Introducing The Subjectogram: A Fast Way To Visualize Pairwise Comparisons

Carson Trego

November 17th, 2024

1 Introduction

The p-values obtained by initial results such as ANOVA, do not reveal how the individual levels perform compared to each other [7, 4]. To reveal further information about the levels a set of post-hoc pairwise comparisons are often conducted, in which each level within a factor is compared to each other level [7, 4]. The number of pairwise comparisons needed to compare all levels can be large, increasing the difficulty of interpreting the results [6, 8, 1, 4]. For example, a set of 18 levels results in 153 individual comparisons, each with an estimate and p-value [6].

Several graphical approaches have been created in an attempt to make pairwise comparisons more easily interpretable than inspecting the result of each pairwise comparison [6, 8, 1, 4]. A non-exhaustive list of methods for displaying pairwise comparisons will be listed in the next section, with the typical approaches involving directly showing the result of each comparison, using a Diffogram, using a Compact Letter Display, or using a Tukey Grouping Display [6, 8, 1, 4, 9].

Direct results, Diffograms, and Compact Letter Displays all require the user to search through the set of comparisons to see how each comparison relates to the complete set of levels [6, 8, 1, 4, 9]. For example, if an analyst is looking at a Compact Letter Display, they may see that the level they are observing belongs to grouping "A", this alone does not indicate if the level is in its unique grouping, or if there are other levels that are not significantly different from the level of interest [6, 8, 1]. In order to see if the level of interest is in its own unique grouping, the analyst has to inspect each level to determine if any other levels belong to grouping "A". If the analyst chose to use the direct comparisons or a Diffogram, the analyst would have to inspect each pairwise comparison with the level of interest to determine if the level of interest shares a grouping with any other levels in the study [6, 8, 1]. In using any of these displays, the analyst has to manually comb through several bits of information to understand which, if any, other levels share a grouping with the level of interest [6, 8, 1, 4, 9, 3].

Tukey Grouping Displays (seemingly) solve this issue by displaying groupings as being connected with a ribbon [6, 8, 1, 4, 9, 3]. To see if the level of interest shares a grouping with any other level in the study, all the analyst has to do is inspect the level of interest on the Tukey Grouping Display, and see if the ribbon connected to the level of interest extends out from the level of interest [6, 8, 1, 4, 9, 3]. This is a simple and fast way to interpret pairwise comparisons, but Tukey Groupings have a fundamental flaw that prevents them from being used in general: if the variance of the sampling distributions of the levels is not equal such that two neighboring levels can be significantly from each other, but neither is significantly different from another level, the ribbon display no longer works [6, 8, 1, 4, 9, 3]. This issue is often handled by software packages, such as SAS's PROC PLM, by adding a footnote that explains which indicated groupings are incorrect

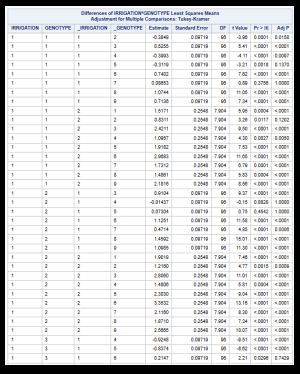
[9, 3]. This results in a display that has been criticized for being misleading, as the results of the graphic directly contradict the footnote information at the bottom of the chart [6, 8, 9, 3].

To address this issue, we have created a new form of display called the Valid Grouping Display. Valid Grouping Displays are almost identical to Tukey Grouping Displays but are modified to allow the ribbons to skip over levels. Rather than having a set of ribbons that indicate a grouping in all levels that they cross over, Valid Groupings have perpendicular lines that indicate which levels are included in the grouping, and which levels are being skipped over. The intended purpose of this new graphic is to allow users to quickly understand how the levels relate to each other, without having to resort to a misleading graphic that is complicated to interpret.

From our review of other studies, there has been no formal test of Valid Groupings or functionally similar graphics [6, 8, 1, 4, 9, 3]. While it may seem that the graphical display can reduce errors by displaying the pairs more intuitively and increase the speed at which a person can interpret a set of pairwise comparisons, these claims have not been formally tested. The following study is intended to investigate Valid Grouping Displays as a more effective alternative to traditional means of displaying pairwise comparisons such as Diffograms, Tukey Groupings, and Compact Letter Displays. The goal of the the new display is to prevent errors in analysis, as well as save time inspecting the results, and thus our experiment will use multiple accuracy and time metrics to compare Valid Grouping Displays to other means of displaying pairwise comparisons.

