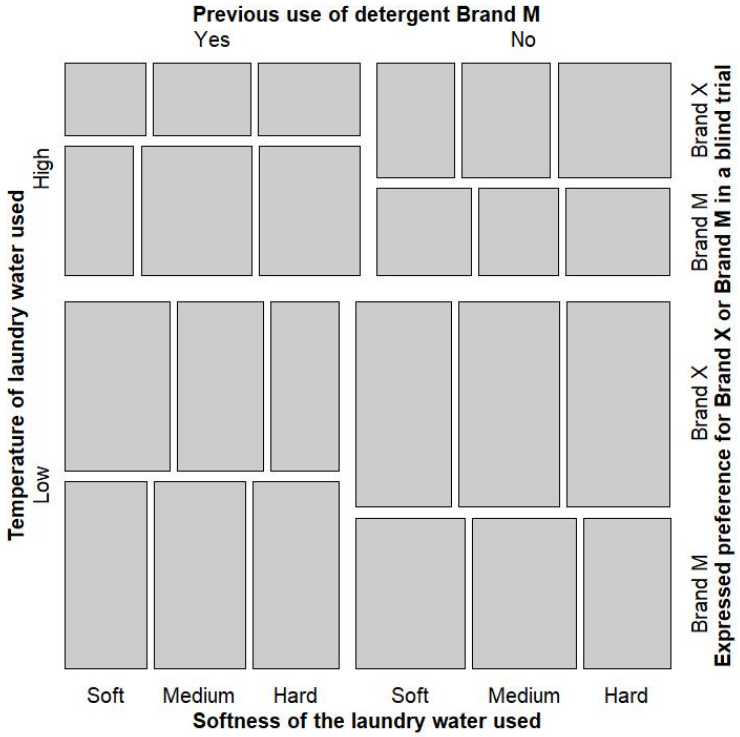
As a last example, we consider the Detergent data initially published in Ries and Smith (1963) found in the R package library `vcdExtra` (Friendly, 2023, dataset `Detergent`). In this example, consumers are classified on four factors: *Softness of water used* (3 levels: soft, medium or hard), *Expressed preference for brand M or X after blind test* (2 levels: Brand M or Brand X), *Previously used brand M* (2 levels: yes or no), and *Temperature of landry water* (2 levels: hot or cold). It is therefore a $3 \times 2 \times 2 \times 2$ design with 24 cells.

Table 4 ■ ANOFA results of the data from Figure 2.

	G	df	$G_{corrected}$	p
Total	266.8894	9		
Aspiration (A)	215.0163	4	214.6684	< 0.0001
Gender (B)	1.9865	1	1.9849	0.1589
$A \times B$	49.8867	4	49.5554	< 0.0001
Decomposition of the interaction and Aspiration effects:				
Gender within Secondary school	0.0081	1	0.0081	0.9282
Gender within Vocational training	2.9089	1	2.9078	0.0881
Gender within Teacher college	3.3868	1	3.3855	0.0658
Gender within Gymnasium	3.2214	1	3.2201	0.0727
Gender within University	42.3478	1	42.3307	< 0.0001

Note. Top part shows the omnibus analysis with main effects and their interaction; the bottom part shows the analysis of the simple effect for each vocational aspiration.

Figure 3 ■ A mosaic plot (Meyer et al., 2006) of the `Detergent` dataset showing the frequencies in each of the 24 cells in the design. Larger rectangles indicates larger frequencies. The factors are read clockwise, starting with the factor on the left size of the plot.



The data can be seen and the plots generated using the Supplementary materials on OSF at https://osf.io/q3yem/. We illustrate the frequencies using a mosaic plot in Figure 3 (Meyer et al., 2006). In a mosaic plot, the area of the squares is proportional to the frequencies in the cells. From that plot, we see that roughly two-thirds of the participants use low temperature laundry water. Regarding *Softness of water*, it seems to be divided evenly between the three levels, suggesting a lack of difference on that factor. Finally, note that the preferred use of Brand M is more than half only when participants were previously using Brand M. This last result suggests an interaction between these two factors. These intuitions will be confirmed by an ANOFA.

The ANOFA analysis is given in Table 5. As seen, the 4-way interaction is not significant ($G(2) = 0., p > .999$). The computations returned a G below zero due to a cumulative rounding error. Examining the four 3-way interactions, none are significant (all $p > .05$). There is one strongly significant two-way interaction involving *Expressed preferences for Brand M or Brand X* and whether *Brand M was used in the past* ($G_{\text{corrected}}(1) = 18.79, p < .001$). Another interaction is significant ($p = .046$) but this last effect is small (effect size below 0.01) and reaching significance only because the sample is quite large ($N = 1008$), so the practical significance of this result is limited and we may ignore it. Finally, the main effect of *Water temperature* is the only significant main effect ($G_{\text{corrected}}(1) = 66.86, p < .0001, \eta^2 = 0.067$).

We show in Figure 4 these two results, the main effect on the left and the interaction on the right. We see that the interaction is a near cross-over effect, with a confirmation of the preferences as those who previously used a certain brand still prefer that brand after a blind test.

On the OSF web site, https://osf.io/q3yem/, you can find the R code to analyze this example.

Various authors have proposed alternative methods. On OSF, you will find the code to perform a GLM analysis with a Poisson link function (as suggested by Agresti, 2013). As you will see, whereas ANOFA analyses effects, indicating their significance, GLM estimates parameters. For this example, there are 24 cells ($3 \times 2 \times 2 \times 2$) so that a GLM model may have up to 24 free parameters, all relative to a baseline

Table 5 ■ ANOFA results of the data from Figure 3.

	G	df	$G_{corrected}$	p	η^2
Total	118.6269	23			
Softness of water (A)	0.5015	2	0.4787	.7871	0.0007
Expressed preference (B)	0.0635	1	0.0580	.8097	0.0001
Previously used (C)	1.9212	1	1.7545	.1853	0.0017
Water temperature (D)	73.2121	1	66.8559	<.0001**	0.0667
$A \times B$	0.3952	2	0.3774	.8281	0.0011
$A \times C$	1.0751	2	1.0263	.5986	0.0030
$A \times D$	6.0991	2	5.8223	.0544	0.0185
$B \times C$	20.5815	1	18.7946	<.0001**	0.0367
$B \times D$	4.3616	1	3.9829	.0460*	0.0085
$C \times D$	1.2531	1	1.1443	.2847	0.0024
$A \times B \times C$	5.2201	2	4.9832	.0828	0.0296
$A \times B \times D$	0.0701	2	0.0669	.9671	0.0004
$A \times C \times D$	1.6757	2	1.5997	.4494	0.0103
$B \times C \times D$	2.2265	1	2.0332	.1539	0.0090
$A \times B \times C \times D$	−0.0294	2	−0.0281	>.9999	−0.0004

Note. *: $p < .05$, **: $p < .01$. The G statistic of the 4-way interaction is negative owing to rounding errors; G statistics cannot be negative.

condition chosen arbitrary by the software. (for example, the effect of high temperature relative to low temperature included in the baseline is 0.69, $p = .0011$ whereas the effect of low temperature relative to high temperature included in a different baseline is $+1.10, p < .00001$; see OSF script for more). From there, finding which parameters to remove to get a parsimonious model is difficult. Fienberg (2007), following Goodman (1971), suggests a stepwise approach, adding and removing parameters based on the model fits. However, Flom and Cassell (2007) argue against this technique which is too dependent on random fluctuations.

We also use the alternative analysis based on GLM with a multinomial link (proposed by Venables & Ripley, 2002). This model returns hundreds of estimates, with no clear indications how to locate a more parsimonious one from the initial estimations.

What these alternative approaches illustrate is that they are not geared toward identifying effects. ANOFA, on the other hand, examines only effects, returning their significance. Among other beneficial consequences, it is easy to verify that the results from an ANOFA analysis are totally unaffected by a change in the baseline condition.

