

PBSTAT: A WEB-BASED STATISTICAL ANALYSIS SOFTWARE FOR PARTICIPATORY PLANT BREEDING

¹Willy Bayuardi Suwarno, ²Sobir, ³Hajrial Aswidinnoor, and ⁴Muhamad Syukur

^{1,2,3,4}Department of Agronomy and Horticulture, Bogor Agricultural University
Jl. Meranti, Wing 13 Level 5, Kampus IPB Darmaga, Bogor 16680 – Indonesia

e-mail : ¹ willy@ipb.ac.id, ² sobir@ipb.ac.id, ³ hajrial@ipb.ac.id, ⁴ msyukur@ipb.ac.id

Abstract. Indonesian agriculture areas were dominated by variation in agro-ecological and socio-economic conditions implies that formal plant breeding (FPB) programs less effective compare to participatory plant breeding (PPB). However to improve genetic gain in PPB, should be supported by simple statistical program. PBSTAT was developed in order to meet the needs of simple statistical software for selection and trials in participatory breeding approach. This software is programmed using PHP scripting language, therefore can be utilized on web platform, and provides easy access for its users to do the statistical analysis. The user can use common spreadsheet software for data entry and web browser to run the analysis. The main feature of PBSTAT is analysis of variance (ANOVA) for analyzing variety trials in one location (using one factor RCBD), multi-locations, and multi-seasons-locations (combined analysis of several RCBD trials across seasons and locations). Mean differences can be compared using Tukey's HSD method. Other important feature of PBSTAT is stability analysis using Finlay-Wilkinson method. This feature is very useful for the breeder conducting several participatory trials to elucidate which genotypes are stable across environments, and which one are environment-specific.

Keywords: statistical software, combined ANOVA, broad-sense heritability, stability analysis

1. Introduction

Indonesian agriculture areas were dominated by variation in agro-ecological and socio-economic conditions implies that formal plant breeding (FPB) programs less effective compare to participatory plant breeding (PPB). PPB is defined as plant breeding program that involving researchers, farmers, and other stakeholders such as consumers, vendors, industries, extension and farmer groups (Sperling *et al.*, 2001).

One of the most important thing to consider in PPB trials is interaction of genotype and environment. The effect of the environment is therefore a distraction in the genetical analysis, and our aim will thus be to isolate it and set it on one side in the analysis rather than to make it subject of analysis in its own right, except of course where genotype and environment interact in producing their effects (Mather and Jinks, 1982). Genotypic main effects (i.e. differences in mean yield between genotypes) provide the only relevant information when genotype x environment (GE) interaction effects are absent or ignored (Annicchiarico, 2002).

In the data analysis point of view, there is many statistical analysis software existed to meet the needs of combined analysis. However to improve genetic gain in PPB, should be supported by simple statistical program. PBSTAT was developed in order to meet the needs of simple statistical software for selection and trials in participatory breeding approach. Utilizing the web technology, this software provides easy access to do PPB's combined analysis.

2. Software Features

The Platform

PBSTAT is using web platform. The reason is because we want this software to be available widely over the internet. It will make its users, especially plant breeders, can use it easily. They don't need to install this software in their PC. They just require a web browser to run PBSTAT. We have tested it with Microsoft® Internet Explorer 6 and Mozilla Firefox 3 web browser. Using web browser, just point to <http://web.ipb.ac.id/~agrohort/pbstat> to access this software.

We are using PHP: Hypertext Preprocessor, a famous web programming language to develop this software. The scripts are processed server-side, and the outputs are sent as HTML page. The software can be hosted over the internet, or distributed in CD using a packaged web server software embedded with PHP, such as Server2Go (Haber Kern, 2008).

The Data File

Here we use a sample data set from rice yield evaluation trials conducted in four locations (Aswidinnoor *et al.* 2007, with one location added). In each location, 20 genotypes are arranged in Randomized Complete Block Design (RCBD) with three replications. Yield observed at each experimental plot and then converted to ton/ha at 14% moisture content.

