# FISHER F TEST WITH SIGNIFICANTLY LOW VALUES IN SMALL SAMPLES*

## *Case study*

**PhD Candidate Alina BARBU**
*Academy of Economic Studies, Bucharest*

**Abstract**

*Testing for differences between groups is usually done with the Fisher-F test. In practice, it has been known to return unusually small, but significant values. This situation is seldom covered by theory, usually being considered a fault of the model or of the analysis. This paper looks at a real-life situation where significantly low F-values were encountered, and presents the way these values were treated.*

**Key words:** Fisher F test; significantly low F-values; additivity; omitted factors; normal distribution; small samples

***

### Testing for differences between groups: theoretical aspects

Testing statistical hypotheses is a range of methods belonging to inductive statistics. Looking at data gathered in experiments or observations, the statistician can identify the shape of the repartition of random variables in the population, or the validity of certain assumptions regarding the parameters of these laws. The statistical test is the criterion for checking the validity of statistical hypotheses; it involves calculating a statistic and establishing a rule for deciding whether to accept or reject the null hypothesis, $H_0$, with a certain probability to take an inaccurate decision when confronting $H_0$ with $H_1$ (Trebici, 1983).

Several tests are available to check for differences between two groups, depending on the type of the variables and samples involved. Table 1 presents types of tests applicable in each situation.

* This article is a result of the project „Doctoral Program and PhD Students in the education research and innovation triangle". This project is co funded by European Social Fund through The Sectorial Operational Programme for Human Resources Development 2007-2013, coordinated by The Bucharest Academy of Economic Studies.

**Statistical tests for checking differences between two groups, depending on the variable and sample type**

**Table 1**

| | Dependent samples | Independent samples |
|---|---|---|
| Nominal scale | ▪ McNemar test | ▪ Fisher<br>▪ $X^2$ |
| Ordinal scale | ▪ Wilcoxon | ▪ Mann- Whitney |
| Interval scale | ▪ t test for means | ▪ Z or t test for means<br>▪ Regression<br>▪ Fisher-F test |

t, Z, Fisher, $X^2$ are parametric tests – the conclusions take into account assumptions on the distribution shape. The performance of parametric tests if the normality hypothesis (requested for t, Z, Fisher) is not fulfilled has been checked over time with Monte-Carlo simulation methods: by generating a large number of data sets which do not follow the normal distribution, the actual error of the tests can be checked. Parametric tests have been proven to have better results than was initially thought, but this does not imply ignoring the normality assumption in all cases.

When applying Fisher's F-test in linear models, most of the theoretical presentations focus on large, significant values, because these support the *rejection* of the null hypothesis, i.e. *rejecting* the assumption that the two samples are statistically equivalent. Few authors mentioned recording significantly low F-values, much lower than they should be if no relation existed between variables.

## References in statistical literature: „Small *F*-ratios: Red Flags in the Linear Model"

The only article which debates significantly low F-values is „Small *F*-ratios: Red Flags in the Linear Model" (Meek, Ozgur and Dunning, 2007). The authors note that the only previous references to F-values were in their own papers: Meek and Turner (1983) present a two-factor model analyzed as a one-factor model, with the conclusion that the significantly low F-ratio occurred because a significant factor had been left out. Meek, Ozgur and Dunning (2005) then present the preliminary results of a larger discussion on significantly low F-ratios.

Meek, Ozgur and Dunning (2007) suggested that there are several causes for significantly low F-ratios and that each situation should be treated as a „ red flag" and analyzed in detail. They present examples of three such situations:

**Interaction in randomized blocks**, exemplified by a two-factor design without replication, involving 3 colleges and 5 study programs, with 5 students

randomly chosen from each cell. The average score in the Graduate Management Admission Test (GMAT) was calculated for the students. The study program resulted in F=0.14, p=0.962 (1-p=0.038), an unusually low, but significant value. The row-column (college-study program) interaction is checked with Tukey's non-additivity test, but the resulting Tukey value did not show additivity effects.

