Annals of Agricultural Science, Moshtohor (ASSJM)
https://assjm.journals.ekb.eg/
www.fagr.bu.edu.eg
© 2023, Faculty of Agriculture, Benha University, Egypt. ISSN:1110-0419
Original Article Vol. 62(4) (2024), 51 – 66 DOI: 10.21608/assjm.2025.324602.1316

Model Assumptions Testing of Annual Ryegrass-Egyptian Clover Cultivars Component in Mixtures and Repeated-Measure Harvests

Magda M. Salem Badr, Hany A. Tageldin, Haroun M. Mosa El Naggar, and Mohamed F. Tolba
Agronomy Department, Faculty of Agriculture, Benha University, 13736, Qalyubia, Egypt
Corresponding author: Magda M. Salem Badr Email: m.badr48700@fagr.bu.edu.eg

Abstract

It is a myth that a dataset *a priori* may not violate the assumptions of univariate general linear model (GLM). Validation of hypothesis testing (HT) is threatened if assumptions are violated. This research aims to check normality and variance homogeneity for between- and within-factor levels. In a 2-year field trial, *Lolium multiflorum* cv 'Local' was seeded in 6 percentage mixtures with 3 *Trifolium alexandrinum* cvs components in mixtures. A factorial model was fitted to forage yield, with 2 between-factors, and one within. Quantile-Quantile (Q-Q) plots and Shapiro-Wilk test were used to check normality. For between-factors, variance homogeneity is tested using plots of residuals and HT's Levene's test. For repeated measures, Mauchly's test is used to estimate sphericity. Six extreme outliers were spotted overall in the 2 years. The Q-Q plots showed that most residuals lay on the fitted lines, implying that normality was not violated. Residual error variance homogeneity was violated in Year 1. Inspecting repeated-harvests variances, Mauchly's W declared a minor violation of sphericity. Violations of model assumptions exist in real-life data.

Keywords: Assumptions, Homogeneity, Normality, Model, Repeated Measures, Residuals, Sphericity, Variance, Violation.

Introduction

Between-Factors ANOVA

The univariate GLM assumptions indicate that the response variable, Y should be predicted from a fitted predictor variable, $\hat{Y}$, comes from a covariance matrix of equal variance and covariance, and is $\sim N$. Put differently, residuals are normal, have a common variance (σ^2), and are independent. Diagnostic tests are carried out via HT and residuals plots.

Reporting whether a test is performed on the response or the residuals is ignored. Kozak & Piepho (2018) confirmed that the residuals are relatively more informative. Moreover, relying on only HT might be misleading. Draper & Smith (1981, p. 141) considered (e_i), from its definition, as the amount of variation a fitted equation failed to explain. If the model is correct, e_i should not decline assumptions. Steyn (2021) explained that interest is in the y's, given that x has already been entered. The interest is in the unexplained variation (residuals). Nonnormal residuals occur due to nonnormal y, but normal y does not guarantee normal residuals.

A dataset is 'messy' if it has outliers, belongs to skewed/kurtotic distributions, and suffers from variance heterogeneity (Shahin, 2017). Diagnostic tests are suggested if violations exist (Lix et al., 1996). Delacre et al. (2020) argued that assumptions are rarely fulfilled. Since variance heterogeneity caused biasedness of F test, they recommended Brown-Forsythe (Brown & Forsythe, 1974) and Welch's ANOVA (Kohr & Games, 1974). Type-1 error is affected by violations even with nonparametric tests (Marcinko, 2014). Transforming data or handling outliers may be practiced (Schützenmeister et al., 2012; Schützenmeister & Piepho, 2012; Debashis, 2013).

Whether a test is robust to one violations has been debated. Blanca et al. (2017) considered that real data is not often N-distributed and/or homoscedastic. They studied how robustness to nonnormality be related to type-1 error. (Blanca et al., 2013) addressed skewness and kurtosis effects on type-1 error and power. Variance heteroscedasticity was addressed (Blanca et al., 2018), and it was also tested related to type-1 error (Rogan & Keselman, 1977).