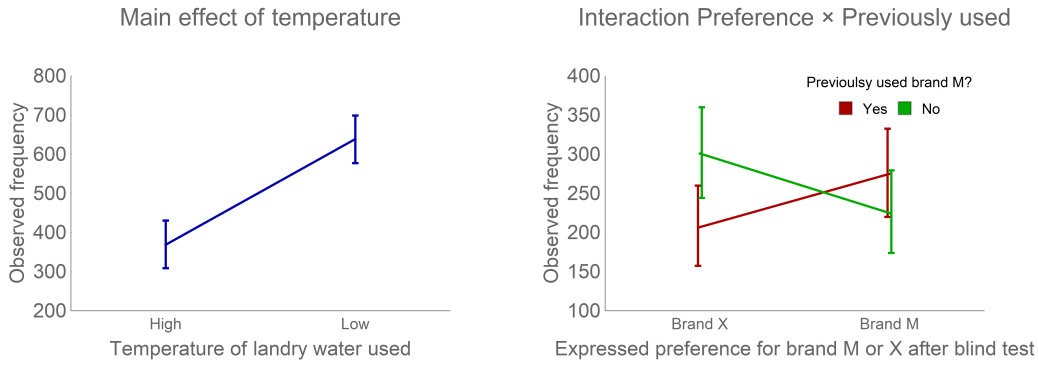
Error bars in plots

Error bars are a most useful addition to any summary statistics plot. However, the G statistic being a global lack-of-fit measure, it is not possible to assign "pieces of misfit" to every frequency category and, therefore, the G value cannot be used to derive specific confidence intervals. We advocate here an approach based on the pivot method developed by Clopper and Pearson (1934) and given in an analytic form in Leemis and Trivedi (1996). Such confidence intervals are commonly non-symmetrical around the estimate; they are also exact or conservative, the width of the interval tending to be too wide when the frequencies are small (Chen, 1990).

Given a total sample size N, an observed frequency n is used to get lower and upper confidence bounds around the proportion $\hat{\pi} = n/N$ with the formula:

$$\hat{\pi}_{\text{low}} = \left(1 + \frac{N-n+1}{nF_{1-\alpha/2}(2n, 2(N-n+1)}\right)^{-1} < \pi < \left(1 + \frac{N-n}{(n+1)F_{\alpha/2}(2(n+1), 2(N-x)}\right)^{-1} = \hat{\pi}_{\text{high}}$$

Figure 4 ■ Illustration of the main effect of *Temperature* on frequencies (left) and of the interaction of Brand used × Previously used Brand M on frequencies (right) in the Detergent dataset illustrated in Figure 3.



in which $F_{1-\alpha}$ denotes the $100(1-\alpha)$ % quantile of an F distribution where $1-\alpha$ is the desired coverage of the interval, often 95%. The interval $\{n_{\text{low}}, n_{\text{high}}\} = N \times \{\hat{\pi}_{\text{low}}, \hat{\pi}_{\text{high}}\}$ is then used as a $100(1-\alpha)$ % confidence interval of the observed frequency n which can be used to compare one frequency to an expected or theoretical frequency. Such an unadjusted confidence interval is termed a stand-alone confidence interval (Cousineau et al., 2021). More commonly, we wish to compare an observed frequency to another observed frequency, thus a difference-adjusted confidence interval is sought. To obtain a difference-adjusted confidence interval, it is required to multiply the interval width by 2,

$$n^*_{\text{low}} = 2(n - n_{\text{low}}) + n$$
$$n^*_{\text{high}} = 2(n_{\text{high}} - n) + n$$

where the asterisk denotes difference-adjusted confidence interval limits.[1] Thus, the interval $\{n^*_{\text{low}}, n^*_{\text{high}}\}$ is the difference-adjusted $100(1-\alpha)$ % confidence interval (Baguley, 2012). The difference-adjusted confidence intervals allow comparing the frequencies pairwise rather than to a theoretical frequency. Just keep in mind that precise inference is only warranted using the relevant G statistic. The error bars were obtained in Figures 1, 2, and 4 using `superb`, a general-purpose tool for generating summary plots with error bars for R (Cousineau et al., 2021). The scripts are found on OSF at https://osf.io/q3yem/.

Better confidence intervals can be obtained knowing that the proportions are not distinct but must sum to one. Such intervals are called *simultaneous confidence intervals* (SCI; Glaz, 1999; Wang, 2000; Hou & Tai, 2003). Hou and Tai (2003) performed a review of these techniques. These improved confidence intervals will be exact or conservative, but when they are conservative, they will be less conservative than the Clopper-Pearson confidence intervals used here, and therefore slightly shorter. Informal comparison found the best SCI to be about 5% shorter than the Clopper and Pearson intervals. Considering that each SCI requires extensive calculations (taking many seconds), we favored the Clopper-Pearson intervals.

Effect size and statistical power for the ANOFA test

The effect size proposed by Cohen (1992) for chi-square family variables is called Cohen's w; it is inspired by the Pearson's chi-square formula. There is however two limitations to the w formula. First, it assumes that the changes in predicted frequencies are additive. This poses a problem for small frequencies, as subtracting a quantity from these frequencies could bring some to negative values. Second, Cohen's w is one more effect size measure: there are already many, and so whenever possible, we should seek to stick to the existing ones. As it turns out, one well-known effect size measure is suitable in the present situation. Indeed, we show hereafter that an effective measure of effect size for the study of frequencies is Cohen's f^2 and its related η^2 (Cohen, 1992). These effect size measures are commonly used in ANOVA settings where they represent the ratio of the effect variance onto the error variance and the propor-

[1] The reason for the multiplication by 2 is two-fold. First, to obtain a difference-adjusted confidence interval (CI), it is necessary to multiply the CI width by $\sqrt{2}$ (under the assumption of homogeneous variance). Second, the observed classifications are correlated and this correlation equals $-1/(C-1)$ where C is the number of class (Cousineau, 2019). As this CI is meant for pair-wise comparisons, C is replaced by 2, resulting in a second, correlation-based, correction of $\sqrt{1-r} = \sqrt{1-(-1/(2-1))} = \sqrt{2}$. To better understand the correlation, consider that if a participant is in a certain class, then he is not in the other classes. Coding class membership with 1 and 0, a perfect correlation of 1 ensues. Both corrections to the CI width are multiplicative.

tion of effect variance onto the total variance, respectively.

In the present situation, we know that the corrected G statistics follows approximately a χ^2 distribution, from which ensues that G divided by its df follows a F distribution with infinite denominator degrees of freedom (e.g., Forbes et al., 2010, p. 71). The standardized F, that is, F divided by the number of observations per group (or its harmonic mean $\tilde{n}$ when class sizes are unequal, which is almost certainly the case with frequencies) returns f^2 and the total sample size times f^2 is the non-centrality parameter (often noted λ) needed to assess statistical power. Once f^2 is known, it can be converted to an eta square (η^2) measure. Thus,

$$\begin{aligned} f^2 &= \frac{1}{\tilde{n}} F = \frac{1}{\nu\ \tilde{n}}\ G \\ \lambda &= N\ f^2 \\ \eta^2 &= f^2/(1+f^2) \end{aligned} \tag{3}$$

Using the data from the first example, we find that the corrected G for the main effect of *Source of financing* is 206.1956. This translates into an F statistic of 51.5489 (the G statistics divided by the degrees of freedom). Dividing this F by the harmonic mean of the marginal number of participants (which are 170, 350, and 33; that is 76.8393), we get $f^2 = 0.670866$ and $\eta^2 = 0.4015$, that is, a huge effect size. The noncentrality parameter, $\lambda = 553 \times 0.670866 = 370.989$, is again, a large noncentrality parameter.

Regarding the interaction effect, the corrected G statistic was 18.6878 for an F of 2.3360. Dividing by the harmonic mean of the number of participants per cell (8.7457), we obtain a f^2 of 0.2671, a moderate-to-large effect size according to Cohen's guidelines. The corresponding noncentrality parameter is $\lambda = 553 \times 0.2671 = 147.707$.

If we have an a priori effect size in mind (from documented or prior test results), it is possible to conduct a power analysis. For example, regarding the interaction effect above, feeding $f = \sqrt{f^2} = 0.5168$ to G*Power (Faul et al., 2009), we get that the power to detect such an interaction is above 99%. Thus, the Landis et al. (2013) study collected a sample of such a size that it had ample statistical power.

When predicted class probabilities are accessible, it is possible to determine $f^2_{\text{predicted}}$. Indeed, suppose that the predicted probability of falling in each i^{th} class is π_i. Computing the G index on these predictions, we find

$$\begin{aligned} G_{\text{predicted}} &= -2 \sum_{i=1}^{C} f_i \left(\log g_i - \log f_i\right) \\ &= 2 \sum_{i=1}^{C} (N\ \pi_i)\ \log\left(C \times \pi_i\right) \\ &= N \times f^2_{\text{predicted}} \end{aligned}$$

in which the f_i are the predicted frequencies, equal to $N \times \pi_i$, and g_i are the expected frequencies under the null hypothesis of equality, that is, $g_i = N/C$. From the last line, we derive an alternate way to compute $f^2_{\text{predicted}}$:

$$f^2_{\text{predicted}} = 2 \sum_{i=1}^{C} \pi_i \log(C \times \pi_i). \tag{4}$$

As an illustration, we consider the scenario where the observations will be classified in one of four classes ($C = 4$). Assuming that the predicted $\boldsymbol{\pi}$s are .35, .25, 25 and .15 for classes 1 to 4, we find $f^2_{\text{predicted}}$ = 0.08228. Feeding this value $f = \sqrt{0.08228} = 0.28685$ into G*Power, we are invited to recruit a total of 140 participants to reach a statistical power of 80%: G*Power recommends 35 participants per group as the software only assumes an integer number of participants per group, but with a noncentral F distribution calculator, the exact number is 34.1181 participants per group, for a total number of participants of ≈ 136. This number is slightly overestimated because, when running a power computation from a f^2, G*Power assumes that the denominator df are taken from a standard ANOVA (i.e., $P(n-1)$), which is not the case for the G statistic where denominator degrees of freedom should be infinite. Using a non-central chi-square calculator instead, we find that the recommended number of participants is $132.5 \approx 132$.