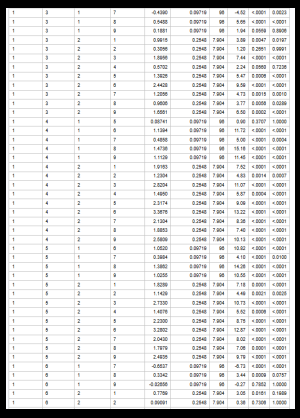
2 Methods of Displaying Pairwise Comparisons

2.1 Direct Output

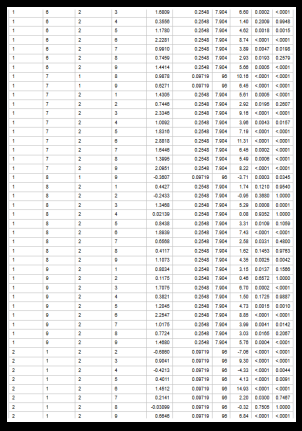


Differences of IRRIGATION*GENOTYPE Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

IRRIGATION	GENOTYPE	_IRRIGATION	_GENOTYPE	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
1	1	1	2	-0.3840	0.09719	95	-3.95	0.0001	0.0158
1	1	1	3	0.5255	0.09719	95	5.41	<.0001	<.0001
1	1	1	4	-0.3993	0.09719	95	-4.11	<.0001	0.0097
1	1	1	5	-0.3119	0.09719	95	-3.21	0.0018	0.1370
1	1	1	6	0.7402	0.09719	95	7.62	<.0001	<.0001
1	1	1	7	0.08853	0.09719	95	0.59	0.7795	1.0000
1	1	1	8	1.0744	0.09719	95	11.05	<.0001	<.0001
1	1	1	9	0.7135	0.09719	95	7.34	<.0001	<.0001
1	1	2	1	1.6171	0.2548	7.904	6.46	0.0004	<.0001
1	1	2	2	0.8311	0.2548	7.904	3.26	0.0117	0.1202
1	1	2	3	2.4211	0.2548	7.904	9.66	<.0001	<.0001
1	1	2	4	1.0887	0.2548	7.904	4.35	0.0027	0.0069
1	1	2	5	1.9182	0.2548	7.904	7.53	<.0001	<.0001
1	1	2	6	2.9863	0.2548	7.904	11.65	<.0001	<.0001
1	1	2	7	1.7312	0.2548	7.904	6.79	0.0001	<.0001
1	1	2	8	3.1881	0.2548	7.904	5.83	0.0004	<.0001
1	1	2	9	2.1816	0.2548	7.904	8.56	<.0001	<.0001
1	2	1	2	0.9104	0.09719	95	9.37	<.0001	<.0001
1	2	1	4	-0.01427	0.09719	95	-0.15	0.8828	1.0000
1	2	1	5	0.07264	0.09719	95	0.75	0.4542	1.0000
1	2	1	6	1.1251	0.09719	95	11.58	<.0001	<.0001
1	2	1	7	0.4714	0.09719	95	4.85	<.0001	0.0006
1	2	1	8	1.4692	0.09719	95	15.01	<.0001	<.0001
1	2	1	9	1.0695	0.09719	95	11.00	<.0001	<.0001
1	2	2	1	1.9919	0.2548	7.904	7.48	<.0001	<.0001
1	2	2	2	1.2190	0.2548	7.904	4.77	0.0015	0.0059
1	2	2	3	2.8090	0.2548	7.904	11.01	<.0001	<.0001
1	2	2	4	1.4806	0.2548	7.904	5.81	0.0004	<.0001
1	2	2	5	2.3030	0.2548	7.904	9.04	<.0001	<.0001
1	2	2	6	3.3632	0.2548	7.904	13.18	<.0001	<.0001
1	2	2	7	2.1150	0.2548	7.904	8.30	<.0001	<.0001
1	2	2	8	1.6710	0.2548	7.904	7.24	<.0001	<.0001
1	2	2	9	2.5885	0.2548	7.904	10.07	<.0001	<.0001
1	3	1	4	-0.9245	0.09719	95	-9.51	<.0001	<.0001
1	3	1	5	-0.8374	0.09719	95	-8.62	<.0001	<.0001
1	3	1	6	0.2147	0.09719	95	2.21	0.0298	0.7429