Under homoscedasticity if factor levels are small, and n goes to ∞, F test is robust to nonnormality (Arnold, 1980; Akritas & Papadatos, 2004). When factor levels go to ∞, ANOVA becomes more complicated (Wang & Akritas, 2006). In a one-factor (Boos & Brownie, 1995) and in mixed models (Akritas & Arnold, 2000), at heteroscedasticity for factor levels =30 with n= 4-15 at $\alpha = 0.05$, type-1 error inflated. At homoscedasticity, it was close to α. Akritas &

Papadatos (2004) addressed a heteroscedasticity when factor levels go to ∞ at small/large n. They might not be homoscedastic if n is small. In balanced /unbalanced cases, F test was sensitive to heteroscedasticity. In Driscoll's (1996), nonnormality affected type-1 error relative to α. Also, with t=2-21 and n=2-20, this resulted in very trivial differences between type-1 error and α under a range of distributions. Approaching normality, the relative difference was only 0.01. As well, increasing both t and n kept the magnitude narrow. Driscoll (1996) emphasized that it is not a faulty decision to apply ANOVA based on a nominal 0.05 where the true one is 0.03 difference.

Does ANOVA's 'claimed' robustness to heteroscedasticity extend to many main and interaction effects? Bathke (2004) assessed the results of (Boos & Brownie, 1995; Akritas & Arnold, 2000) since they have only assumed homoscedasticity. Therefore, he tested F test for main effect of one factor, in a factorial ANOVA, of maybe, many factors, under heteroscedasticity for all factors except for X_1, he simulated changes in type-1 error, at $\alpha = 0.05$, for 2 factors when levels of X_1 were 4 and 20, X_2=2, and σ^2 increased by 4- , 9-, 25- and 100-fold, among the variances of factor X_2. At homoscedastic σ^2, type-1 error was 0.048 regardless of X_1 levels. At $\sigma^2 = 100$, it was 0.073 for X_1=4 and 0.055 for X_1=20.

Inspecting outliers is crucial for their may influence inferences. Outliers might distort parameter estimates (Wainer, 1976; Lind & Zumbo, 1993), "*Fringeliers*" lie $\geq 3\,SD$ (Wainer, 1976). Osborne & Overbay (2004) considered a fringelier is within the domain of outliers. Draper & Smith (1981, pp.152-153) defined an outlier if its $|e_i|$ lies farther $\bar{e}_i$ by $\geq 3 - 4\,s$, to agree with Jones's (2019) where it lies $> \lambda\sigma$ from $\bar{x}$, $|x_i| > (\bar{x} + \lambda\sigma)$, λ= 2-3.

Three valid questions arise: i) What reasons do outliers? ii) How does one handle them? and iii) What are the consequences of their keeping/deleting? Neither Draper & Smith (1981) nor (Halldestam, 2016) supported deletion, perhaps outliers might have resulted from significant lurking variables (Draper & Smith, 1981; Orr et al., 1991), or outliers are yet as credible as any observation (Halldestam, 2016). Contrarily, Osborne & Overbay (2004) reported some researchers who adopted their outright removal.

Outliers may affect type 1 & 2 errors (Osborne & Overbay, 2004; Halldestam, 2016; Jones, 2019). To examine how a parameter estimate is 'robust' [1] and outliers affect inferences Halldestam (2016) simulated 3 treatments using n= 33, 100, and 1000 with one outlier in all 3, type-1 error was < α, indicating that 1-way ANOVA is not very sensitive to a single outlier regardless of n. A change in type-1 error implies a change in type 2. This is to conclude that type-2 error needs investigation.

Outliers influence extraction in *Exploratory Factor Analysis* (EFA) [2]. Liu et al. (2012) explored outliers' magnitude and number relation to extracted factors by conducting a 3-way ANOVA (3 outlier levels, 4 magnitudes, and 4 variables w/ outliers) on factor number. They calculated (η^2)[3] to find that sources with highest η^2 contained an outlier component. Generally, extreme points are either 'influential' or 'outlier'. Orr et al. (1991) emphasized that outliers do not threat test validity.

Outliers must be inspected (Wainer, 1976; Vandierendonck & De Soete, 1983; Lind & Zumbo, 1993), or be waived by nonparametric tests (Snell & Sprent, 1995; Kvam & Vidakovic, 2007). Arnold (1980) phrased, "… *asymptotically, no observation has a nontrivial effect on the estimation of its mean*."