Field-collected data inputted in a Microsoft® Excel worksheet (Figure 1). The first row is used only for the name of factors or variables, and the following rows could be contain either labels or observation data. The name of factors or variable must be all in lowercase or uppercase, and without spacing. Note that PBSTAT is using some reserved letters or word to define the session, location, replication, and genotype factors, which is Y, L, REP, and G respectively. If we want to add another character, such as plant height and 100 g seed weight, simply input the data in the right column after YIELD.

The data file has to be saved in Microsoft® Excel 2000/XP/2003 format, with the “xls” extension. In this example, name of data file is “COMBINED RICE 4 LOC.xls”. The Excel data file can be directly imported by PBSTAT. Note that data file must be closed before imported.

	A	B	C	D
1	L	REP	G	YIELD
2	1	1	1	7.8
3	1	1	2	6.1
4	1	1	3	8.5
5	1	1	4	5.8
6	1	1	5	6.2
7	1	1	6	6.0
8	1	1	7	7.4
9	1	1	8	5.1
10	1	1	9	5.3
11	1	1	10	4.3
12	1	1	11	3.4
13	1	1	12	4.5
14	1	1	13	4.1
15	1	1	14	5.3
16	1	1	15	7.3
17	1	1	16	5.7
18	1	1	17	4.9
19	1	1	18	6.1
20	1	1	19	5.4
21	1	1	20	6.3
22	1	2	1	4.2
23	1	2	2	4.6
24	1	2	3	7.7
25	1	2	4	5.0

Figure 1. Yield data obtained from multi-locations trial

The Interface

Because of its specialized feature, the first screen of PBSTAT 1.0 software directly shows an query form for PPB’s data analysis (Figure 2). In this form, we have to browse data file (in Microsoft® Excel format), choose type of trial, and define response variable(s) according to the data file’s column name(s). Those form elements are mandatory. Moreover, we can select further data analysis, those are estimation of broad sense heritability (h^2_{bs}) and Finlay-Wilkinson stability analysis. Finally, a click on “Show” button will run the program and outputs the result.

Figure 2. Data analysis query form

The Output

Output of ANOVA presented in Figure 3. In this example, the dependent variable is YIELD. To make a “common” heritability estimation by using the Expected Mean Squares in combined analysis (Comstock and Moll, 1963; Darrah and Mukuru, 1977), G and L here are assumed as random factors. Therefore, G is tested to G*L and G*L is tested to Error (Annicchiarico, 2002). In SAS program, we have to do this way using “test h = ... e = ...” statement after MODEL in PROC ANOVA (SAS Institute, Inc., 2003). The summary of ANOVA table presented after the series of ANOVA tables. If there is more than one variable analyzed, the summary table will contains ANOVA’s summary of all variables.

Source	df	SS	MS	Counted F	Tabulated F		P Value
					5%	1%	
L	3	176.87	58.96	41.50**	2.66	3.91	0.0000
REP*L	8	11.36	1.42	1.42 ^{ns}	2.00	2.63	0.1913
G	19	177.94	9.37	1.80**	1.66	2.03	0.0000
G*L	57	296.61	5.20	5.21**	1.41	1.63	0.0000
Error	152	151.82	1.00				
Corrected Total	239	814.60					

cv = 21.11%

Summary of ANOVA			
Karakter	G	G*L	cv (%)
YIELD	**	**	21.11

* = significant at P < 0.05
 ** = significant at P < 0.01

Figure 3. Output of ANOVA

However, for precise result of combined analysis, it's suggested to do the ANOVA for each locations first, and then check the homogeneity of variance among locations using chi-square test. If the variances are homogene, we can use the pooled error mean square in combined analysis. (Gomez and Gomez, 1984; Koopmans, 1987). Unfortunately, the “automatic” ANOVA for each location using combined data set, as using “BY” statement in SAS’s PROC ANOVA (SAS Institute, Inc., 2008), has not supported by PBSTAT yet.

Below the summary of ANOVA, PBSTAT also outputs GxL means (Figure 4). The means presented in two-way tables, with the mean of each G and L showed on the right and bottom, respectively. If there is a significant effect of G, L, or GxL factor the mean number is followed by HSD letter to show the differences between means. We limit the HSD comparisons to 20 sample means, which is the same as the maximum number of treatment means in q table (May *in* Steel and Torrie, 1980). Therefore, the letters doesn't appear in this example's GxL means (Figure 4).

