E

UPOV

TWC/30/38 Rev.
ORIGINAL: English
DATE: June 21, 2012

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

Geneva

TECHNICAL WORKING PARTY ON AUTOMATION AND COMPUTER PROGRAMS

Thirtieth Session
Chisinau, Republic of Moldova, June 26 to 29, 2012

F-RATIO TEST FOR PLANT VARIETAL DISTINCTNESS WITH CATEGORICAL CHARACTERISTICS

Document prepared by experts from China

1. This document presents a method for an *F*-ratio test for distinctness of variety protection trials. This test method estimates potential differences in the sensitivity of varieties to environment effect. Two types of dummy example were conducted to demonstrate the application of the method in distinctness, uniformity and stability (DUS) test.

INTRODUCTION

2. In new variety protection trials, many of the characteristics observed are categorical. The categorical data, or class data, only can use the non-parameter method to analyze. The Chi-square test is a widely used non-parameter method. The standard formula for the chi-square statistic used in such analysis is:

$$\chi^2 = \sum \frac{(\text{Observed value of a class - Expected value of a class})^2}{\text{Expected value of a class}}$$

3. To use the Chi-square analysis for plant breeder rights' (PBR) purposes, how to arrive at certain conclusions about distinctness should be considered by formulating certain hypotheses using the classification data. One of the most important hypotheses is that the variety must be distinct on one or more characteristics from all other reference varieties on the list. Results vary from plant to plant, plot to plot and year to year and statistical criteria are required to separate genuine varietal differences from random variation, or experimental errors.

4. DUS tests, biologists wish to ascertain the relative effects of reference varieties and candidate varieties. In this paper, a method of ratio-test of two Chi-squares was put forward to compare the relative effects of candidate variety to reference variety. The former Chi-square is the Chi-square of goodness of fit of frequency distribution of candidate varieties fitting the theory frequency distribution, or the frequency distribution of reference varieties. The distribution of characteristics observed for this reference variety is considered to be the expected distribution. The latter Chi-square is the interaction between characteristics of reference variety and repeats of plot, or year. The interaction Chi-square can be considered as the heterogeneity Chi-square, or error of experiment, from the contingency table. Because the *F*-distribution is a derivative from two chi-square variables, we consider that the *F*-ratio test be used for comparing two Chi-squares from plant variety testing.

F-DISTRIBUTION AND *F*-STATISTICS

5. Statisticians have shown that the ratio of two chi-square variables follows a new distribution known as the *F*-distribution. If we have one χ^2 variable with n_1-1 degree of freedom, and another with n_2-1 degree of freedom then the ratio has an *F*-distribution with n_1-1 degrees of freedom for the numerator and n_2-1 degrees of freedom for the denominator.

6. In other words, if χ_1 and χ_2 are both chi-squares with v_1 and v_2 degrees of freedom respectively, then the statistic F belongs to *F*-distribution.

$$F(v_1, v_2) = \frac{\chi_1^2 / v_1}{\chi_2^2 / v_2}$$

7. The two parameters, v_1 and v_2, are the numerator and denominator degrees of freedom. That is, v_1 and v_2 are the number of independent pieces of information used to calculate χ_1 and χ_2, respectively.

8. The *F*-distribution provides a function for comparing the ratios of two chi-square variables associated with different source factors. In DUS test, two variables with chi-squared distribution are derived from the source of variance. The first is the heterogeneity chi-square, derives from the interaction chi-square of contingency table of characteristic-by-repeat (year). The χ^2 value can be considered as the pooled error of experiment. The χ^2 value is denoted by χ_H^2 and its degree of freedom by df_H. The second is the chi-square of goodness of fit, derives from the difference between the frequency distribution of candidate variety and the expected frequency distribution of characteristics of reference variety, and the expected frequency distribution is not the exact or theoretical distribution. We denote the chi-square of goodness of fit by χ_F^2. Similarly, we denote its degree of freedom by df_F.