Within-Factors ANOVA

Since harvests are on the same unit, these harvests might suffer from pairwise non-zero covariances and heterogeneous variances in Σ matrix. To test ≥2 Σ equality, Box's M[4] is employed (Abdi, 2010; Zaiontz, 2023). Sphericity should not be violated (Lane, 2016) by testing variance of differences (Kim, 2015). If it did, this would inflate type-1 error (Lane, 2016; Haverkamp & Beauducel, 2017). Haverkamp & Beauducel (2017) tested sphericity, n=20-100, and within-subjects number (t=3, 6, 9) on ANOVA-no correction, ANOVA Greenhouse-Geisser (GG) correction $\hat{\varepsilon}$ [5], and ANOVA- Huynh-Feldt (HF) correction $\tilde{\varepsilon}$ [6]. With sphericity, for t=3, type-1 error was close to α, and 'n' did not affect the 3 ANOVAs. If violated, for ANOVA-no correction, type-1 errors started at 0.07, n=20, by reaching 100, a negligible increase was detected, with not much shift between n=40, 60, and 80. For ANOVA (GG and HF), they were close to α. Depending on its severity, Lane (2016) held that F test's df got higher to reduce p and inflated type-1 error. Multiplying by GG's $\hat{\varepsilon}$, or HF's $\tilde{\varepsilon}$, df are lowered. Epsilon reflects how sphericity is violated (Box, 1954 as cited in Lane, 2016), it ranges $((t\text{-}1)^{-1}$ - 1.0) (Lane , 2016). Quintana & Maxwell (1994) suggested HF's $\tilde{\varepsilon}$ if $\tilde{\varepsilon} > 0.75$, yet GG's $\hat{\varepsilon}$ otherwise.

Both Pillai's Trace and Wilki's Lambda (Λ) are associated with MANOVA (Ateş et al., 2019). For both, if $p < \alpha$, H_0 is rejected; however, they are

[1] Robustness results in satisfying estimators if dataset is unaffected by outliers and/or little violations assumptions.

[2] EFA, is a multivariate procedure to identify factors that explain the order & structure among variables (Liu et al., 2012; Watkins, 2018).

[3] η^2 is SS of a factor divided by the total SS.

[4] For *m* independent populations the Box's M is to test H_0: $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_m$

[5] The Greenhouse-Geisser procedure estimates epsilon ($\hat{\varepsilon}$) to correct F test's df.

[6] The Huynh-Feldt correction estimates epsilon ($\tilde{\varepsilon}$) to correct F test's df.

interpreted differently. Pillai's Trace ranges 0.0-1.0; the closer to 1.0, the higher the contribution of a predictor; for Wilki's Λ, the closer to 0.0, the higher the contribution.

This research aims at diagnosing violation of normality, variance homogeneity and sphericity for a between- and within-factor model.

Materials and Methods

Field Experiment Layout, management, and Data Collection

Annual ryegrass (*Lolium multiflorum*) cv. Local, was seeded in mixtures with 3 Egyptian clover (*Trifolium alexandrinum*) cultivars: Helaly, Giza 6, and Gemiza. The Seed mixture was broadcast at seeding rates:12.0 kg and 20.0 kg fa^{-1} for ryegrass and clover. On a seed-weight basis, percentage ryegrass: clover mixtures were 20:80, 80:20, 30:70, 50:50, 60:40, and 40:60.

Trials were laid out factorially in 3 clover components x 6 mixtures in 4 RCBD in 2021 and 5 in 2022. Treatments were in three tiers per block. Plot area was ~ 6.75 m^2 (1.85 x 3.65). Planting was on 9 November 2021 and on 30 October 2022 at Benha University Farm (30^0 21`10.2 N Lat., $31^0$ 13`36.43 E Long.) on [silty clay (fine clayed, mixed typic fluvude)] soil. The Previous crop was carrot (*Daucus carota*) in 2021 and *Zea mays* L. in 2022. Trials were cut 3 times in 2021 and 4 in 2022. Following each cut, 25 kg calcium superphosphate (15.5% P_2O_5) was added. A random sample of fresh yield was collected to estimate dry matter.