G	L1	L2	L3	L4	Mean of G
G1	6.67	0.77	4.00	5.03	4.12 ^{def}
G2	5.37	5.33	4.27	5.43	5.10 ^{bcde}
G3	7.13	6.03	3.23	2.60	4.75 ^{bcdef}
G4	5.50	0.73	3.33	3.73	3.33 ^f
G5	6.47	6.87	4.73	6.77	6.21 ^{ab}
G6	7.00	4.07	2.70	4.63	4.60 ^{cdef}
G7	4.83	6.97	2.70	1.83	4.08 ^{def}
G8	4.97	2.33	3.77	4.10	3.79 ^{ef}
G9	6.13	5.57	4.03	2.70	4.61 ^{cdef}
G10	6.70	4.70	2.27	3.67	4.33 ^{def}
G11	5.70	5.50	3.20	3.77	4.54 ^{cdef}
G12	6.87	6.20	5.73	4.87	5.92 ^{abc}
G13	3.87	5.10	2.60	5.13	4.18 ^{def}
G14	4.20	6.87	2.23	1.80	3.78 ^{ef}
G15	8.00	7.87	4.47	6.60	6.73 ^a
G16	5.23	7.30	3.57	4.30	4.40 ^{bcde}

Figure 4. Output of GxL means and HSD test

The estimation of broad sense heritability is presented below the GxL tables, followed by the Finlay-Wilkinson stability analysis (Figure 5). The estimation of broad sense heritability showed genetic variance (V_G), interaction between genetic and location variance (V_{GxL}), phenotypic variance (V_p), and the broad sense heritability (h^2_{bs}) which is the ratio of V_G and V_p in percent (Darrah and Mukuru, 1977). For advanced breeding lines, the higher h^2_{bs} showed the better repeatability across environments.

The Finlay-Wilkinson stability analysis presented the genotype number, followed by its yield, b_i , and SDi. Finlay-Wilkinson proposed the regression coefficient for each genotype, b_i , as a stability parameter. The observed value are regressed on environmental indices defined as the difference between the marginal mean of the environments and oer all mean. A genotype considered to be stable if its response to environment is parallel to the mean response of all genotypes in the trial (Lin *et al.*, 1985). Genotype has $b_i = 1.0$ considered dynamically stable. The b_i value greater than 1.0 expect the genotype is suitable for more favorable environments, otherwise the b_i value less than 1.0 expect the genotype is suitable for less favorable environments.