For the interested reader, we placed in the OSF site https://osf.io/q3yem/ a script, `ComputePower.R`, that automatizes this exact search.

We checked this result by running a simulation. We generated datasets from a multinomial distribution under the above π_i's as the population proportions, generating 132 observations per sample, and ran an ANOFA test with the correction factor. We repeated this 500,000 times and recorded the number of rejections of the null hypothesis of equal probabilities. We found 80.7% rejection, well in line with the theoretical power analysis.

With 136 participants, power is estimated to be 81.8% and, with 140 participants, it rises to 83.1%. Hence, in the current context, G*Power recommendations are a bit conservative compared to the exact computation.

The major benefits of using $f^2$and its related η^2 (rather than w or other effect size measures) is that it is better known, is used in other statistical procedures and is interpreted in the exact same manner and implies exactly the same power planning procedure.

A short examination of Type I error rates, specificity and statistical power

Because the ANOFA test is based on formal mathematical arguments (Mood et al., 1974), because its likelihood ratio is asymptotically chi-square distributed (Wilks, 1938) and because a correction is available for small samples (Williams,

1976), there really is no reason to question this procedure for testing inference. Nonetheless, we decided to run some Monte Carlo simulations examining three issues: (*i*) the Type I error rate, that is, the fact that, for a certain decision threshold α, the proportion of rejection of the null hypothesis under no-difference conditions should not exceed this α: (*ii*) the specificity of the test, that is, in the presence of a complex design where some of the effects may be statistically significant, will the test of other null effects still respect the type-I error rate and, finally (*iii*): the statistical power, that is, the test's ability to reject the null hypothesis if an effect is indeed present.

We first considered the case with a single classification factor. However, the Type I error rates and statistical power results did not differ from the two-classification design and, as the issue of specificity is not relevant when there is a single factor, we chose not to report these results.

Methodology

Type I error rates. To examine Type I error rates in a two-way classification design, we generated in one simulation a total of N observations that were classified in one of the $C \times R$ cells of the design using a multinomial random number generator. In this simulation, all the observations had an equal chance $1/(C \times R)$ of falling in any cell so that the null hypothesis of no difference between the cells is true. We then ran an ANOFA on the simulated data for the main and interaction effects, recording the decisions made. In distinct set of simulations, we varied the following conditions: the total sample size ($N = 50, 75, 100, 150, 200, 250, 300, 500, 750, 1000, 2000$: 11 levels), the number of levels C of the first factor (3, 4, 5, 8: 4 levels), and the number of levels R of the second factor (3, 4, 5: 3 levels), for a total of $11 \times 4 \times 3 = 132$ conditions. The simulations within a given condition were repeated 250,000 times. Each simulation was run twice, the first in which the statistical significance is calculated with no correction factor on the G statistic, the second with Williams (1976) correction.

Specificity. To test for specificity, we ran the exact same conditions as above with one exception: we added an effect on the first factor. The effect was simulated in the population by increasing by Δ_p the proportion of the first level of the first factor, and decreasing the proportion of the last level of that factor by the same amount. The effect size used was Cohen's $w = 0.1$, which was converted to Δ_p with $w/\sqrt{2\,C\,R}$ where C is the number of levels of the first factor and R, of the second factor. This corresponds to a small-to-medium effect size $f \approx 0.100$ (the subsequent decimals depend on the number of classes). The second factor is unaffected as well as the interaction, both complying with the null hypothesis being tested. Except for the variation added to the first factor, everything else is as in the Type I error rate simulations.

Statistical power. Finally, we also ran a power analysis, that is, how frequently the null hypothesis was rejected in each simulation context. To keep the number of conditions manageable, we removed the second factor and introduced instead an effect size w with three levels ($w = 0.1, 0.2$ or 0.3) which correspond to f 's of approximately 0.100, 0.20, and 0.3 depending on the number of classes (for example, for $w = 0.3$, f goes from 0.03036 to 0.3104). The first is labeled (arbitrarily) small by Cohen (1992) whereas the third is declared medium. Everything else is as in the previous simulations. In this round of simulations, we also recorded the decisions reached by the Pearson's chi-square test to compare it to those of the ANOFA test.

Results

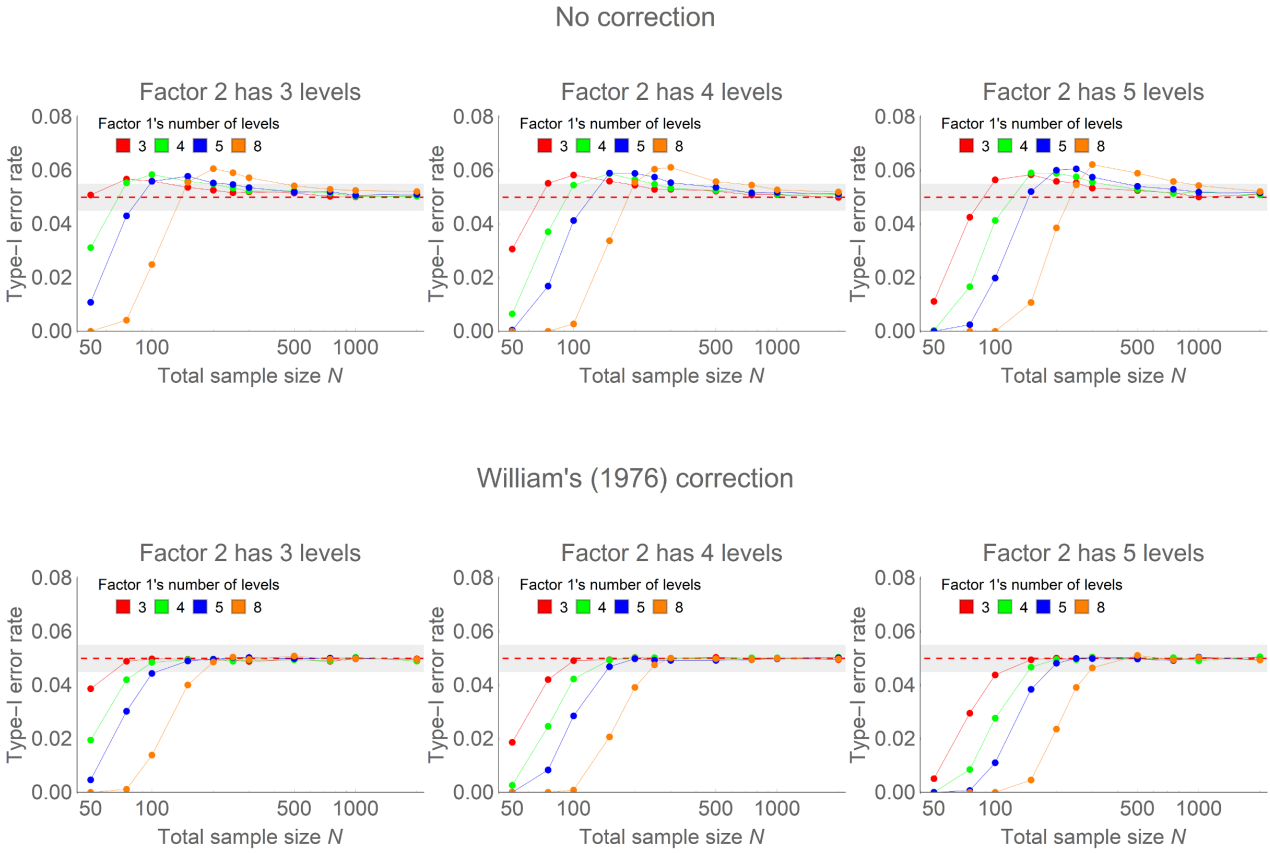
The results regarding Type I error rates are seen in Figure 5. As seen, with smaller sample sizes, the uncorrected ANOFA test tends to overshoot, having Type I error rates too large compared to the decision threshold of .05, reaching in some cells up to 6.2% of false rejection of the null hypothesis. As the samples get large, the situation is corrected, as expected since likelihood ratio tests are asymptotic tests (i.e., exact for large samples). Using Williams correction, the situation is entirely corrected, the test never exceeding the .05 limit. When the samples are too small, the test is very conservative, not rejecting the null on almost all simulations (something to be seen also in the power simulations below).