STATISTICAL HYPOTHESIS OF DISTINCTNESS TEST

1. Test the heterogeneity, or interaction between characteristic and repeat (year)

H_0 is that characteristics and repeats (years) are independent, not associated or interactive. The statistic used is based on the Chi-square distribution with df_H degree of freedom. H_0 is rejected if the calculated statistic χ^2_H is greater than $\chi^2_\alpha(df_H)$ where $\chi^2_\alpha(df_H)$ is the percentile of the distribution corresponding to a cumulative probability of (1- α) and α is the significance level. If H_0 is rejected, there is interaction between the characteristic and repeat. At this time it would not be appropriate to make further distinctness test.

2. Test distinctness between candidate variety and reference variety

H_0 is that to test the hypothesis of frequency distribution of characteristics from candidate variety fitted to expected distribution of reference variety the statistic used is based on the F distribution. If the null hypothesis H_0 is true, then the statistic

$$F_0 = \frac{\chi^2_F / df_F}{\chi^2_H / df_H}$$

follows the F distribution with df_F degree of freedom in the numerator and df_H degrees of freedom in the denominator. H_0 is rejected if the calculated statistic, F_0, is such that:

$$F_0 > f_\alpha(df_F, df_H)$$

where $f_\alpha(df_F, df_H)$ is the percentile of the distribution corresponding to a cumulative probability of (1- α) and α is the significance level.

Example 1

The following data are dummy example (Table 1). It was considered as a disease scoring of two candidate varieties and four repeats of a reference variety. The scoring was on 3 class scale (data from TGP/8/1 Draft 13).

Table 1 Frequencies of Classified Categories of both Candidate and Reference Varieties

Characteristic	Reference variety				Candidate varieties	
	Repeat 1	Repeat 2	Repeat 3	Repeat 4	1	2
1	12	6	1	7	34	32
2	23	20	18	22	6	8
3	9	19	9	15	6	4
Total	44	45	28	44	46	44

1. Compute the χ^2_H and degrees of freedom of interaction of repeat by characteristic

1.1 Fill all the given information in the Table 2 and compute the row totals (R), column totals (C), and grand total (G).

Table 2 Frequencies of Classified Categories of Reference Variety

Characteristic	Repeat 1	Repeat 2	Repeat 3	Repeat 4	Total
1	12	6	1	7	26
2	23	20	18	22	83
3	9	19	9	15	52
Total	44	45	28	44	161

1.2 Compute the expected value of each of the $r \times c$ cells as:

$$E_{ij} = \frac{R_i C_j}{G}$$

Where E_{ij} is the expected value of the (i, j)th cell, R_i is the total of the ith row, C_j is the total of the jth column, and G is the grand total. For our example, the expected value of the first cell is computed as:

$$E_{ij} = \frac{R_1 C_1}{G} = \frac{26 \times 44}{161} = 7.11$$

The results for all 12 cells are shown as follows (Table 3).

Table 3 The Expected Frequencies of Classified Categories of Reference Varieties

Class	Repeat 1	Repeat 2	Repeat 3	Repeat 4
1	7.11	7.27	4.52	7.11
2	22.68	23.20	14.43	22.68
3	14.21	14.53	9.04	14.21

1.3 The χ_H^2 value is the interaction of repeat-by-characteristic in contingency table and is computed as:

$$\chi_H{}^2 = \frac{(12-7.11)^2}{7.11} + \frac{(23-22.68)^2}{22.68} + \cdots + \frac{(15-14.21)^2}{14.21}$$
$$= 11.01045$$

And the degrees of freedom, df_H, is (r-1)(c-1)= (3-1)(4-1)=6

For our example, the tabular χ^2 value with 6 degrees of freedom is 12.59 at the 5% level of significance. Because the computed heterogeneity χ_H^2 value, 11.01 is smaller than the corresponding tabular χ^2 value at 5% level of significance, the hypothesis of no heterogeneity existed cannot be rejected. Then the distinctness between candidate variety and reference variety can be compared.