1	3	1	7	-0.4390	0.09719	95	-4.52	<.0001	0.0023
1	3	1	8	0.5409	0.09719	95	5.56	<.0001	<.0001
1	3	1	9	0.1851	0.09719	95	1.94	0.0559	0.8606
1	3	2	1	0.9815	0.2548	7.904	3.85	0.0047	0.0197
1	3	2	2	0.3086	0.2548	7.904	1.26	0.2851	0.9991
1	3	2	3	1.8986	0.2548	7.904	7.44	<.0001	<.0001
1	3	2	4	0.5702	0.2548	7.904	2.24	0.0563	0.7238
1	3	2	5	1.3826	0.2548	7.904	5.43	0.0006	<.0001
1	3	2	6	2.4426	0.2548	7.904	9.59	<.0001	<.0001
1	3	2	7	1.2896	0.2548	7.904	4.73	0.0015	0.0061
1	3	2	8	0.9508	0.2548	7.904	3.77	0.0058	0.0289
1	3	2	9	1.6561	0.2548	7.904	6.50	0.0002	<.0001
1	4	1	5	0.80741	0.09719	95	8.66	0.2707	1.0000
1	4	1	6	1.1394	0.09719	95	11.72	<.0001	<.0001
1	4	1	7	0.4868	0.09719	95	5.00	<.0001	0.0004
1	4	1	8	1.4738	0.09719	95	15.16	<.0001	<.0001
1	4	1	9	1.1139	0.09719	95	11.46	<.0001	<.0001
1	4	2	1	1.9183	0.2548	7.904	7.52	<.0001	<.0001
1	4	2	2	1.2354	0.2548	7.904	4.85	0.0014	0.0057
1	4	2	3	2.8224	0.2548	7.904	11.07	<.0001	<.0001
1	4	2	4	1.4870	0.2548	7.904	5.87	0.0004	<.0001
1	4	2	5	2.3174	0.2548	7.904	9.09	<.0001	<.0001
1	4	2	6	3.3876	0.2548	7.904	13.22	<.0001	<.0001
1	4	2	7	2.1324	0.2548	7.904	8.36	<.0001	<.0001
1	4	2	8	1.6883	0.2548	7.904	7.40	<.0001	<.0001
1	4	2	9	2.5839	0.2548	7.904	10.13	<.0001	<.0001
1	5	1	6	1.0530	0.09719	95	10.83	<.0001	<.0001
1	5	1	7	0.3904	0.09719	95	4.16	<.0001	0.0100
1	5	1	8	1.3852	0.09719	95	14.26	<.0001	<.0001
1	5	1	9	1.0255	0.09719	95	10.55	<.0001	<.0001
1	5	2	1	1.8299	0.2548	7.904	7.18	0.0001	<.0001
1	5	2	2	1.1439	0.2548	7.904	4.49	0.0021	0.0035
1	5	2	3	2.7330	0.2548	7.904	10.73	<.0001	<.0001
1	5	2	4	1.4078	0.2548	7.904	5.52	0.0008	<.0001
1	5	2	5	2.2300	0.2548	7.904	8.75	<.0001	<.0001
1	5	2	6	3.2902	0.2548	7.904	12.91	<.0001	<.0001
1	5	2	7	2.0430	0.2548	7.904	8.02	<.0001	<.0001
1	5	2	8	1.7979	0.2548	7.904	7.06	0.0001	<.0001
1	5	2	9	2.4835	0.2548	7.904	9.75	<.0001	<.0001
1	6	1	7	-0.6637	0.09719	95	-6.72	<.0001	<.0001
1	6	1	8	0.3342	0.09719	95	3.44	0.0059	0.0751
1	6	1	9	-0.02855	0.09719	95	-0.27	0.7852	1.0000
1	6	2	1	0.7769	0.2548	7.904	3.05	0.0161	0.1469
1	6	2	2	0.09991	0.2548	7.904	0.39	0.7008	1.0000



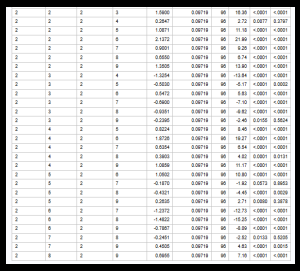


Figure 1: Example of the direct output produced by *SAS PROC GLIMMIX* for a set of pairwise comparisons given 18 levels based on a simulated drought-stress experiment [2]. From the 18 levels, 153 comparisons are produced, which may make interpretation difficult [6].

Software designed to process multiple comparisons often reports each comparison individually with a tabular format [2]. This non-graphical method of displaying multiple comparisons is automatically available from the analysis but can be difficult to interpret when a large number of comparisons are being performed.