How large should a sample be depends on the number of levels, but a rough rule of thumb would suggest 10 times the number of cells in the design. The first example had 533 participants for 15 cells (a 5×3 design). This number is far sufficient to render the correction factor optional (and as seen, the corrected G statistics were barely different from the uncorrected G statistics).

The specificity results are presented in Figure 6. The rejection rate of the effect on the first factor (not reported) is increasing with sample sizes so that the test is behaving as expected (see next section on statistical power). In Figure 6, we show the rejection of the interaction effect (the rejections of the main effect of the second factor were also examined and found similar). As before, the uncorrected test may exceed the .05 limit, whereas it never does once corrected. We see in general that the error rate is now conservative, not quite reaching the .05 limit, hovering around .04. Thus, the test is not lured by the presence of an effect in the design, but it loses a bit of sensitivity on the remaining null effects, making it more 'cautious'.

The power results are presented in Figure 7. As seen, rejection rates increase steadily as the sample sizes get larger. They also rise faster when the effect sizes are larger, as

Figure 5 ■ Type I error rate for ANOFA test at the .05 level as a function of the total sample size simulated N and as a function of the number of levels in the first classification factor (color) and the number of levels in the second classification factor (panels). The top row is done without correction factor; the bottom row with Williams (1976) correction factor. The red dashed line at .05 is the decision threshold used.



expected. Indeed, the summands in the G statistics are composed of two terms, a deviance term $(\log \pi_{0i} - \log \hat{\pi}_i)$ which indicates the effect size in the i^{th} group, and a magnifier, n_i, based on the sample size in that group. When any of these is increased, the total G statistic deviates more from the G expected under the null hypothesis, increasing the chance of a correct rejection.

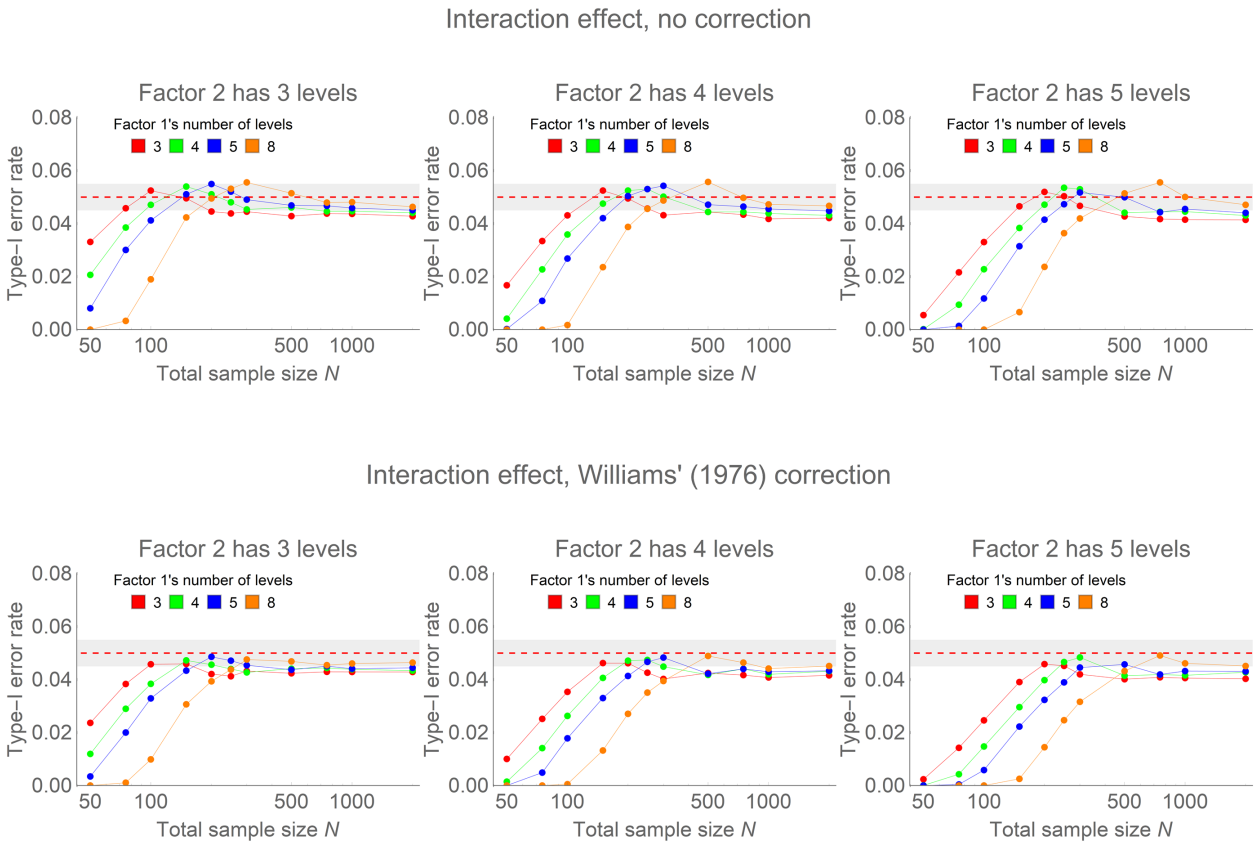
The dashed lines and + symbols show the chi-square test rejection rates. As seen, they are barely different from the rejection rates of the ANOFA test. The only place where we see a difference is when there are many groups, when the population effect size is large and when the sample is small. In this optimal situation, the ANOFA test has a power advantage reaching about 5%. We placed in the OSF site https://osf.io/q3yem/ an alternate version of Figure 7 where the effect size is plotted along the x-axis and the sample size is seen across panels (file `Figure7-AsAFunctionOfN.png`).

Discussion of the simulation study

As seen, the ANOFA test with Williams correction respects the Type I error rate dictated by the decision threshold; it gets slightly conservative when other effects are present in the dataset or when the sample is small. In terms of statistical power, it equals or slightly surpasses the chi-square test in the conditions tested herein (Hoeffding, 1965). It has been documented in the past that the performance of frequency tests can be impaired in the presence of sparse data; this is one situation we did not test here (Agresti, 2013).

Note that the reason for choosing the ANOFA test is not transparent in this section. It is therefore worth recapitulating it: The reason to use this test is its aptitude to perform complementary tests that cannot be done with a χ^2 test, such as main effects, interaction analyses, simple effects,

Figure 6 ■ Specificity of an ANOFA test examined by the Type I rejection rate of a null interaction effect when there is a main effect in the population as a function of the total sample size N and as a function of the number of levels in the first classification factor (color) and the number of levels in the second classification factor (panels). The top row is done without correction factor; the bottom row with Williams (1976) correction factor. The red dashed line at .05 is the decision threshold used.



contrasts, etc., by performing an additive decomposition of the total test statistic. To our knowledge, no other test can be decomposed in any of these ways.

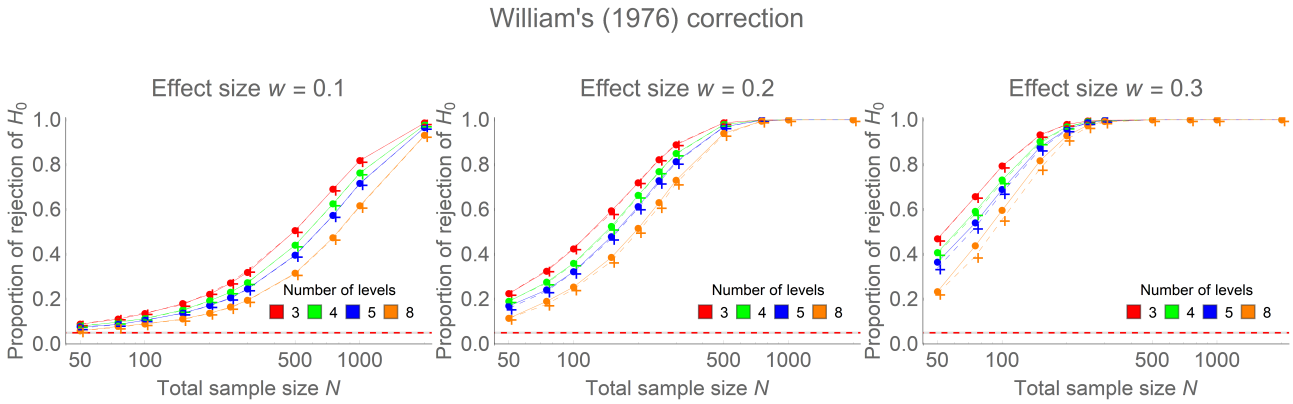
The interested reader will find in Laurencelle (2022) a thorough study of the comparative performances of the G and chi-square statistics for analyzing frequency tables: The G variable stands out as being more precise and discriminating, due to its finer (logarithmic) scale, giving it more granularity.