2. Compute the χ_F^2 value for Candidate 1 fitting the expected distribution of reference variety.

2.1 Compute the probability associated with each class based on contingency table of reference variety Compute the row totals (R) and grand total (G), and the ratio of the total of the ith row(R_i) to the grand total (G) is the probability associated with each class (Table 4).

Table 4 Probability Distribution of Classified Categories of Reference Variety

Class	Repeat 1	Repeat 2	Repeat 3	Repeat 4	Total	Probability
1	12	6	1	7	26	0.16
2	23	20	18	22	83	0.52
3	9	19	9	15	52	0.32
Total	44	45	28	44	161	1.00

2.2 Compute the expected frequency of Candidate 1 and its Chi-square of goodness of fit to the probability distribution of reference variety (Table 5).

Table 5 Frequency Distribution of Classified Categories of Candidate Variety 1

Class	Reference variety			Candidate 1	$\frac{(O_i - E_i)^2}{E_i}$
	Total	Probability	Observed frequency	expected Frequency	
1	26	0.16	*34*	7.43	95.02
2	83	0.52	*6*	23.71	13.23
3	52	0.32	*6*	14.86	5.28
Total	161	1.00	46	46	113.53

The χ_F^2 value is the goodness of fit, and the degrees of freedom, df_F, is (r-1)= (3-1)=2

2.3 Similarly the calculated χ_F^2 for Generation 2 is 103.97 and the degrees of freedom, df_F, is also (r-1)= (3-1)=2.

3. Compute the F value, or F-ratio for testing the distinctness between candidate 1 and reference variety as:

$$F = \frac{\chi_F^2 / df_F}{\chi_H^2 / df_H}$$

Put data into the above formula, results are as Table 6.

Table 6 F-ratio Statistics and Significance p-value of Candidate Varieties

Candidate Variety	F-Ratio	Degree of freedom	p-value
1	30.94	(2,6)	0.0007
2	28.33	(2,6)	0.0009

4. Compare the computed F value with the tabular F values with f_1= df_F and f_2= df_H and make conclusions, or make conclusions by p-value. At α=0.01, the tabular value of $F_{(2,6)}$ is 13.74. The calculated distinctness F-ratio of candidate 1 is more than the tabulated $F_{(2,6)}$ value. Therefore, we reject the null hypothesis that candidate 1variety has a similar reaction to the disease as the reference variety. Similarly the calculated distinctness F-ratio for candidate 2 is greater than the tabulated F value of 9.21. Hence, the variety is also significantly different from the candidate variety 1.

Example 2 Analysis of Over Years

To take into account the effects of years, we will have three categorical variables, Year, Class and Variety and the test of distinctness will be conducted with three way contingency table. The following data in Table 7 are dummy example. We have here a case of Reference Variety 1 and Candidate 1 to demonstrate the process of statistical tests.

Table 7 Frequencies of Classified Categories of both Candidate and Reference Varieties over Three Years

Year	Characteristic	Reference variety 1	Reference variety 2	Reference variety 3	Reference variety 4	Candidate 1	Candidate 2
Year 1	1	12	6	1	7	34	32
	2	23	33	18	27	6	8
	3	9	19	9	15	6	4
Year 2	1	10	6	1	7	27	37
	2	21	23	18	26	7	10
	3	7	19	11	17	5	4
Year 3	1	12	8	1	9	27	38
	2	23	23	14	15	7	12
	3	8	16	10	15	5	3

1. Compute the χ_H^2 and degrees of freedom of interaction of year by characteristic

1.1 To get the contingency table of Characteristic × Year cell counts, we cannot consider the different varieties using cross sections of the two-way contingency table Characteristic ×Year, and we called it a Characteristic ×Year marginal table (Table 8).