Two-Between Factors and One-Within Factor Model

A model was fitted with two between-factors, and one within-factor. The two factors are 3 clover components in 6 mixture ratios; the one within-factor is the number of harvests. The LAM is,

$$Y_{ijkl} = \mu + \rho_l + \alpha_i + \beta_j + (\alpha\beta)_{ij} + (\rho\alpha\beta)_{ijl} + \psi_k + (\rho\psi)_{lk} + (\alpha\psi)_{ik} + (\beta\psi)_{jk} + (\alpha\beta\psi)_{ijk} + \varepsilon_{ijkl}$$

where, Y_{ijkl} is the l^{th} total dry forage yield (kg unit $area^{-1}$) for block, ρ_l, and of i, j, k levels of forage mixture yield, α_i, clover in a mixture, β_j, and harvest forage yield, ψ_k. And $(\rho\alpha\beta)_{ijl}$ is error (a), and ε_{ijkl} is the residual error ~N and iid (mean zero and σ_ε^2).

Model Assumptions Diagnostics

The model assumes that the observations of Y_{ijkl} are assumed independent and normally distributed (IND) with error variance (σ_ε^2). The expected mean response is $E(y_{ijkl}) = \mu + \rho_l + \alpha_i + \beta_j + (\alpha\beta)_{ij} + (\rho\alpha\beta)_{ijl} + \psi_k + (\rho\psi)_{lk} + (\alpha\psi)_{ik} + (\beta\psi)_{jk} + (\alpha\beta\psi)_{ijk}$ and a residual error. However, repeated forage harvests might suggest that residuals may not have a common variance. Hence, these urges testing sphericity assumption.

Table 1. ANOVA for two-between factors and one-within factor repeated measure model. (Block is random, A, B, and C are fixed).

Source	DF	EMS
Block, R	R-1	$\sigma_e^2 + c\sigma_\delta^2 + abc\sigma_\rho^2$
Mixture, A	A-1	$\sigma_e^2 + c\sigma_\delta^2 + rbc\sum_i (\alpha)_i^2/(a-1)$
Cultivar, B	B-1	$\sigma_e^2 + c\sigma_\delta^2 + rac\sum_j (\beta)_j^2/(b-1)$
A x B	(A-1) (B-1)	$\sigma_e^2 + c\sigma_\delta^2 + rc\sum_{i,j} (\alpha\beta)_{ij}^2/(a-1)(b-1)$
Error (a), RAB	(R-1) (AB-1)	$\sigma_e^2 + c\sigma_\delta^2$
Harvest, C	C-1	$\sigma_e^2 + rab\sum_k \psi_k^2/(k-1)$
RC	(R-1) (C-1)	$\sigma_e^2 + ab\sigma_\theta^2$
A x C	(A-1) (C-1)	$\sigma_e^2 + rb\sum_{i,k} (\alpha\psi)_{ik}{}^2/(a-1)(c-1)$
B x C	(B-1) (C-1)	$\sigma_e^2 + ra\sum_{j,k} (\beta\psi)_{jk}{}^2/(b-1)(c-1)$
A x B x C	(A-1) (B-1) (C-1)	$\sigma_e^2 + r\sum_{i,j,k} (\alpha\beta\psi)_{ijk}{}^2/(a-1)(b-1(c-1)$
Residual error, RABC	(R-1) (C-1) (AB-1)	σ_e^2
Total, RABC	RABC-1	

Residuals Checking

For the between-factor, $\hat{e}_{ijl} = y_{ijl} - \hat{y}_{ijl}$ to estimate $\varepsilon_{ijl} = Y_{ijl} - E(Y_{ijl})$, and the standardized residuals, $st.\hat{e}_{ijl}$, are estimated. Hence, both hypotheses testing model assumptions, as well as the visual residual plot analyses, are all only limited to the estimated MS error variance resulting from the univariate model. For the within-factor, $\hat{e}_{ijkl} = y_{ijkl} - \hat{y}_{ijkl}$ to estimate $\varepsilon_{ijkl} = Y_{ijkl} - E(Y_{ijkl})$, and the $st.\hat{e}_{ijkl}$ are estimated. Both mean errors, $\bar{\hat{e}}_{ijl}$ & $\bar{\hat{e}}_{ijKl} = 0$. Std. residuals are plotted vs. any factor (Rawlings et al., 1998) to check assumptions. Normal quantile-quantile (Q-Q) plots of std. residuals vs. normal theoretical z are used to check violation of normality. Outliers (Onoz & Oguz, 2003; Zimmerman, 2010; Liu et al., 2012; Jones, 2019) are spotted in a Q-Q plot. Boxplots are used to check outliers which are measured in s units from median. Skewness, γ_1, and kurtosis, γ_2, (Blanca et al., 2013; Blanca et al., 2017; Roser et al., 2020) are to check normality. Both stem from 3rd and 4th moments. According to Blanca et al. (2013),

$$Skewness\ \gamma_1 = \frac{\sqrt{n(n-1)}}{n-2}\frac{m_3}{m_2^{3/2}}$$

The more γ_1 is closer to 0, the more the distribution is symmetrical.