Discussion

We presented a general test to analyze frequencies, the ANOFA. Some aspects of this test have been documented in the past; others, such as the analyses of interactions, have never been examined whereas analysis of orthogonal contrasts was only considered in Black and Laurencelle (1987). We found the ANOFA to be very flexible, allowing a range of analyses from higher-order interactions to simple effects and contrasts to be easily performed and intuitively interpreted. Also, its ability to support power planning also speak in favor of ANOFA. One fundamental asset of ANOFA is its logic, the same as that of ANOVA: anyone acquainted with ANOVA will transfer to ANOFA without much difficulty.

The most critical characteristic of ANOFA is that the total G statistics can be decomposed additively. This characteristic implies that the decompositions are done without loss of information. None of the alternative analyses has this important characteristic. The Pearson Chi-square test, using a division by the expected frequency for its calculation, is a non-linear test statistic so that looking for useful decompositions is utopian (Shaffer, 1973a, 1973b; Sharpe, 2015). Similarly, GLM-based approaches violate the assumption of homogeneity of variances as discussed

Figure 7 ■ Power of an ANOFA and a χ^2 tests for effect size w increasing from 0.1 to 0.3 (panels) as a function of the total sample size N, and the number of of levels in the first classification factor (colors). ANOFA shown with full lines and circle symbols; χ^2 tests shown with dashed lines and + symbols. Only the results with the Williams (1976) correction are presented. The red dashed line at .05 is the decision threshold used.



in Gart et al. (1985). Informal simulations show that error variance can increase five-fold with population proportion moving away from .50. Heterogeneity of variances implies reduced statistical power (Petscher & Schatschneider, 2011). The ANOFA has no "variance" parameter and is therefore immune to this limitation. As demonstrated by Hoeffding (1965), a G statistic results in more powerful statistical test. By assembling all these characteristics in a unified framework (G statistics, additive decomposition), we obtain the ANOFA.

One reason for its attractiveness is that ANOFA analyses effects. These effects are main effects, interaction effects, or any other desired effects defined by contrasts. This is in stark contrast with alternative approaches. For example, GLM is a regression method and is focused on parameters. Consider a 5×5 design. There are three possible effects to consider: two main effects and one interaction effect. However, a GLM analysis of this dataset returns 25 parameters, all relative to a baseline condition chosen arbitrarily. Browsing through such a list of parameter estimates, it is very unlikely that the effects will reveal easily. Consequently, various, more parsimonious, models must be tested but this involves trials and errors or stepwise testing which are known to be very sensitive to the specifics of the data (e.g. Flom & Cassell, 2007).

One possible reason that the G statistics has not been popularized previously may have to do with the log transform. In the booming days of statistical methods (1925 to 1950), logarithms were cumbersome to compute, necessitating conversion tables. Nowadays, with the generalized use of computers, that reason is no longer relevant, so that the consensual *a priori* preference for the chi-square test should dwindle. When the G statistic is coupled with procedures which aim at detecting effects, we obtain the ANOFA framework described herein. Another reason that this did not occur earlier may be the strong impetus that was given to the Pearson's chi-square test in these early days so that nowadays not a single statistics textbook could be found that does not extensively promote the older test.

ANOFA has limitations. It is based on the assumption that there is a single sample of participants who are then classified according to one or more classification factors. Hence, ANOFA cannot be used to compare two distinct samples. That would be the equivalent, in the ANOVA nomenclature, of a mixed design where the participants are assigned to groups prior to be classified. ANOFA, in a sense, is only for a "one (global) sample design". Also, ANOFA does not support cluster randomized sampling. In cluster randomized sampling, the participants are not selected entirely are random; instead, chunks of participants are taken from groups, and it is the groups that are selected at random. See Cousineau and Laurencelle (2016) for the proper handling of clustered (continuous) variables in ANOVA. For similar reasons, ANOFA does not handle "repeated measures" data tables, those where each data source is measured more than once.

Using the present framework, frequencies are just a regular dependent variable that can be analyzed in the same way that continuous variables are, with the same vocabulary, and the same emphasis on effects. They have a

universally-known measure of effect size and plots with decent error bars are at our fingertips. As the French saying goes, "on serait fou de s'en passer".

References

Agresti, A. (2013). *Categorical data analysis* (3rd.). John Wiley & Sons.

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for anova. *Behavior Research Methods*, *44*, 158–175. doi: 10.3758/s13428-011-0123-7.

Black, P., & Laurencelle, L. (1987). Le test g pour les tableaux de fréquences et sa décomposition orthogonale. *Lettres Statistiques*, *8*, 97–114.

Bresnahan, J., & Shapiro, M. (1966). A general equation and technique for the exact partitioning of chi-square contingency tables. *Psychological Bulletin*, *66*, 252–262. doi: 10.1037/h0023728.

Castellan, J. (1965). On the partitioning of the contingency tables. *Psychological Bulletin*, *64*, 330–338. doi: 10.1037/h0022528.

Chen, H. (1990). The accuracy of approximate intervals for a binomial parameter. *Journal of the American Statistical Association*, *85*, 514–518. doi: 10.2307/2289792.

Clopper, C., & Pearson, E. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*, 404–413. doi: 10.1093/biomet/26.4.404.

Cochran, W. (1936). The chi-square distribution for the binomial and poisson series with small expectations. *Annals of Eugenics*, *7*, 207–217. doi: 10.1111/j.1469-1809.1936.tb02140.x.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. doi: 10.1037/h0045186..

Cousineau, D. (2019). Correlation-adjusted standard errors and confidence intervals for within-subject designs: A simple multiplicative approach. *The Quantitative Methods for Psychology*, *15*(3), 226–241. doi: 10.20982/tqmp.15.3.p226.

Cousineau, D. (2020). How many decimals? rounding descriptive and inferential statistics based on measurement precision. *Journal of Mathematical Psychology*, *97*, 102362–102362 8. doi: 10.1016/j.jmp.2020.102362.

Cousineau, D., Goulet, M.-A., & Harding, B. (2021). Summary plots with adjusted error bars: The superb framework with an implementation in r. *Advances in Methods and Practices in Psychological Sciences*, *4*(3), 1–46. doi: 10.1177/25152459211035109.

Cousineau, D., & Laurencelle, L. (2016). A correction factor for the impact of cluster randomized sampling and its applications. *Psychological Methods*, *21*, 121–135. doi: 10.1037/met0000055.

D'Ambra, L., Beh, E., & Amenta, P. (2005). Catanova for two-way contingency tables with ordinal variables using orthogonal polynomials. *Communications in Statistics—Theory and Methods*, *34*, 1755–1769. doi: 10.1081/STA-200066325.

Fagen, R., & Mankovich, N. (1980). Two-act transitions, partitioned contingency tables, and the 'significant cells' problem. *Animal Behaviour*, *29*, 1017–1023. doi: 10.1016/S0003-3472(80)80090-X.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. doi: 10.3758/BRM.41.4.1149.

Fienberg, S. (2007). *The analysis of cross-classified categorical data* (second). Springer.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A*, *222*, 309–368. doi: 10.1098/rsta.1922.0009.

Flom, P., & Cassell, D. (2007). *Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use*. North-East SAS User Group (NESUG.

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2010). *Statistical distributions*. Wiley.

Friendly, M. (2023). Vcdextra: 'vcd' extensions and additions. R package version 0.8-2. https://search.r-project.org/CRAN/refmans/vcdExtra/html/Detergent.html

Gart, J., Pettigrew, H., & Thomas, D. (1985). The effect of bias, variance estimation, skewness and kurtosis of the empirical logit on weighted least squares analyses. *Biometrika*, *72*(1), 179–190.

Glaz, S. (1999). Simultaneous confidence intervals for multinomial proportions. *Journal of Statistical Planning and Inference*, *82*, 251–262.

Goodman, L. (1971). The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics*, *13*, 33–61.