2.2 Diffograms

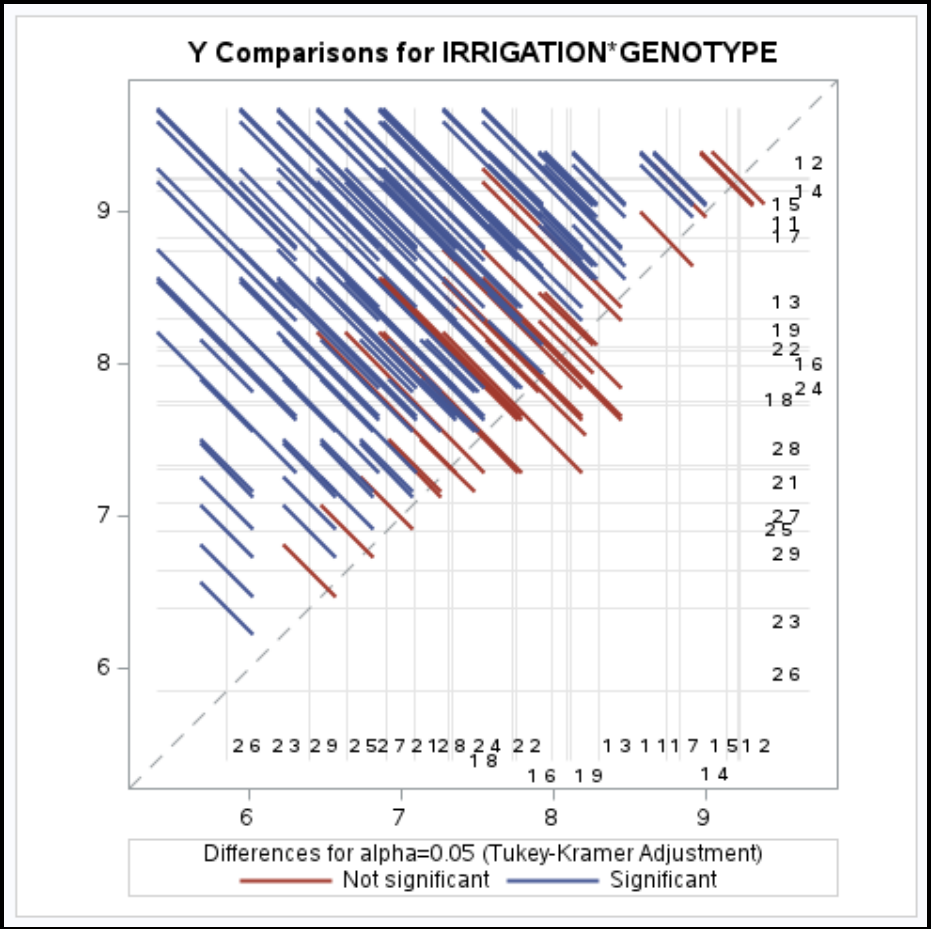


Figure 2: Pairiwse comparisons from the simulated drought-stress experiment displayed with a Diffogram given by *SAS PROC GLIMMIX* [2]. With 153 pairwise comparisons, many of the intervals are tightly packed together, potentially making reading this graphic difficult.

Diffograms display pairwise comparisons by plotting each level on the x-axis, and each level again on the y-axis [9, 2, 3]. From each level, a line perpendicular to the axis, and where each line intersects the pairwise difference confidence intervals are displayed diagonally. The zero line is also displayed diagonally [9, 2, 3]. If a difference confidence interval crosses the zero line, the color of the confidence interval is changed to emphasize that there is no significant difference between the two levels [9, 2, 3]. Similar to the direct output of the pairwise comparisons, a Diffogram requires that there is a direct result of each pairwise combination, making it difficult to display pairwise comparisons with several levels [9, 2, 3].

2.3 Compact Letter Displays

GENOTYPE	IRRIGATION	emmean	SE	df	lower.CL	upper.CL	.group
6	2	5.86	0.49	7	3.66	8.05	a
3	2	6.40	0.49	7	4.20	8.60	b
9	2	6.64	0.49	7	4.44	8.84	bc
5	2	6.91	0.49	7	4.71	9.10	cde
7	2	7.09	0.49	7	4.89	9.29	defg
1	2	7.31	0.49	7	5.11	9.51	fg
8	2	7.34	0.49	7	5.14	9.54	fg
4	2	7.73	0.49	7	5.53	9.93	hij
8	1	7.75	0.49	7	5.55	9.95	cd f h
2	2	7.99	0.49	7	5.79	10.19	hijkl
6	1	8.08	0.49	7	5.89	10.28	defghi
9	1	8.11	0.49	7	5.91	10.31	e g i
3	1	8.30	0.49	7	6.10	10.50	g i
7	1	8.74	0.49	7	6.54	10.94	j
1	1	8.82	0.49	7	6.63	11.02	jk
5	1	9.14	0.49	7	6.94	11.33	klm
2	1	9.21	0.49	7	7.01	11.41	lm
4	1	9.22	0.49	7	7.02	11.42	m

Figure 3: Pairiwse comparisons from the simulated drought-stress experiment displayed with a Compact Letter Display where each level is given one of more letters to signify a grouping given by *R multcomp* [5]. Levels with the same letter label are not significantly different from each other [5, 1, 6].

A popular method of displaying pairwise comparisons is the Compact Letter Display (CLD) [6, 1]. Compact letter displays work by assigning one or more letters to each level [6, 1]. In these displays, levels that are labeled with the same letter indicate that the levels are not significantly different from each other [6, 1]. In order to effectively use a Compact Letter Display, a reader must select a single level, and find which levels share the same letter label as the primary level of interest [6, 1]. Compared to inspecting the results of each comparison, which is equal to the number of combinations produced by the levels, Compact Letter Displays only require each level to show up once [6, 1]. While Compact Letter Displays can potentially decrease the cognitive load on a viewer interpreting the results, the viewer still needs to inspect every level individually to find all levels that are not significantly different from the primary level of interest [6, 1].