The experiment was defined again, as a two-factor design with replication; 10 random students were chosen from each college and 2 were allocated to each program. This enables the authors to evaluate the college-program interaction itself. The new ANOVA table registered F=3.09, p=0.048 for the program and F=3.15, p=0.026 for the interaction, both significant at $\alpha$=0.05. The authors conclude that the significantly low F-value in the first model indicated a problem with the design of the experiment, but that the interaction effect may not have been the cause for the low F-value.

**Omitting significant factors**, exemplified by the number of days spent by a woman in hospital after childbirth. The design involves 4 hospitals with 9 observations each, and the one-factor analysis returns F=0.08, p=0.971 (1-p=0.029), unusually low, significant F-value. A graphical representation of the data is presented; the observations tend to cluster in two categories within each hospital. Hartlett's F max test indicates that equal variances are registered between hospitals, but the variables do not seem to follow the normal distribution (as seen in the histogram). The fact that observations cluster in distinct groups suggests that an important factor has been omitted.

The type of birth is introduced as a second factor (caesarian, natural and medically assisted) and the analysis is run again. Significant differences are found between types of birth (F=560.67, p=0.0) and, at $\alpha$=0.1, between hospitals (F=2.83, p=0.06). In this case, omitting a significant factor led to a significantly low F-value in the original model.

**Non-linearity or lack of fit** is presented through a linear model analysis on VHS sales between 1995-2004. The resulting F-value is 0.00, p=0.984 (1-p=0.016) and the graphical representation of the data clearly indicates that they follow a non-linear distribution. Using a quadratic model, the authors obtain F=41.15, p=0.0, thus demonstrating the original (linear) model was unfit to describe the available data.

Meek, Ozgur and Dunning's conclusion is that **the F-value should never be close to 0; all significant F-values close to 0 indicate a problem in the model or the analysis** and should be investigated as much as a significantly large F-value. Apart from the three causes with extended presentations, the authors also note that significantly low F-values may occur because of: violations of distributional assumptions, multicolinearity in regression and/or false data.

***

Investigating significantly low F-values started from two actual studies from medical marketing research: in two surveys, physicians were asked to evaluate the performance of pharmaceutical companies' medical representatives by rating them with marks from 1 to 10 on several attributes. Each survey sample was divided in two sub-samples depending on the physicians' specialty. The F-test was used to check for differences between specialties. Tables 2 and 3 present the mean evaluations of physicians; for confidentiality issues, the specialties and attributes cannot be stated.

**Case study 1: results on the F-test for attributes A, B, C and specialties 1 and 2**

**Table 2**

| *mean* | **Attribute A** | **Attribute B** | **Attribute C** |
|---|---|---|---|
| Specialty 1 (N=56) | 9.50 | 9.52 | 9.66 |
| Specialty 2 (N=33) | 9.52 | 9.68 | 9.67 |
| Total sample* (N=89) | 9.51 | 9.57 | 9.66 |
| *F*-test | *F=0.01; p=0.086* | *F=1.32; p=0.253* | ***F=0; p=0.037*** |

**Case study 2: results on the F-test for attributes D, E and specialties 3 and 4**

**Table 3**

| *mean* | **Attribute D** | **Attribute E** |
|---|---|---|
| Specialty 3 (N=30) | 8.79 | 8.74 |
| Specialty 4 (N=21) | 8.75 | 8.75 |
| Total sample ** (N=51) | 8.77 | 8.74 |
| *F*-test | *F=0.01; p=0.077* | ***F=0; p=0.02*** |

* Full answers; the original sample, including missing answers, has 150 respondents. We opted for eliminating missing answers in order to increase accuracy, and also because there were enough full answers (>50).

** Out of which 31 full answers; because of the small sample, missing values were replaced by the global mean in order to minimise information loss. We are aware of the implications of a high non-response rate on the validity of the conclusions, but keeping only full answers would generate too small a sample for any statistical analysis.