Hoeffding, W. (1965). Asymptotically optimal tests for multinomial distributions. *Annals of Mathematical Statistics*, *401*(doi), 10 1214 1177700150.

Hou, C., & Tai. (2003). A family of simultaneous confidence intervals for multinomial proportions. *Computational Statistics and Data Analysis*, *43*(1), 29–45.

Jaccard, J., & Guilamo-Ramos, V. (2002). Analysis of variance frameworks in clinical child and adolescent psychology: Issues and recommendations. *Journal of Clinical Child & Adolescent Psychology*, *31*, 130–146. doi: 10.1207/S15374424JCCP3101_15.

Landis, S., Barrett, M., & Galvin, S. (2013). Effects of different models of integrated collaborative care in a family medicine residency program. *Families, Systems and Health*, *31*, 264–273. doi: 10.1037/a0033410.

Laurencelle, L. (2022). Les distributions multinomiales, leur mesure par les tests khi2 et g , leur approximation par la loi khi-carré et l'analyse des tableaux de fréquences par le test g. *The Quantitative Methods for Psychology*, *18*, 1–20. doi: 10.20982/tqmp.18.1.p001.

Leemis, L., & Trivedi, K. (1996). A comparison of approximate interval estimators for the Bernoulli parameter. *The American Statistician*, *50*(1), 63–68.

Light, R., & Margolin, B. (1971). An analysis of variance for categorical data. *Journal of the American Statistical Association*, *66*, 534–544. doi: 10.1080/01621459.1971.10482297.

McCullagh & Nelder. (1989). *Generalized linear models* (second). Chapman & Hall.

Meyer, D., Zeileis, A., & Hornik, K. (2006). The strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software*, *17*(3), 1–48. doi: 10.18637/jss.v017.i03.

Mood, A., Graybill, F., & Boes, D. (1974). *Introduction to the theory of statistics (3rd.ed.* McGraw-Hill.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, 175–240.

Petscher, Y., & Schatschneider, C. (2011). A simulation study on the performance of the simple difference and covariance-adjusted scores in randomized experimental designs. *Journal of Educational Measurement*, *48*, 31–43.

Ries, P., & Smith, H. (1963). The use of chi-square for preference testing in multidimensional problems. *Chemical Engineering Progress*, *59*, 39–43.

Shaffer, J. (1973a). Defining and testing hypotheses in multidimensional contingency tables. *Psychological Bulletin*, *79*, 127–141. doi: 10.1037/h0033865.

Shaffer, J. (1973b). Testing specific hypotheses in contingency tables: Chi-square partitioning and other methods. *Psychological Reports*, *33*, 343–348. doi: 10.2466/pr0.1973.33.2.343.

Sharpe, D. (2015). Your chi-square test is statistically significant: Now what? *Practical Assessment Research & Evaluation*, *20*, 1–10. doi: 10.7275/tbfa-x148.

Venables, W., & Ripley, B. (2002). *Modern applied statistics with S* (4th). Springer.

Wang. (2000). Exact confidence coefficients of simultaneous confidence intervals for multinomial proportions. *Journal of Multivariate Analysis*, *99*(5), 896–911.

Wilks, S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, *9*, 60–62. doi: 10.1214/aoms/1177732360.

Williams, D. (1976). Improved likelihood ratio tests for complete contingency tables. *Biometrika*, *63*, 33–37. doi: 10.1093/biomet/63.1.33.

Appendix: ANOFA Formulas

This technical appendix provides the relevant equations for four designs: a one-way ANOFA, a two-way ANOFA, a three-way ANOFA, and a four-way ANOFA. In the two- and three-way designs, main effects and interaction effects equations are also provided, but other decompositions can be performed, such as interaction decomposition (as was done in the first illustration), simple effects (as we showed in the first two illustrations), orthogonal contrasts (as was done in the second illustration), polynomial analysis, etc.

In what follows, indices i, j, k, ... identify a level of a factor whereas uppercase roman letters $\mathrm{A}, \mathrm{B}, \mathrm{C}, \ldots$ represent the names of these factors; in italics, they represent the number of levels of these factors. Variables n_i, n_{ij}, n_{ijk}, etc. represent the observed cell frequencies in a single-factor design, a two-factor design (a table), a three-factor design (a table with multiple layers), etc.; finally, e_i, e_{ij}, e_{ijk}, etc., are the expected cell frequencies in these same designs. The variables $n_\bullet$, $n_{\bullet\bullet}$, $n_{\bullet\bullet\bullet}$, etc., denote the total observed frequencies in these designs (it is often noted in short N) whereas $n_{i\bullet}$ et $n_{j\bullet}$ are the marginal frequencies in a two-factor design. In the case of a 3-factor design, $n_{i\bullet\bullet}$, $n_{\bullet j\bullet}$ etc. are the first-order marginal frequencies, and $n_{ij\bullet}$, $n_{i\bullet k}$, etc. are the second-order marginal frequencies.

Design with a single factor

In a design with a single factor, there is a single test statistic, G_A, which is also the total test statistic, G_Total. It is obtained with

$$G_\mathrm{A} = -2\sum_{i=1}^{A} n_i \left(\log(e_\bullet) - \log(n_i)\right)$$

in which the expected frequency under the null hypothesis is given by $e_\bullet = n_\bullet/A$. This formula simplifies to $G_A = 2\sum_{i=1}^{A} n_i log(n_i) - 2\ n_\bullet\ log(n_\bullet/A)$ in the present design. This statistic follows asymptotically a chi-square distribution with $A-1$ degrees of freedom (df).

Contrasts can be tested as long as they are orthogonal. For each degree of freedom, there can be a contrast.

Designs with two factors

In a two-factor design, the total test statistic can be decomposed into two main effect statistics A and B, and an interaction statistic A × B, with

$$
\begin{aligned}
G_{\text{Total}} &= -2\sum_{i=1}^{A}\sum_{j=1}^{B} n_{ij}\left(\log(e_{\bullet\bullet}) - \log(n_{ij})\right)\\
G_{\text{A}} &= -2\sum_{i=1}^{A} n_{i\bullet}\left(\log(e_{i\bullet}) - \log(n_{i\bullet})\right)\\
G_{\text{B}} &= -2\sum_{j=1}^{B} n_{\bullet j}\left(\log(e_{\bullet j}) - \log(n_{\bullet j})\right)\\
G_{\text{A}\times\text{B}} &= -2\sum_{i=1}^{A}\sum_{j=1}^{B} n_{ij}\left(\log(e_{ij}) - \log(n_{ij})\right)
\end{aligned}
$$

in which the expected frequencies are given by $e_{\bullet\bullet} = n_{\bullet\bullet}/(A \times B)$, $e_{i\bullet} = n_{\bullet\bullet}/A$, $e_{\bullet j} = n_{\bullet\bullet}/B$ et $e_{ij} = n_{i\bullet} \times n_{\bullet j}/n_{\bullet\bullet}$. As some readers may have noticed, the structure of this last formula is also found in the chi-square test on contingency tables. These expected cell counts, as shown in Mood et al. (1974), are maximizing the likelihood of the relevant factors.

We demonstrate that the total test statistic equals the sum of the elements, that is $G_{\text{Total}} = G_{\text{A}} + G_{\text{B}} + G_{\text{A}\times\text{B}}$. Dividing the components by 2 everywhere simplifies the demonstration.

$$
\begin{aligned}
&(G_{\text{A}} + G_{\text{B}} + G_{\text{A}\times\text{B}})/2\\
&\quad= \sum_{i=1}^{A} n_{i\bullet}(\log n_{i\bullet} - \log e_{i\bullet}) + \sum_{j=1}^{B} n_{\bullet j}(\log n_{\bullet j} - \log e_{\bullet j})\\
&\qquad+ \sum_{i=1}^{A}\sum_{j=1}^{B} n_{ij}(\log n_{ij} - \log e_{ij})\\
&\quad= \sum_{i=1}^{A} n_{i\bullet}\log n_{i\bullet} - n_{\bullet\bullet}\log n_{\bullet\bullet} + n_{\bullet\bullet}\log A\\
&\qquad+ \sum_{j=1}^{B} n_{\bullet j}\log n_{\bullet j} - n_{\bullet\bullet}\log n_{\bullet\bullet} + n_{\bullet\bullet}\log B\\
&\qquad+ \sum_{i=1}^{A}\sum_{j=1}^{B}\log n_{ij} - \sum_{i=1}^{A} n_{ij}\log n_{i\bullet} - \sum_{j=1}^{B} n_{ij}\log n_{\bullet j}\\
&\quad= \sum_{i=1}^{A}\sum_{j=1}^{B} n_{ij}\log n_{ij} - n_{\bullet\bullet}\log n_{\bullet\bullet}/(AB)\\
&\quad= \sum_{i=1}^{A}\sum_{j=1}^{B} n_{ij}(\log n_{ij} - \log e_{\bullet\bullet}) = G_{\text{Total}}/2
\end{aligned}
$$