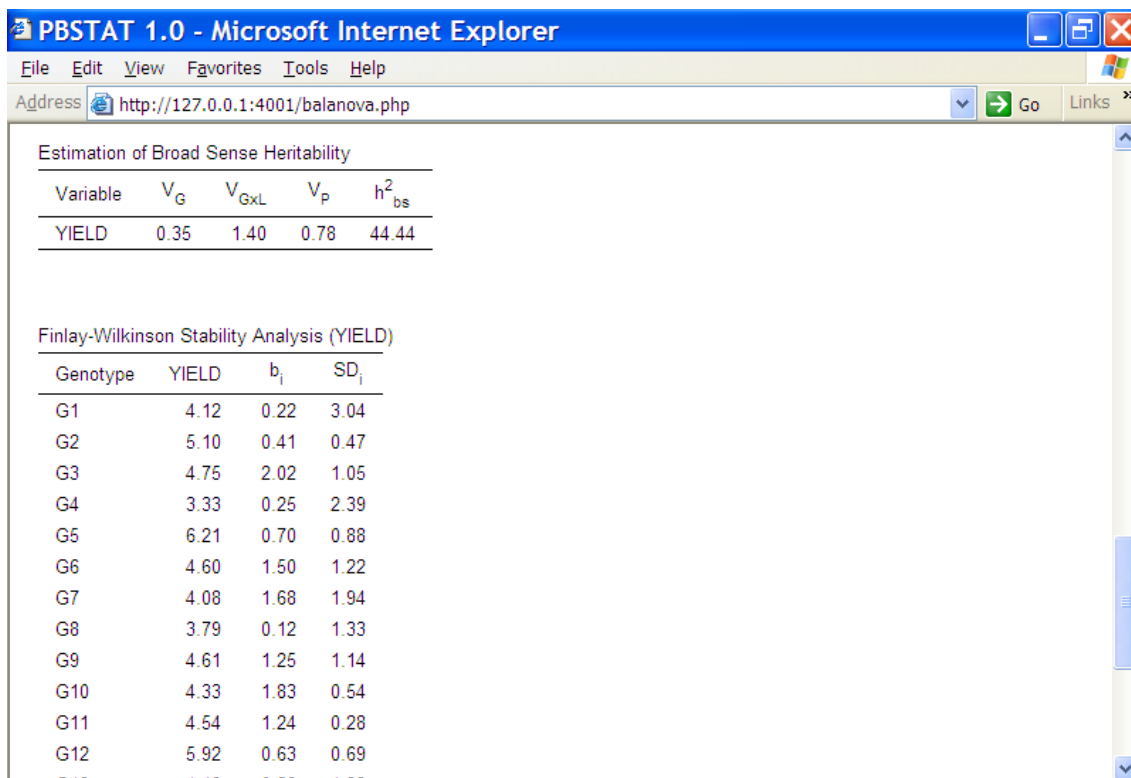


Figure 5. Output of Estimation of Broad Sense Heritability and Finlay-Wilkinson's Stability Analysis

3. PHP Functions

Data Reading and Statistical Tables

Excel data file are imported using PHP-ExcelReader utility (Tkachenko *et al.*, 2008). The F table and P-value are provided by PDL Library (Meagher *et al.*, 2008). The qtukey algorithm (Trujillo-Ortiz and Hernandez-Walls, 2003) is used to estimate the q studentized range critical value for HSD test.

PHP functions mostly used to store, read, and manipulate data are the array functions. For example, *foreach* statement is frequently used to do some calculation on array keys or values (Bakken *et al.*, 2008; Schwendiman, 2001).

Estimating Broad Sense Heritability

PHP functions to estimate broad sense heritability for multi-locations trial is presented in Figure 6. We named it *heritability2*, where the *heritability1* and *heritability3* function will estimate heritability for single and multi-seasons-locations, respectively. The *heritability2* function simply required the number of replication and location, as well as G, G*L, and Error mean square. It will calculate the genetic variance (\$var_G), interaction between genetic and location variance (\$var_GL), phenotypic variance (\$var_P), and the broad sense heritability (\$h_bs). Finally, it will return those three values in one-dimensional array. We can access the array's values and present it in tabular format like Figure 5 above.

```
function heritability2($r, $l, $ms_G, $ms_GL, $ms_E)
{
    $M1 = $ms_E; $M2 = $ms_GL; $M3 = $ms_G;
    $var_E = $M1; $var_G = ($M3-$M2)/($r*$l);
    $var_GL = ($M2-$M1)/$r;
    $var_P = $var_G + ($var_GL/$l) + ($var_E/($r*$l));
    $h_bs = ($var_G / $var_P) * 100;
    $ret = array($var_G, $var_GL, $var_P, $h_bs); return $ret;
}
```

Figure 6. PHP function for estimating broad sense heritability in multi-locations trial

Finlay-Wilkinson's Stability Analysis

We create PHP function `fw_stability` to do the Finlay-Wilkinson's stability analysis (Figure 7), based on its formula (Lin *et al.*, 1986). The function will require two parameters, `$G_value` and `$L_value`. Both of them are two-dimensional array. In `$G_value` array, `[G1][L1]` is the yield mean of Genotype 1 in Location 1. This value will be paired with the same element (`[G1][L1]`) in `$L_value` array, which contains the Location 1 mean over all genotypes. The `fw_stability` function will outputs `$fw_parameter`, which is an one-dimensional array. The array contains three values: genotype mean (`$G_mean`), b value (`$b`), and SDbi value (`$se`). Same as heritability functions, we can output those values in tabular format like Figure 5 above.

```
function fw_stability($G_value, $L_value)
{
    foreach($G_value as $key => $value)
    {
        $n = count($value);
        $G_mean[$key] = array_sum($value) / count($value);
    }

    foreach($L_value as $key => $value)
    {
        $L_mean[$key] = array_sum($value) / count($value);
    }

    foreach($G_value as $key => $value)
    {
        foreach ($value as $key2 => $value2)
        {
            $ypow[$key] += pow($value2 - $G_mean[$key], 2);
            $xpow[$key] += pow($L_value[$key][$key2] - $L_mean[$key], 2);
            $xy[$key] += ($value2 - $G_mean[$key]) * ($L_value[$key][$key2] -
                $L_mean[$key]);
        }

        $b[$key] = $xy[$key] / $xpow[$key];
        $se[$key] = sqrt((1/($n-2)) * ($ypow[$key] - (pow($xy[$key], 2)/$xpow[$key])));
    }

    $fw_parameter = array($G_mean, $b, $se);
    return $fw_parameter;
}
```