Two aspects can be pointed out:

- Samples are small (89 and 51 evaluations), which is a characteristic of medical surveys compared to generic „consumer" surveys.
- Evaluation on a scale from 1 to 10 is not recommended by marketing research theory, because it does not differentiate well among several attributes; in the previous examples, we can see that mean evaluations are quite similar among surveys and specialties.

Low F-values occur on attributes A, C, D and E, but they are significant on 95% probability level only on attributes C and E, indicated by the $p \leq 0.05$. These situations are unusual and can potentially be treated as „red flag" as indicated by Meek, Ozgur and Dunning (2007). Before treating these values as random occurrences, we will evaluate whether they can be explained by any of the solutions identified by these authors.

**Omitting significant factors**

There is no universal method to identify missing factors, because actual relations between variables are complex and not always easy to see. In the two studies presented, we only have one other socio-demographic variable available: locality, with three values (Bucharest, large cities, medium cities). We will test whether locality significantly influences the physicians' evaluation.

For $\alpha=0.05$, the F-test reveals:

- A significant relation between locality and evaluations on attribute A ($F=3.87$, $p=0.024$).
- No significant relation between locality and evaluations on attribute C ($F=2.45$, $p=0.09$), but it becomes significant if the confidence level is lowered. The sample is small and disproportionate: N=45 in Bucharest, N=87 in large cities, N=18 in medium cities (because the sample was not designed to be representative at regional level), therefore we do not recommend lowering the confidence level.
- No significant relation between locality and evaluations on attribute D ($F=0.44$, $p=0.345$).
- No significant relation between locality and evaluations on attribute E ($F=0.3$, $p=0.25$).

This time, the resulting F-values are „normal"; this may mean that locality is a better factor than specialty. But **because there is no general significant relation between locality and evaluations, we cannot state that omitting significant factors** was the cause for the significantly low F-values.

**Interaction (additivity)**

Additivity is the property of independent variables to interact significantly in order to influence a dependent variable. Several tests check for additivity, among them: Tukey, Mandel, Johnson-Graybill, *locally best invariant* (LBI) and Tussel. Šimečková, Šimeček and Rasch (2008) compare these five tests in order to evaluate the actual type I risk registered based on computer simulations.

Tukey's test (1949) is a common option to check for additivity effects; it detects additivity in which the interaction is directly related to the row and column effects. Šimečková, Šimeček and Rasch found it particularly accurate for this type of interaction, with less than 4% of computer simulations registering $\alpha>0.05$. The statistic of the test is a ratio between the mean squares of the interaction and error effects:

$$S_T = \frac{MS_{interaction}}{MS_{error}}$$

It is F-distributed with *1; (a-1)(b-1)* degrees of freedom, where *a* is the number of rows and *b* is the number of columns.

The initial model is one-factor, therefore interaction is not applicable. Instead, we will test whether a two-factor model (specialty-locality) can fit this situation. Additivity is not recommended for case study 2 because:

- Distributing the answers by rows (locality) and columns (specialty) generates few values in each cell: 4 cells out of 6 contain no more than 6 observations.
- At point (a) we noted that there was no significant relation between locality and evaluations on attributes D and E.

For case study 1, the answers are distributed as follows:

| *Count* | **Specialty 1** | **Specialty 2** | **Total** |
|---|---|---|---|
| Bucharest | 16 | 8 | 24 |
| Large cities | 37 | 18 | 55 |
| Medium cities | 3 | 7 | 10 |
| **Total** | **56** | **33** | **89** |

| *Mean* | **Specialty 1** | **Specialty 2** | **Total** |
|---|---|---|---|
| Bucharest | 19.81 | 9.75 | **9.79** |
| Large cities | 9.62 | 9.78 | **9.67** |
| Medium cities | 9.33 | 9.29 | **9.30** |
| **Total** | **9.66** | **9.67** | **9.66** |

The ANOVA analysis reveals that the locality effect is not significant, but very close, for $\alpha$=0.05 (F=17.03, p=0.055, $F_{critic}$=19), while specialty registers a low F-value, not significant (F=0.04, p=0.84).