The df are summarized in this diagram:

Term	Degrees of freedom (df)
G_{Total}	$A \times B - 1$
⌙ G_{A}	$A - 1$
⌙ G_{B}	$B - 1$
⌙ $G_{\text{A}\times\text{B}}$	$(A-1)(B-1)$

If simple effects are desired (say, along factor A on the levels of factor B), then for each B-factor level, the ANOFA test statistic is given by:

$$G_{\mathrm{A|B}=j} = -2\sum_{i=1}^{A} n_{ij}\left(\log(e_{i|\mathrm{B}=j}) - \log(n_{ij})\right)$$

in which $e_{i|\mathrm{B}=j} = n_{i\bullet}/B$. As can be demonstrated, $G_{\mathrm{A|B}=1} + \cdots + G_{\mathrm{A|B}=B} = G_{\mathrm{A}} + G_{\mathrm{A}\times\mathrm{B}}$ and consequently, $G_{\mathrm{Total}} = G_{\mathrm{A|B}=1} + \cdots + G_{\mathrm{A|B}=B} + G_{\mathrm{B}}$ which is analogous to the usual decomposition in a regular ANOVA.

Designs with three factors

With three factors, the omnibus ANOFA decomposes the total test statistic

$$G_{\mathrm{Total}} = -2\sum_{i=1}^{A}\sum_{j=1}^{B}\sum_{k=1}^{C} n_{ijk}\left(\log(e_{\bullet\bullet\bullet}) - \log(n_{ijk})\right)$$

into three main effects A, B, C. Three first-order interaction (A × B, A × C, B × C) and one second-order interaction (A × B × C). The three main effects statistics are

$$\begin{aligned}
G_{\mathrm{A}} &= -2\sum_{i=1}^{A} n_{i\bullet\bullet}\left(\log(e_{i\bullet\bullet}) - \log(n_{i\bullet\bullet})\right)\\
G_{\mathrm{B}} &= -2\sum_{j=1}^{B} n_{\bullet j\bullet}\left(\log(e_{\bullet j\bullet}) - \log(n_{\bullet j\bullet})\right)\\
G_{\mathrm{C}} &= -2\sum_{k=1}^{C} n_{\bullet\bullet k}\left(\log(e_{\bullet\bullet k}) - \log(n_{\bullet\bullet k})\right)
\end{aligned}$$

in which the overall cell frequency is $e_{\bullet\bullet\bullet} = n_{\bullet\bullet\bullet}/(A \times B \times C)$, and the first-order marginal cell frequencies are $e_{i\bullet\bullet} = n_{\bullet\bullet\bullet}/A$, $e_{\bullet j\bullet} = n_{\bullet\bullet\bullet}/B$, and $e_{\bullet\bullet k} = n_{\bullet\bullet\bullet}/C$ across factors A, B, and C respectively. The first-order interactions are given by

$$\begin{aligned}
G_{\mathrm{A}\times\mathrm{B}} &= -2\sum_{i=1}^{A}\sum_{j=1}^{B} n_{ij\bullet}\left(\log(e_{ij\bullet}) - \log(n_{ij\bullet})\right)\\
G_{\mathrm{A}\times\mathrm{C}} &= -2\sum_{i=1}^{A}\sum_{k=1}^{C} n_{i\bullet k}\left(\log(e_{i\bullet k}) - \log(n_{i\bullet k})\right)\\
G_{\mathrm{B}\times\mathrm{C}} &= -2\sum_{j=1}^{B}\sum_{k=1}^{C} n_{\bullet jk}\left(\log(e_{\bullet jk}) - \log(n_{\bullet jk})\right)
\end{aligned}$$

in which $e_{ij\bullet} = n_{i\bullet\bullet} \times n_{\bullet j\bullet}/n_{\bullet\bullet\bullet}$, $e_{i\bullet k} = n_{i\bullet\bullet} \times n_{\bullet\bullet k}/n_{\bullet\bullet\bullet}$ and $e_{\bullet jk} = n_{\bullet j\bullet} \times n_{\bullet\bullet k}/n_{\bullet\bullet\bullet}$ are the second-order expected marginal cell frequencies across the pairs of factors A × B, A × C and B × C respectively. Finally, the second-order interaction is given by

$$G_{\mathrm{A}\times\mathrm{B}\times\mathrm{C}} = -2\sum_{i=1}^{A}\sum_{j=1}^{B}\sum_{k=1}^{C} n_{ijk}\left(\log(e_{ijk}) - \log(n_{ijk})\right)$$

in which

$$e_{ijk} = n_{\bullet\bullet\bullet} \times \frac{n_{ij\bullet}n_{i\bullet k}n_{\bullet jk}}{n_{i\bullet\bullet}n_{\bullet j\bullet}n_{\bullet\bullet k}}$$

is the second-order expected marginal cell frequencies.

The df for the various terms of the omnibus test are:

Term	Degrees of freedom (df)
G_{Total}	$A \times B \times C - 1$
G_{A}	$A - 1$
G_{B}	$B - 1$
G_{C}	$C - 1$
$G_{\text{A}\times\text{B}}$	$(A-1)(B-1)$
$G_{\text{A}\times\text{C}}$	$(A-1)(C-1)$
$G_{\text{B}\times\text{C}}$	$(B-1)(C-1)$
$G_{\text{A}\times\text{B}\times\text{C}}$	$(A-1)(B-1)(C-1)$

.

As before, it is straightforward to demonstrate that $G_{\text{Total}} = G_{\text{A}} + G_{\text{B}} + G_{\text{C}} + G_{\text{A}\times\text{B}} + G_{\text{A}\times\text{C}} + G_{\text{B}\times\text{C}} + G_{\text{A}\times\text{B}\times\text{C}}$. Finally, analogously to ANOVA, we can decompose any of the global test statistics into simple effects, as desired.

Design with four factors

The ANOFA is extensible at will. Here we provide the equations for 4 factors. The four factors, noted with subscripts i, j, k and l have A, B, C and D levels respectively. The observed frequencies in the cells are noted with n_{ijkl}. The omnibus ANOFA decomposes the total test statistic

$$G_{\text{Total}} = -2\sum_{i=1}^{A}\sum_{j=1}^{B}\sum_{k=1}^{C}\sum_{l=1}^{D} n_{ijk}\left(\log(e_{\bullet\bullet\bullet\bullet}) - \log(n_{ijkl})\right)$$

into four main effects

$$G_{\text{A}} = -2\sum_{i=1}^{A} n_{i\bullet\bullet\bullet}\left(\log(e_{i\bullet\bullet\bullet}) - \log(n_{i\bullet\bullet\bullet})\right)$$

$$G_{\text{B}} = -2\sum_{j=1}^{B} n_{\bullet j\bullet\bullet}\left(\log(e_{\bullet j\bullet\bullet}) - \log(n_{\bullet j\bullet\bullet})\right)$$

$$G_{\text{C}} = -2\sum_{k=1}^{C} n_{\bullet\bullet k\bullet}\left(\log(e_{\bullet\bullet k\bullet}) - \log(n_{\bullet\bullet k\bullet})\right)$$

$$G_{\text{D}} = -2\sum_{l=1}^{D} n_{\bullet\bullet\bullet l}\left(\log(e_{\bullet\bullet\bullet l}) - \log(n_{\bullet\bullet\bullet l})\right)$$

in which the overall expected cell frequency is $e_{\bullet\bullet\bullet\bullet} = n_{\bullet\bullet\bullet\bullet}/(A \times B \times C \times D)$, and the first-order marginal cell frequencies are $e_{i\bullet\bullet\bullet} = n_{\bullet\bullet\bullet\bullet}/A$, $e_{\bullet j\bullet\bullet} = n_{\bullet\bullet\bullet\bullet}/B$, $e_{\bullet\bullet k\bullet} = n_{\bullet\bullet\bullet\bullet}/C$ and $e_{\bullet\bullet\bullet l} = n_{\bullet\bullet\bullet\bullet}/D$ across factors A, B, C and D respectively. The total test statistic also includes six first-order (also called 2-way) interaction terms which are given by

$$G_{\text{A}\times\text{B}} = -2\sum_{i=1}^{A}\sum_{j=1}^{B} n_{ij\bullet\bullet}\left(\log(e_{ij\bullet\bullet}) - \log(n_{ij\bullet\bullet})\right)$$