2.4 Tukey Grouping Displays

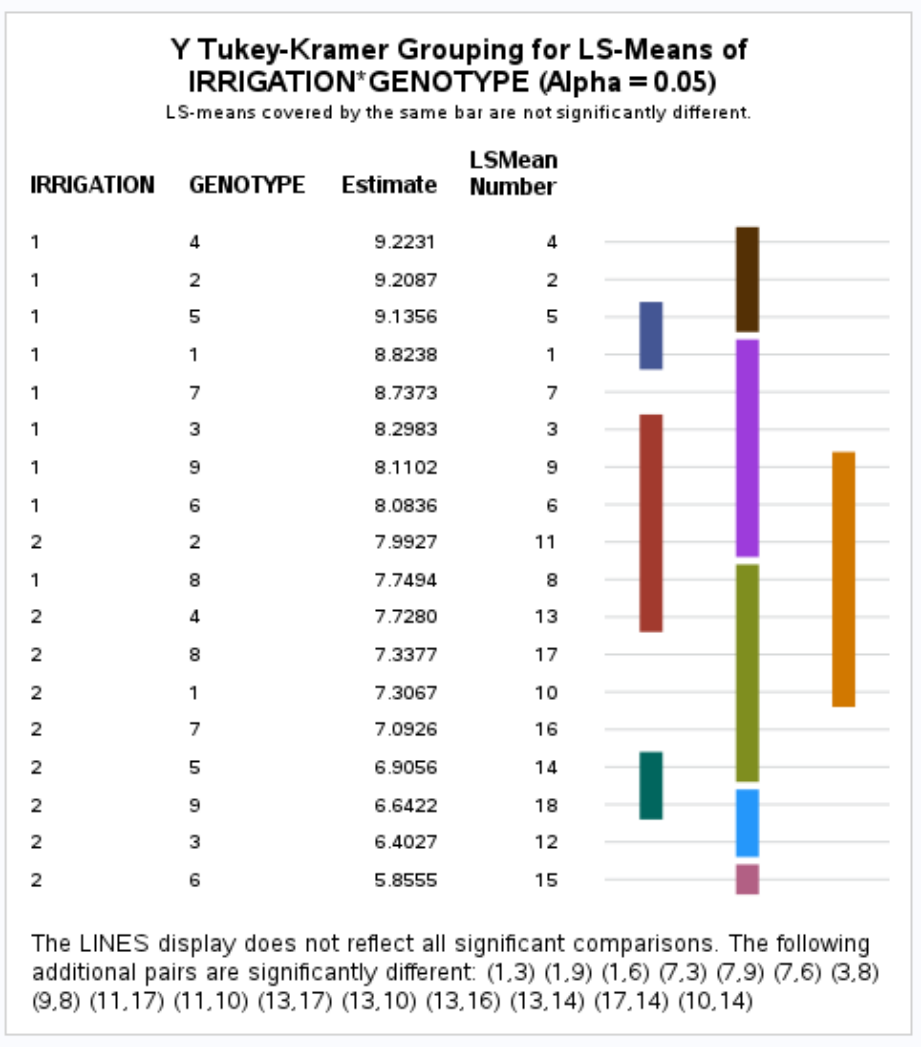


Figure 4: Pairiwse comparisons from the simulated drought-stress experiment displayed with a Tukey Grouping Display given by *SAS PROC PLM LINES* [3]. As with the Compact Letter Display, all levels under the same grouping are not significantly different from each other but are displayed using a set of lines instead of letters [6, 1, 3]. This plot assumes there are no gaps across the groupings, which does not hold when the variance of the sampling distributions for each level is not equal [3, 8, 6]. The display attempts to correct misleading results with a footnote at the bottom [3].

Tukey Groupings are similar to compact letter displays in that they display groups of levels that are not significantly different from each other but display this information graphically using a set of lines [1, 6, 8, 3]. This form of display makes seeing the groups of levels that are not significantly different immediately clear, as all levels that are not significantly different are connected by a ribbon,

allowing the reader to trace the ribbon and see all levels that belong to the same group [1, 6, 8, 3]. While this even further decreases the cognitive load on the viewer, the line display alone is only valid when a set of assumptions are met [1, 6, 8, 3]. The line plot assumes that if two levels are not significantly different, then any levels with estimates between the values of the two levels must also be not significantly different [1, 6, 8, 3].

This assumption is easily broken when the variance within the levels is not equal [8, 6, 3]. To address this problem, these displays will often be programmed with an additional footnote, declaring that the line plot does not display all significantly different levels [3]. This is the default method used by many plotting software packages, such as SAS's proc PLM *LINES* display [3]. While including the footnote indicating which significant comparisons are not shown allows for all significant differences to be displayed in some form, the graphical display itself can be misleading, and the small footnote below may make the interpretation of the display as a whole difficult [8, 6, 3]. Even worse, the footnote may be missed altogether. Due to the inherent misleading results produced by these line displays, many authors have argued that these displays should not be used [8, 6].

2.5 Valid Grouping Displays

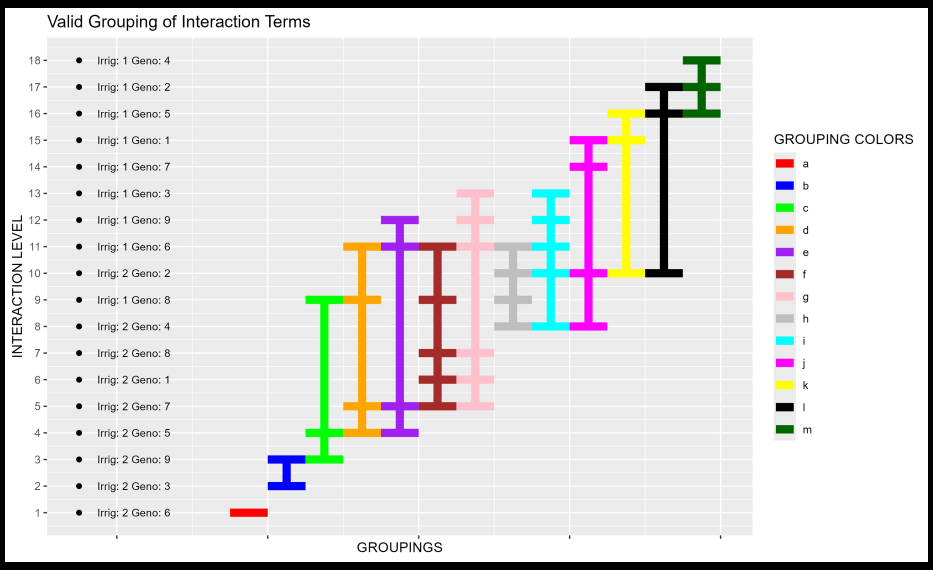


Figure 5: Pairwise comparisons from the simulated drought-stress experiment displayed with a Valid Grouping Display. Groupings in which all included levels are not significantly different from each other are indicated by color-coded perpendicular lines. In the above plot, Irrig 2 Geno 9 is not significantly different from Irrig 1 Geno 8, but Irrig 2 Geno 9 is significantly different from Irrig 2 Geno 7. All of the groupings indicated are valid even when the variance of the sampling distribution is not equal, so there is no need for a footnote as in the case of Tukey Grouping Displays.