The Tukey test calculated for this data is T=-3207,14, lower than the critical $F_{0,01;1,1}$=4052,18 – but not for $F_{0,05;1,1}$=161,45. Considering the size of sub-samples, the Tukey test and the ANOVA results, we can conclude that **there is not enough evidence to support the theory of interaction between variables**.

**Mis-specified model**

Looking at the means, it is unlikely that specialty and evaluation are connected by a non-linear relation. A problem may occur when applying the (parametric) F test if the distributions do not follow the normal distribution – especially in small samples, because in large samples the Central Limit Theorem ensures that the variable mean will follow the normal distribution even when the sample variable is not normally distributed.

The normal distribution is usually checked with:

**- the Kolmogorov-Smirnov test (K-S)**, which is well-known;

**- the Shapiro-Wilk W test**, increasingly popular due to its good power compared with other tests, especially in medium and small samples (Shapiro and Wilk, 1965; Shapiro and Wilk, 1968). The power of the W test is greater than K-S because it also detects deviations caused by skewness and kurtosis. The statistic of the test is:

$$W = \frac{\left(\sum_{i=1}^{n} c_i X_i\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

The coefficients $c_i$ and critical values of the W statistic can be found in tables, calculated by Shapiro and Wilk up to N=50. In its original version, the test was not extended above N=50 because the calculations were difficult, but the Shapiro-Francia extension presents coefficients up to N=99.

The test is not implemented in all statistical packages, which maintains its status as difficult to calculate.* Burdenski (2000) recommends K-S for samples higher than N=25 and W for samples lower than N=25.

K-S indicates deviations from the normal distribution for attributes A and C (89 full answers) and normality for attributes D and E (31 full answers).

* An online calculator of the Shapiro-Wilk test is available at: http://dittami.gmxhome.de/shapiro/. Data can be simply copied in and the result, along with critical values and rejection/ acceptance are automatically calculated. According to the original presentation, the calculator accepts samples between 5-50 units.

Though the samples are higher than 25, the low F values raise suspicion concerning the last two attributes. We apply the W test for attributes D and E, the results being found in Table 4.

Checking the normality of the distribution
for attributes D and E using the W test

Table 4

| | **Attribute D** | **Attribute E** |
|---|---|---|
| $W_{calc}$ | 0.834 | 0.830 |
| $W_{critic}$ for α=0.05 | 0.902 | 0.902 |
| Decision | *Not normal* | *Not normal* |

**For all attributes with low F-values, there are problems with the normality of the distribution.** Despite Burdenski's recommendation, K-S does not provide good results for sample 2 (N=31 full answers), Shapiro-Wilk performing better.

The Central Limit Theorem does not apply for these samples because their size is too small. Therefore, non-parametric tests should be used to evaluate attributes A, C, D and E:

**a) Mann-Whitney's U** is the most common non-parametric test for ordinal variables (we are using an „inferior" scale), calculated based on the sum of ranks on each sub-sample:

$$U = \max\left( n_1 n_2 + \frac{n_i\left(n_i+1\right)}{2} - R_i \right)$$, where the sum of ranks is $R_i = \sum_{j=1}^{n_i} r_{ji}$ and i=1,2.

The null hypothesis is rejected if the calculated value $U' > U_{critic}$. Table 5 presents the results of the Mann-Whitney test for attributes A, C, D and E.

**b) The median test** is another non-parametric option, being calculated based on the number of observations lower and higher than the median in each sub-sample:

$$\mathrm{X}^2 = \sum_{i=1}^{2} \frac{\left(n_{1i} - n_{1i}^{'}\right)^2 + \left(n_{2i} - n_{2i}^{'}\right)^2}{n_{1i}} \approx \mathrm{X}_1^2$$

The median test is weaker than Mann-Whitney, but is more suitable if the evaluations contain numerous extreme values (here: mark 10 evaluations).

**Testing for differences using Mann-Whitney**

Table 5

| | Attribute A | Attribute C | Attribute D | Attribute E |
|---|---|---|---|---|
| N | *89* | *89* | *31* | *31* |
| U' | *911* | *890* | *103.5* | ***124.5*** |
| ($U_{critic}$) | *(<963)* | *(<990)* | *(<128.5)* | ***(>107)*** |

For **attributes A, B, C, D we cannot claim there are significant differences** between evaluations by the medical specialties.