Figure 7. PHP function for Finlay-Wilkinson's stability analysis

4. References

- Annicchiarico, P. 2002. Genotype x Environment Interactions - Challenges and Opportunities for Plant Breeding and Cultivar Recommendations. FAO. Rome.
- Aswidinnoor, H., W. B. Suwarno, I. G. Cempaka, R. Indriani, W. S. Nurhidayah. 2007. Uji Daya Hasil Galur-galur Harapan Padi Sawah di Tiga Lokasi. Prosiding Seminar Nasional yang Dibiayai oleh Hibah Kompetitif. Bogor.
- Bakken, S. S., D. Beckham, G. Hojtsy, M. Jansen, J. Kosek, P. Olson, A. Tehtonik, J. Vrana, and J. v. Wolffelaar. 2008. PHP Documentation. The PHP Documentation Group.
- Comstock, R. E. and R. H. Moll. 1963. Genotype-Environment Interactions. *In*: Hanson, W. D. and H. F. Robinson (*Eds*). Statistical Genetics in Plant Breeding. NAS – NRC Pul. Symposium.
- Darrah, L. L. and S. Z. Muku. 1977. Recurrent Selection Methods for Maize Improvement: the East African Experience. East African Agriculture and Forestry Research Organization. Muguga, Nairobi. 20p.

- Gomez, K. A. and A. A. Gomez. 1984. Statistical Procedures for Agricultural Research. John Wiley & Sons. New York.
- Haberkern, T. 2008. Server2Go. <http://www.server2go-web.de>.
- Koopmans, L. H. 1987. Introduction to Contemporary Statistical Methods. Second ed. Duxbury Press. Boston. 683p.
- Lin, C. S., M. R. Binns, and L. P. Lefkovich. 1986. Stability analysis: where do we stand? Crop Sci. 26: 894-900.
- Mather, K. and J. L. Jinks. 1982. Biometrical Genetics. Third ed. 396p.
- Meagher, P., M. Hale, J. v. Kooten, M. Bommarito, J. Castagnetto, T. Lumley, K. Sigrist, D. Duehring, Taygata, G. S. Fishman, P. L'Ecuyer, R. Simard, J. C. Pezullo. 2008. PDL Library. <http://www.phpmath.com/build02/PDL/docs/download.php>.
- SAS Institute, Inc. 2003. SAS OnlineDoc® 9.1. SAS Institute, Inc. Cary, NC.
- Schwendiman, B. 2001. PHP4 Developer's Guide. The McGraw-Hill Companies, Inc. USA. 775 p.
- Sperling, L., J. A. Ashby, M. E. Smith, E. Weltzien and S. McGuire. 2001. A framework for analyzing participatory plant breeding approaches and results. Euphytica 122: 439-450.
- Steel, R. G. D. and J. H. Torrie. 1980. Principles and Procedures of Statistics: A Biometrical Approach. McGraw-Hill. New York. 633 p.
- Tkachenko, V., D. Haiduchonak, Mmp, D. Sanders, T. Harris. 2008. PHP-ExcelReader. <http://sourceforge.net/projects/phpexcelreader/>.
- Trujillo-Ortiz, A. and R. Hernandez-Walls. 2003. qtukey: Tukey's q studentized range critical value. A MATLAB file. <http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=3469&objectType=FILE>.