Valid Groupings are a new method of displaying multiple comparisons, which combine the visual display of Tukey Groupings with the validity of a Compact Letter Display. This is made possible by incorporating a set of perpendicular lines in the Tukey Grouping lines, indicating a level that is not significantly different from every other level marked by the group. In contrast to the original Tukey

Grouping, Valid Groupings are able to skip over levels, eliminating the need for a footnote detailing which statistically significant pairs were missed [8, 6, 3].

3 General Design and Purpose

While the Valid Groupings were designed to make seeing multiple comparisons less difficult, it is unclear if a graphical display is any more effective than more conventional such as the direct pairwise comparison output, Diffograms, or compact Letter Displays. The Valid Groupings display was designed to modify the Tukey Grouping plot less misleading with a more direct interpretation, Valid Groupings have not been formally tested. To provide an evidence-based recommendation for a method of displaying pairwise comparisons, we will conduct a study to measure participants' ability to interpret pairwise results based on several metrics, including the false positive rate, false negative rate, and average completion time.

A series of questions will be designed which require a pairwise comparison to find the answer. Each question will be based on simulated data, in which there will be one-factor categorical factor with multiple levels and one quantitative response variable. Questions will be subset based on the number of levels and if the variance is assumed to be equal within levels or if the variance is not assumed to be equal within levels.

Participants will be randomly assigned to one of six groups. The first group will be given problems with displays that directly show the multiple comparisons output, which includes the two levels being compared, an estimated value for the difference, and whether or not they are significantly different. For simplicity of interpretation, the statistical significance will be displayed as a binary outcome: either statistically significant or not, and the exact p-value or statistical significance "star rating" will not be displayed. The second group will be given the same questions, but the comparisons will be displayed with a Diffogram. The third group will be given the questions and will have to answer using a Compact Letter Display. The fourth group will be given the questions and will be expected to answer using a Tukey Grouping. Due to the inherent misleading aspect of the Tukey Grouping display, each Tukey Grouping will be paired with a footnote declaring what statistically significant pairings were not shown in the display. The fifth and final group will be assigned the questions using our Valid Groupings design. Since the Valid Groupings do not produce the misleading results that the Tukey Groupings do, there will not be any footnotes included on the Valid Groupings display.

Each display of data will be paired with a multiple-choice question with one intended "correct" answer. Participants will be asked to use the display to determine the relationship between two levels, and have the choice between saying if the two levels are significantly different or not, and if so, in which direction. Each question will also be paired with an "I am unsure" option, allowing participants to state when they believe they are unable to solve the question. An example of a question given to participants will be as follows:

```
Given the above display, determine the relationship between LEVEL C and LEVEL F

A - LEVEL C and LEVEL F are SIGNIFICANTLY DIFFERENT:  LEVEL C is GREATER
B - LEVEL C and LEVEL F are SIGNIFICANTLY DIFFERENT:  LEVEL C is LESSER
C - LEVEL C and LEVEL F are NOT SIGNIFICANTLY DIFFERENT: We cannot conlcude whether
    or not LEVEL C or LEVEL D is greater.
D - I am unsure how to answer this question.
```

3.1 Performance Measures

While participants are answering the questions, three performance measures will be recorded. The first measure of performance is the *False Positive Rate*, which indicates the proportion of times a participant indicated that two levels were significantly different when the participant was given two levels that are NOT significantly different. The second measure of performance is the *False Negative Rate*, which is the proportion of times a participant indicated that two levels were NOT significantly different when the participants were given two levels that are significantly different. The third measure of performance, *Null Time*, will be the time it takes a participant to complete a question when a statistically significant difference is NOT present, and the fourth measure of performance, *Alt Time*, will be the time it takes to complete a question when a statistically significant difference is present.

3.2 Null Hypothesis

For any graphical display used, we set out the null hypothesis to be that all graphical methods perform the same for the False Positive Rate, False Negative Rate, Null Time, and Alt Time. Conversely, we consider the null hypothesis to be rejected if there is a statistically significant difference in any of the graphical displays within a performance metric. A statistically significant difference between two graphical displays for a performance metric implies that a difference of performance between the two graphical displays is likely, while a lack of a statistically significant difference between two graphical displays neither implies that a difference in performance or lack thereof is established. Two graphical displays may fail to achieve a statistically significant difference in performance for one of the performance metrics, but this does not imply that the two displays have equivalent performance.