For attribute E, the F test did not reveal significant differences, but Mann-Whitney does. The median test is not applicable because, after eliminating values equal to the median, less than 20 values are left (20 is the minimum sample size for the median test). The actual mean values (8.74 and 8.75) do not support the existence of significant differences, therefore **for attribute E we conclude that the statistical results are inconclusive** mostly because of the small sample. We cannot accept significant differences on this attribute either.

***

At first sight, the examination is not justified because the evaluations are very close and sample sizes are quite small. The case studies are real though, and they show situations which are not covered by statistical theory. The medical field generally deals with smaller samples because the populations themselves are smaller; clearing the theoretical situation of significantly low F-values helps complete a gap in the knowledge on hypotheses testing (the potential problem of the F test in small and medium samples) and also to a better grasp of actual occurrences of low F-values.

The only theoretical paper debating the issue at large is Meek, Ozgur and Dunning (2007), who consider that significantly low F-values may indicate problems with the model and should be checked for possible causes before being considered random occurrences. The suggestions presented by these authors are adapted to the current case study. We concluded that the main cause for the significantly low F-value was the application of the F test on data not normally distributed, coming from small-medium samples. We also noted that Shapiro-Wilk's W test is more suitable than Kolmogorov-Smirnov to test normality in small samples (N<50, for safety).

Still, the issue of significantly low F-values is not entirely cleared. The small sample size itself may affect statistical tests: in this case, N=89 and N=51, while Meek, Ozgur and Dunning presented case studies with 10-50 observations. In these cases, the Central Limit Theorem cannot be applied

and this may be the main cause for the low F-values. In smaller samples, non-parametric tests (e.g. Mann-Whitney) as a general rule.

Future research in this field should focus on three main questions:

- Identifying the apparition frequency of significantly low F-values in larger samples, i.e. larger than N=100-150.;
- If significantly low F-values do occur, the normality of the distributions should be evaluated and the performance of K-S and W tests should be compared;
- If the variables are not normally distributed, the applicability of the Central Limit Theorem should be tested.

If significantly low F-values are shown to occur only in smaller samples, we can consider the exclusive application of non-parametric tests in these cases; the limit sample size should be identified, the point at which low F-values cease to occur.

**Bibliography**


**- Burdenski, T.** (2000), ***Evaluating Univariate, Bivariate and Multivariate Normality using Graphical and Statistical Procedures***. Multiple Linear Regression Viewpoints, nr 26,

**- Meek, G., Ozgur, C., Dunning, K.** (2005). ***Some implications of significantly small F-ratios***. Proceedings of the 2005 Annual National Meeting of the Decision Sciences Institute

**- Meek, G., Ozgur, C., Dunning, K.** (2007). ***Small F-ratios: Red Flags in the Linear Model***. Journal of Data Science, nr. 5

**- Meek, G., Turner, S.** (1983). ***Statistical Analysis for Business Decisions***. Boston: Houghton Mifflin.

**- Shapiro, S.S., Wilk, M.B.** (1965). ***An analysis of variance test for normality (complete samples).*** Biometrika, nr. 52

**- Shapiro, S.S., Wilk, M.B., Chen, H.J.** (1968). ***A comparative study of various tests for normality***. Journal of the American Statistical Association, nr. 63

**- Šimečková, M., Šimeček, P., Rasch, D. (**2008) ***Tests of Additivity in two-way ANOVA Models with Single Subclass Numbers***, Statistical Papers of the Union of Czech Mathematicians and Physicists – JČMF

**- Trebici, V.** (coord) (1985), ***Mică Enciclopedie de Statistică***, Bucureşti: Editura Ştiinţifică şi Enciclopedică

**- Tukey, J.W.** (1949). ***One Degree of Freedom for Non-Additivity***. Biometrics, nr. 5