Table 8 Marginal Table of Classified Category by Year

Characteristic	Year1	Year2	Year3
1	26	24	30
2	101	88	75
3	52	54	49

1.2 Compute the row totals (R) , column totals (C), and grand total (G) of marginal table and the expected value of each of the r × c cells as:

$$E_{ij} = \frac{R_i C_j}{G}$$

Where E_{ij} is the expected value of the (i, j)th cell, R_i is the total of the ith row, C_j is the total of the jth column, and G is the grand total. For our example, the expected value of the first cell is computed as:

$$E_{ij} = \frac{R_1 C_1}{G} = \frac{80 \times 179}{499} = 28.70$$

The results for all nine cells are shown in Table 9.

Table 9 The Expected Frequencies of Marginal Table

Characteristic	Year 1	Year 2	Year 3
1	28.70	26.61	24.69
2	94.70	87.82	81.47
3	55.60	51.56	47.84

1.3 The χ_H^2 value is the interaction of characteristic-by-year in contingency table and is computed as:

$$\chi_H^2 = \frac{(24-28.70)^2}{28.70} + \frac{(101-94.70)^2}{94.70} + \cdots + \frac{(49-47.84)^2}{47.84}$$
$$= 2.9630$$

And the degrees of freedom, df_H, is (r-1)(c-1)= (3-1)(3-1)=4

For our example, the tabular χ^2 value with 4 degree of freedom is 9.49 at the 5% level of significance. Because the computed heterogeneity χ_H^2 value, 2.963 is smaller than the corresponding tabular χ^2 value at 5% level of significance, the hypothesis of no heterogeneity existed cannot be rejected. Then we can compare the distinctness between candidate variety and reference variety.

2. Compute the χ_F^2 value for Candidate 1 fitting the expected distribution of reference variety1.

2.1 To compute the probability associated with each characteristic based on contingency table of reference varieties, we get first the contingency table of Characteristic ×Year. For reference variety 1, the two-way contingency table Characteristic ×Year is as Tab. 10.

Compute the row totals (R) and grand total (G), and the ratio of the total of the ith row(R_i) to the grand total(G) is the probability associated with each class.

Table 10 Probability Distribution of Classified Categories of Reference Variety 1

Characteristic	Year1	Year2	Year3	**Total**	**Probability**
1	12	10	12	34	0.272
2	23	21	23	67	0.536
3	9	7	8	24	0.192
Total	44	38	43	**125**	**1.00**

2.2 Compute the expected frequency of Candidate 1 and Chi-square fitting to the probability distribution of reference variety 1.
To compute the expected frequency of Candidate 1, we get first the contingency table of Characteristic ×Year (Table 11). For candidate 1, the two-way contingency table Characteristic ×Year is as follows (Table 12).

Table 11 Frequencies of Classified Categories of Candidate 1

Characteristic	Year1	Year2	Year3	Total
1	34	27	27	88
2	6	7	7	20
3	6	5	5	16

Table 12 Frequency Distribution of Class of Candidate 1 Fitted to Reference Variety 1

Characteristic	Reference variety 1		Total of Candidate 1		$\frac{(O_i - E_i)^2}{E_i}$
	Total	Probability	Observed Frequency	expected Frequency	
1	34	0.272	88	33.728	87.33
2	67	0.536	20	66.464	32.48
3	24	0.192	16	23.808	2.56
Total	**125**	**1.00**	124	124	122.37

And the degrees of freedom is (r-1)= (3-1)=2
The χ_F^2 value is the goodness of fit, and the degrees of freedom, df_F, is also (r-1)= (3-1)=2.