3.3 Equivalence

With the statistical analysis methods being used, only two outcomes can be achieved when comparing a performance metric across two types of graphical displays. For a given performance metric, two types of displays can either be significantly different or not. The lack of a significant difference will result in the null hypothesis not being rejected, but this does not imply that two graphical displays perform equivalently. In order to establish the equivalence of performance, we must set standards to define what meets the criteria for equivalence. Prior to the study being performed, a margin of equivalence will be set for each performance metric.

For this study, we will define the margin of equivalence for both time metrics to be within two seconds. This means that if two graphical displays are compared, and the difference confidence interval lies entirely between negative two seconds and positive two seconds, the two displays will be considered equivalent for that time metric. If either bounds of the difference confidence interval lie beyond negative two or positive two, the test for equivalence will not be considered proven. Margins of equivalence will also be set for the False Positive Rate and the False Negative Rate, in which the margin of equivalence will be within negative two percent and positive two percent for both the False Positive Rate and False Negative Rate. These margins are based on what is considered a practical and meaningful difference in performance but are ultimately subjective. To address any concerns about how equivalence is defined, the definition of equivalence and the margins used will be included clearly in the analysis section of the paper.

4 Participant Instruction and Question Filtering

The topic of this study is fairly complex, and the design of the study should aim to be as accessible as possible to new participants. Participants may have misconceptions about statistical significance, or be completely unfamiliar with the topic. To address this issue, the study will begin with a short instructional video and question set that aims to quickly introduce the participants to pairwise statistical significance. Participants will not be required to do calculations, and the interpretation of statistical significance will be restricted to whether or not we know a level is different from another. With these restrictions on the amount of interpretation the participants need to engage in, we believe that participants should be able to understand how to answer the questions correctly after a short training exercise, regardless of their background.

To train participants, they will be given a short video that is tailored to the experimental group they are placed in. The video will first treat "significantly different" as an operational definition for the purpose of the study, and participants will be told that the direct numeric results of a comparison are sometimes deceptive, and so we consider a result *significantly different* when we can confirm that the difference actually exists, and is not the result of noise. This is not the standard definition of statistical significance, but this definition was chosen to make the interpretation of the graphics easier for people who are inexperienced with statistical analysis. To acknowledge that there is further nuance in interpreting statistical significance, we will strongly indicate that the definition we are using is for the purpose of the study, and not a definition used in the field. To indicate that this is an operational definition for the experiment, we will say at the beginning that we are defining the word significance to mean there is a known difference, and acknowledge that this is not the standard definition directly. To show how significance would be conveyed to the participants, we will use the following script, which will be displayed as text and as audio in a video presentation:

```
For the purpose of this study, we will define ``significantly different'' as a phrase to describe when we are able to establish that two levels are different from eachother.

Sometimes our tools can indicate that two levels are different, but they are not different in reality. For example, two people may weigh the exact same, but the scale may say that PERSON A is 150.000 pounds and PERSON B 150.001 pounds. The scale is not very good at measuring weight consistently and is only accurate within a pound of what it says on the scale, so while it may seem like PERSON B weighs more than PERSON A, our scale is not accurate enough to make that claim, so we will instead say that PERSON B is not significantly different than PERSON A.

Now let's say that we use the same scale, but PERSON C weighs 300 pounds. Since our scale is accurate within a pound of the weight it displays, and PERSON C appears to be 150 pounds heavier than PERSON A, we can say that PERSON C and PERSON A are significantly different, with PERSON C having a greater weight than PERSON A.

Overall, we can take the weights of PERSON A,B, and C, and compare them against each other to find that

-PERSON A is not significantly different than PERSON B.
-PERSON B is significantly different than PERSON C, in which PERSON B is LESSER.
-PERSON C is significantly different than PERSON A, in which PERSON C is GREATER.

Let's look at how this would be displayed graphically:
```

After being introduced to the content, the participants will take part in a training set of questions, in which they will be told after answering a question if they were correct, and how the correct answer could be obtained using the graphic. Each correct answer adds one point to a progress bar, and

incorrect questions remove one point from the progress bar. The progress bar cannot go below zero. After the participant can obtain 5 points, the study will continue. The amount of tries required for each participant to obtain 5 points, as well as if they gave up prior to obtaining 5 points, will be recorded.

When using anonymous surveys, active subject participation is a major concern, and the study must be designed to take into account participants recording answers randomly simply to get the reward of the study. To mitigate this issue, participants will be given a set of simple questions throughout the study. Some of these questions will be evaluating comparisons that are simple and only have two levels, while other questions will be unrelated to the study, but will have answers that will be obvious if the participant is reading the questions and putting effort into answers. For example, the question may be "Select answer D" and if answer D is not selected, this will be noted on the participant's result. Failing to answer the basic questions correctly will not void the participant from getting the reward, but these answers will be removed from the dataset during analysis.

With the rise of large language models, it is now possible for participants to set up AI agents to fill out several surveys for them, maximizing potential rewards. There are two options that can be used to limit this risk, the first one will include simple questions designed to be difficult for large language models to solve. These questions will be based around spatial awareness, and participants will have to do a spatial task such as aligning an arrow to the direction of a 3D model plane. The other alternative to limit the use of large language models would be to have the study performed in person using computers that are set up in advance and are not connected to the internet. Assuming that having the examination in person will not exceed the budget of the experiment, the latter option will be used, as having the experiment take place in person will make it more difficult for the reward to be "gamed" by participants.