$$G_{\text{A}\times\text{C}} = -2\sum_{i=1}^{A}\sum_{k=1}^{C} n_{i\bullet k\bullet}\left(\log(e_{i\bullet k\bullet}) - \log(n_{i\bullet k\bullet})\right)$$

$$G_{\text{A}\times\text{D}} = -2\sum_{i=1}^{A}\sum_{l=1}^{D} n_{i\bullet\bullet l}\left(\log(e_{i\bullet\bullet l}) - \log(n_{i\bullet\bullet l})\right)$$

$$G_{\mathrm{B\times C}} = -2\sum_{j=1}^{B}\sum_{k=1}^{C} n_{\bullet jk\bullet}\left(\log(e_{\bullet jk\bullet}) - \log(n_{\bullet jk\bullet})\right)$$

$$G_{\mathrm{B\times D}} = -2\sum_{j=1}^{B}\sum_{l=1}^{D} n_{\bullet j\bullet l}\left(\log(e_{\bullet j\bullet l}) - \log(n_{\bullet j\bullet l})\right)$$

$$G_{\mathrm{C\times D}} = -2\sum_{k=1}^{C}\sum_{l=1}^{D} n_{\bullet\bullet kl}\left(\log(e_{\bullet\bullet kl}) - \log(n_{\bullet\bullet kl})\right)$$

in which $e_{ij\bullet\bullet} = n_{i\bullet\bullet\bullet} \times n_{\bullet j\bullet\bullet}/n_{\bullet\bullet\bullet\bullet}$, $e_{i\bullet k\bullet} = n_{i\bullet\bullet\bullet} \times n_{\bullet\bullet k\bullet}/n_{\bullet\bullet\bullet\bullet}$, $e_{i\bullet\bullet l} = n_{i\bullet\bullet\bullet} \times n_{\bullet\bullet\bullet l}/n_{\bullet\bullet\bullet\bullet}$, $e_{\bullet jk\bullet} = n_{\bullet j\bullet\bullet} \times n_{\bullet\bullet k\bullet}/n_{\bullet\bullet\bullet\bullet}$ $e_{\bullet j\bullet l} = n_{\bullet j\bullet\bullet} \times n_{\bullet\bullet\bullet l}/n_{\bullet\bullet\bullet\bullet}$, and $e_{\bullet\bullet kl} = n_{\bullet\bullet k\bullet} \times n_{\bullet\bullet\bullet l}/n_{\bullet\bullet\bullet\bullet}$ are the second-order expected marginal cell frequencies across the pairs of factors A × B, A × C, A × D, B × C, B × D, and C × D respectively. It also includes the four second-order (also called 3-way) interaction terms given by

$$G_{\mathrm{A\times B\times C}} = -2\sum_{i=1}^{A}\sum_{j=1}^{B}\sum_{k=1}^{C} n_{ijk\bullet}\left(\log(e_{ijk\bullet}) - \log(n_{ijk\bullet})\right)$$

$$G_{\mathrm{A\times B\times D}} = -2\sum_{i=1}^{A}\sum_{j=1}^{B}\sum_{l=1}^{D} n_{ij\bullet l}\left(\log(e_{ij\bullet l}) - \log(n_{ij\bullet l})\right)$$

$$G_{\mathrm{A\times C\times D}} = -2\sum_{i=1}^{A}\sum_{k=1}^{C}\sum_{l=1}^{D} n_{i\bullet kl}\left(\log(e_{i\bullet kl}) - \log(n_{i\bullet kl})\right)$$

$$G_{\mathrm{B\times C\times D}} = -2\sum_{j=1}^{B}\sum_{k=1}^{C}\sum_{l=1}^{D} n_{\bullet jkl}\left(\log(e_{\bullet jkl}) - \log(n_{\bullet jkl})\right)$$

in which

$$e_{ijk\bullet} = n_{\bullet\bullet\bullet\bullet} \times \frac{n_{ij\bullet\bullet}n_{i\bullet k\bullet}n_{i\bullet\bullet\bullet}}{n_{i\bullet\bullet\bullet}n_{\bullet j\bullet\bullet}n_{\bullet\bullet k\bullet}},$$

$$e_{ij\bullet l} = n_{\bullet\bullet\bullet\bullet} \times \frac{n_{ij\bullet\bullet}n_{i\bullet\bullet l}n_{\bullet j\bullet l}}{n_{i\bullet\bullet\bullet}n_{\bullet j\bullet\bullet}n_{\bullet\bullet\bullet l}},$$

$$e_{i\bullet kl} = n_{\bullet\bullet\bullet\bullet} \times \frac{n_{i\bullet k\bullet}n_{i\bullet\bullet l}n_{\bullet\bullet kl}}{n_{i\bullet\bullet\bullet}n_{\bullet\bullet k\bullet}n_{\bullet\bullet\bullet l}},$$

and

$$e_{\bullet jkl} = n_{\bullet\bullet\bullet\bullet} \times \frac{n_{\bullet jk\bullet}n_{\bullet j\bullet l}n_{\bullet\bullet kl}}{n_{\bullet j\bullet\bullet}n_{\bullet\bullet k\bullet}n_{\bullet\bullet\bullet l}}$$

are the second-order expected marginal cell frequencies. Finally, it includes the third-order interaction (also called the 4-way interaction) given by

$$G_{\mathrm{A\times B\times C\times D}} = -2\sum_{i=1}^{A}\sum_{j=1}^{B}\sum_{k=1}^{C}\sum_{l=1}^{D} n_{ijkl}\left(\log(e_{ijkl}) - \log(n_{ijkl})\right)$$

in which

$$e_{ijkl} = \frac{1}{n_{\bullet\bullet\bullet\bullet}} \times \frac{n_{i\bullet\bullet\bullet}n_{\bullet j\bullet\bullet}n_{\bullet\bullet k\bullet}n_{\bullet\bullet\bullet l}n_{ijk\bullet}n_{ij\bullet l}n_{i\bullet kl}n_{\bullet jkl}}{n_{ij\bullet\bullet}n_{i\bullet k\bullet}n_{i\bullet\bullet l}n_{\bullet jk\bullet}n_{\bullet j\bullet l}n_{\bullet\bullet kl}}$$

is the third-order expected marginal cell frequencies.

The df for the various terms of the omnibus test are:

Term	Degrees of freedom (df)
G_{Total}	$A \times B \times C \times D - 1$
G_{A}	$A - 1$
G_{B}	$B - 1$
G_{C}	$C - 1$
G_{D}	$D - 1$
$G_{\text{A}\times\text{B}}$	$(A-1)(B-1)$
$G_{\text{A}\times\text{C}}$	$(A-1)(C-1)$
$G_{\text{A}\times\text{D}}$	$(A-1)(D-1)$
$G_{\text{B}\times\text{C}}$	$(B-1)(C-1)$
$G_{\text{B}\times\text{D}}$	$(B-1)(D-1)$
$G_{\text{C}\times\text{D}}$	$(C-1)(D-1)$
$G_{\text{A}\times\text{B}\times\text{C}}$	$(A-1)(B-1)(C-1)$
$G_{\text{A}\times\text{B}\times\text{D}}$	$(A-1)(B-1)(D-1)$
$G_{\text{A}\times\text{C}\times\text{D}}$	$(A-1)(C-1)(D-1)$
$G_{\text{B}\times\text{C}\times\text{D}}$	$(B-1)(C-1)(D-1)$
$G_{\text{A}\times\text{B}\times\text{C}\times\text{D}}$	$(A-1)(B-1)(C-1)(D-1)$

As before, it is straightforward to demonstrate that $G_{\text{Total}} = G_{\text{A}} + G_{\text{B}} + G_{\text{C}} + G_{\text{D}} + G_{\text{A}\times\text{B}} + G_{\text{A}\times\text{C}} + G_{\text{A}\times\text{D}} + G_{\text{B}\times\text{C}} + G_{\text{B}\times\text{D}} + G_{\text{C}\times\text{D}} + G_{\text{A}\times\text{B}\times\text{C}} + G_{\text{A}\times\text{B}\times\text{D}} + G_{\text{A}\times\text{C}\times\text{D}} + G_{\text{A}\times\text{B}\times\text{C}\times\text{D}}$ These terms can be decomposed into simple effects, as desired.

Open practices

The *Open Material* badge was earned because supplementary material(s) are available on osf.io/q3yem/

Citation

Laurencelle, L., & Cousineau, D. (2023). Analysis of frequency data: The ANOFA framework. *The Quantitative Methods for Psychology*, *19*(2), 173–193. doi: 10.20982/tqmp.19.2.p173.

Received: 11/01/2023 ~ Accepted: 01/06/2023