3. Compute the F value, or Distinctness F-ratio for testing the Distinctness between Candidate 1 and reference variety as:

$$F = \frac{\chi_F^2 / df_F}{\chi_H^2 / df_H} = \frac{122.37/2}{2.96/4} = 82.68$$

Compare the computed F value with the tabular F values with f_1= df_F and f_2= df_H and make conclusions. At α=0.01, the tabular value of $F_{(2,4)}$ is 21.20. The calculated distinctness F-ratio of candidate 1 is more than the tabulated $F_{(2,4)}$ value. Therefore, we reject the null hypothesis that candidate variety 1 has a similar as the reference variety 1.

When F statistic and its p value of significance level being computed by computer program, we can also use the p value to make conclusions.

COMPUTING BY PROGRAM DUST

9. DUST is a computer program for DUS test for new plant variety. It has been developed for use in China. The functions of DUST consisted of DUS tests, Outlier test, ANOVA, T-test, Fisher exact probability, and COYD with categorical characteristics. The user interface was showed as follow.

10. For demonstration data from TWC/30/29 (Table 1) were conducted by COYD with categorical characteristics. All operations in DUST are only carried out on the area of the array which you have selected (marked). If you try to run a function which expects data, and no area has been selected, you will get an error message. The area within the array can be selected by 'dragging out' the area (shadow area). Then select the *F-test of COYD for over years* From the *COYD with categorical characteristics* menu.

Year	Color	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20	C1	C2
Year 1	1 Green	0	1	0	30	33	72	3	82	52	50	0	0	0	25	0	0	0	83	54	0	30	5
	2 White	17	7	5	0	12	2	4	2	16	17	12	9	12	0	0	0	6	5	12	6	9	9
	3-5 Red	31	71	80	30	16	3	37	7	0	5	58	74	58	17	65	75	53	3	3	71	15	48
	7 Orange	52	21	20	40	39	23	56	9	32	28	30	17	30	58	35	25	41	9	31	23	46	38
Year2	1 Green	3	0	3	28	25	76	2	82	7	37	0	0	0	22	0	0	0	92	30	0	21	9
	2 White	12	8	0	4	2	4	2	0	33	9	2	8	10	10	10	0	1	1	13	18	1	5
	3-5 Red	35	77	72	30	24	2	29	5	44	12	56	69	65	11	64	55	61	1	4	63	25	46
	7 Orange	50	15	25	38	49	18	67	13	16	42	42	23	25	57	26	45	38	6	53	19	53	40

11. Because some varieties had notes with zero plants in both years, a small value 0.5 will be add to these varieties by computer program for meeting the requirement of COYD test. For our example, heterogeneity χ_H^2 value is 8.95 and its *p* value is 0.0299. At α=0.01, the hypothesis of no heterogeneity existed cannot be rejected. Then the *F* values and the *P* values for testing the hypothesis of no difference between candidate and reference varieties were calculated. The *F* values and the *P* values are showed as follow.

Class	Yr1	Yr2	Yr1	Yr2		
1	485	407	0.1115	0.1112		
2	144	147	0.0364	0.0363		
3	757	779	0.1920	0.1915		
4	619	667	0.1607	0.1603		
Heterogeneity Chi-Square=8.9520			df=3	p=0.0299		
Candidate variety	Fitted Chi-square	F Value	Degree of Freedom		P Value	P_{dif} Value†
C1 - R1	790.7789	88.3351	3	3	0.0020	0.0062
C1 - R2	2690.7275	300.5716	3	3	0.0003	0.0033
C1 - R3	943.2095	105.3626	3	3	0.0015	0.0063
C1 - R4	22.1653	2.4760	3	3	0.2380	0.6575
C1 - R5	3.3627	0.3756	3	3	0.7787	0.9224
C1 - R6	393.2898	43.9330	3	3	0.0056	0.0036
C1 - R7	440.7920	49.2393	3	3	0.0047	0.0073