Poor performance on the general questions will not be used as a justification to remove data, and outliers that perform much better or worse than average will not be removed unless there is external evidence of an issue. An example of external evidence of an issue would be if the computer malfunctions and fails to display images correctly, and the participant answers the questions incorrectly as a result.

5 Privacy and Informed Consent

The involvement of human participants requires certain ethical standards to be upheld. While this form of experiment has a minimal risk to the participants' well-being, the design must be approved by the International Review Board (IRB) prior to the experiment taking place.

One specific ethical consideration is participant privacy. To maintain the privacy of the participants in the study, participants will be given an option to not submit their data as part of the research study. If the participants choose to not submit the data, they may still participate in the experiment and receive the planned reward, but this information will be deleted before any analysis is conducted.

Before the questions are administered to the participants, they will be given some information about the purpose of the experiment, and what sort of questions to expect. Participants will also be given information about how the data will be analyzed, and what it means to choose not to send data. As part of informing participants what the option to submit data entails, they will be told that choosing to not submit their data will not void the reward, nor will the proctor be aware of their decision.

After the end of the experiment has concluded, the proctor will ask the participants if they have any questions related to the experiment. Following addressing any questions the participant may

have, the proctor will give the participants the option to be added to an email list, which will notify the participant when the results have been published.

6 Data Simulation

For this experiment, there will be six different treatment groups, with the treatment groups being the Direct Output group, the Diffogram Group, the Compact Letter Display group, the Tukey Grouping Display group, and the Valid Grouping Display group. Each group will receive a set of questions with the same dataset generation methods and analysis assumptions. Half of the data will be simulated with levels having nearly the same variance, and the analysis techniques will assume equal variance, while the other half of the data will be simulated with unequal variance of levels, with the methods used to analyze those datasets not assuming equal variance of the levels. After the pairwise comparisons are completed, they will be displayed using one of the six methods being considered. Each participant will be assigned one method of displaying pairwise comparisons, and they will only be given displays using that method. This was chosen to avoid overloading participants with instructions on how to use six different methods at once.

The simulated datasets will have three different amount of levels being used. One group of datasets will have 3 levels and three 3 comparisons, the second group of datasets will have 9 levels and 36 levels of comparisons, and the third group will have 18 levels and 153 comparisons.

Each treatment group will receive a mix of all comparison levels and variance levels, with each participant answering 5 questions of each type, or 30 questions total in addition to their introductory questions and attention check questions.

7 Analysis

(Note, the model and analysis choices are dependent on how the data turns out, but in lieu of actual data, we will state the model assumptions and assume they are met.)

Each of the four performance metrics, False Positive Rate, False Negative Rate, Null Time, and Alt Time will be its own response variable with its own model. In each model, the number of levels given to the participant, the type of display shown to the participant, and whether or not variance was considered equal when displaying the pairwise comparison analysis will be fixed effects. We consider it possible that some display types may be more affected by increased comparisons and the involvement of unequal variances, so the interaction terms for the three effects will be included in our model.

8 Goals

The Valid Grouping Display may be helpful as a means to visualize pairwise comparisons, but as have seen with many other types of graphics, it may be misleading, confusing, or difficult to use. Before recommending the use of valid grouping displays in place of other displays, it is important to gather some evidence that Valid Grouping Displays work as intended. The results of this study may not fully inform those decisions but could introduce others to the Valid Grouping Display, as well as prompt further research.

References

[1] John M. Ennis, Charles M. Fayle, and Daniel M. Ennis. Compact letter displays to concisely represent pairwise sensory information. In *Compact Letter Displays to concisely represent pairwise sensory information*, 2011. URL: https://api.semanticscholar.org/CorpusID:63443861.

[2] SAS Institute. Sas/stat® 12.3 users guide the glimmix procedure. Online, 2013. URL: https://support.sas.com/documentation/onlinedoc/stat/123/glimmix.pdf.

[3] SAS Institute. Sas/stat® 14.1 users guide the plm procedure. Online, 2015. URL: https://support.sas.com/documentation/onlinedoc/stat/141/plm.pdf.

[4] Penn State's Department of Statistics. *Analysis of Variance and Design of Experiments*. The Pennsylvania State University, 2023.

[5] orsten Hothorn ORCID iD. multcomp: Simultaneous inference in general parametric models. Online, 2010. URL: https://cran.r-project.org/web/packages/multcomp/index.html.

[6] Hans-Peter Piepho. An algorithm for a letter-based representation of all-pairwise comparisons. *Journal of Computational and Graphical Statistics*, 13(2):456–466, 2004. doi:10.1198/1061860043515.

[7] Cahyono St. Design and analysis of experiments, 06 2022.

[8] Peter H. Westfall. *Multiple Comparisons and Multiple Tests using the SAS System*. Sas Institute Inc, 1999. URL: https://support.sas.com/resources/papers/proceedings/proceedings/sugi24/Stats/p264-24.pdf.

[9] Rick Wicklin. The diffogram and other graphs for multiple comparisons of means. Online, 2017. URL: https://blogs.sas.com/content/iml/2017/10/18/diffogram-multiple-comparisons-sas.html.