C1 - R8	444.6931	49.6751	3	3	0.0047	0.0004
C1 - R9	86.6767	9.6823	3	3	0.0472	0.1361
C1 - R10	67.8746	7.5820	3	3	0.0651	0.1621
C1 - R11+0.5	5211.6370	582.1735	3	3	<0.0001	
C1 - R12+0.5	5315.4726	593.7727	3	3	<0.0001	
C1 - R13+0.5	5249.9543	586.4538	3	3	<0.0001	
C1 - R14	7.7094	0.8612	3	3	0.5474	0.8896
C1 - R15+0.5	5237.0772	585.0154	3	3	<0.0001	
C1 - R16+0.5	5408.8145	604.1995	3	3	<0.0001	
C1 - R17+0.5	5206.1159	581.5568	3	3	<0.0001	
C1 - R18	884.9295	98.8523	3	3	0.0017	<0.0001
C1 - R19	180.2143	20.1311	3	3	0.0172	0.1202
C1 - R20+0.5	5303.0751	592.3878	3	3	<0.0001	
C2 - R1	65.6178	7.3299	3	3	0.0680	0.1432
C2 - R2	237.7694	26.5604	3	3	0.0116	0.1404
C2 - R3	105.3115	11.7640	3	3	0.0363	0.2866
C2 - R4	77.6460	8.6736	3	3	0.0546	0.0522
C2 - R5	107.4157	11.9990	3	3	0.0354	0.0786
C2 - R6	1749.5812	195.4395	3	3	0.0006	<0.0001
C2 - R7	55.2089	6.1672	3	3	0.0847	0.1143
C2 - R8	912.0739	101.8846	3	3	0.0016	<0.0001
C2 - R9	134.8902	15.0681	3	3	0.0259	0.0189
C2 - R10	416.4703	46.5224	3	3	0.0052	0.0051
C2 - R11+0.5	176.6288	19.7306	3	3	0.0177	
C2 - R12+0.5	224.9730	25.1309	3	3	0.0126	
C2 - R13+0.5	192.1111	21.4601	3	3	0.0157	
C2 - R14	192.2460	21.4751	3	3	0.0157	0.0847
C2 - R15+0.5	187.5148	20.9466	3	3	0.0163	
C2 - R16+0.5	356.0433	39.7723	3	3	0.0065	
C2 - R17+0.5	180.8525	20.2024	3	3	0.0171	
C2 - R18	2448.3867	273.5006	3	3	0.0004	<0.0001
C2 - R19	1144.8876	127.8913	3	3	0.0012	0.0027
C2 - R20+0.5	218.9958	24.4632	3	3	0.0131	

† Probability values from document TWC/30/29

CONCLUSION

12. Applying *F*-ratio analysis to distinctness test of variety protection trials expands the application to categorical data. The method proposed here is similar to the analysis of variance (ANOVA) of quantitative data, which is different from the previous Chi-square test of categorical data. The method can also be applied to testing the distinctness of categorical characteristics of biology in the field of bioinformatics research.

ACKNOWLEDGEMENTS

13. This work was partially supported by the National Basic Research Program of China (Grant No. 2010CB126201). The authors thank Dr. Wang Xiao-wei for suggestions.

REFERENCE

Agresti, A. 2002: Categorical Data Analysis 2nd edition. Wiley, New York

K. Krishnamoorthy, 2006: Handbook of statistical distributions with applications, London/Boca Raton: Chapman & Hall/CRC

M. W. Birch. 1963: Maximum Likelihood in Three-Way Contingency Tables, Journal of the Royal Statistical Society. Series B (Methodological), 25(1):220-233

Patterson, H.D. & Weatherup, S.T.C. 1984: Statistical criteria for distinctness between varieties of herbage crops. J. Agric. Sci. Camb. 102, 59-68.

Sokal, R. R. and F. J. Rohlf. 1995: Biometry: the principles and practice of statistics in biological research. 3rd edition. W. H. Freeman and Co., New York

UPOV, 2009: Trial design and techniques used in the examination of distinctness, uniformity and stability, TGP/8/1 Draft 13, Geneva.

[End of document]