$$Kurtosis\ \gamma_2 = \frac{(n-1)}{(n-2)(n-3)}\left\{(n+1)\left(\frac{m_4}{m_2^2} - 3\right) + 6\right\}$$

where, n=sample size, $m_k = \sum_{i=1}^{n}(y_i - \bar{y})^k$.

When γ_2 is closer to 0, this indicates that the distribution is peaked as $\sim N(\mu = 0, \sigma^2 = 1)$, and if $\gamma_2 > 0$, it is more peaked and if $\gamma_2 < 0$, it is more flattened.

In a univariate ANOVA model, for testing whether the estimated residual error variances $\hat{\sigma}_e^2$ are homogeneous among treatment groups, the $std.\hat{e}_{ijl}$, plotted vs. fitted $\hat{y}_{ijl}$ (Kim, 2019). It is also tested using Levene's test (Kim & Cribbie, 2018; Zhou et al., 2023). Levene's W tests:

$H_0: \sigma_i^2 = 0\ vs.\ H_A: \sigma_i^2 \neq 0,$

$i = 1 \ldots . k\ populations\ at\ \alpha$

$$W = \frac{(N-k)}{(k-1)}\frac{\sum_{i=1}^{k} N_i(\bar{Z}_{i.} - \bar{Z}_{..})^2}{\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{ij} - \bar{Z}_{i.})^2}$$

where N_i is size of i[th] treatment, and $Z_{ij} = |Y_{ij} - \bar{Y}_{i.}|$, where $\bar{Y}_{i.}$ is the i[th] treatment mean. $W < F_{\alpha,k-1,N-k}$, H_0 is failed to reject.

For repeated-measure ANOVA, sphericity is to check variance homogeneity. Mauchly's W test of sphericity is a $\chi^2_{\alpha,k-1}$ (Haverkamp & Beauducel, 2017), (Blanca, Arnau, García-Castro, et al., 2023). If sphericity is met, H_0 is failed to reject, $H_0: \sigma^2_{(Y_i - Y_{\acute{i}})} = 0$, $H_A: \sigma^2_{(Y_i - Y_{\acute{i}})} \neq 0$ $\quad i \neq \acute{i}$ where $\sigma^2_{(Y_i - Y_{\acute{i}})}$ is the variance of a difference between pairwise treatment, $\sigma^2_{(Y_i - Y_{\acute{i}})} = \sigma^2_{Y_i} + \sigma^2_{Y_{\acute{i}}} - 2\rho\sigma_{Y_i}\sigma_{Y_{\acute{i}}}$, where ρ is the correlation coefficient.

If sphericity is violated, there is a resolution to apply a correction factor, the Greenhouse & Geisser Epsilon (ϵ) (Lane, 2016), to reduce df in the F test. ϵ value ranges (1-k)$^{-1}$-1.0). The ϵ is estimated by,

$$\hat{\epsilon}_{G-G} = \frac{K^2(\overline{diag\ s} - \bar{s})^2}{(K-1)(\sum_{i=1}^{k}\sum_{j=1}^{k} S_{i,j}^2 - 2K\sum_{j=1}^{k}\bar{S}_{i.}^2 + K^2\bar{S}^2)}$$

where i & j are the rows and columns of S, the covariance matrix, and $\bar{S}$ mean of all elements in S, $\overline{diag\ S}$=diagonal mean, mean of variances.

In addition, Pillai's Trace and Wilki's Lambda, Λ, are estimated (Ateş et al., 2019).

$$Pillai's\ Trace\ \boldsymbol{V} = trace\ [\boldsymbol{H}(\boldsymbol{E} + \boldsymbol{H})^{-1}]$$

where, **H** is the treatment SS and cross product matrix, and **E**= error SS and CP matrix. For p x p **A** matrix, the Trace (**A**)=$\sum_{i=1}^{p} a_{ii}$.

$$\mathbf{T} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{Y_{ij}} - \bar{\boldsymbol{y}}_{i.})(\boldsymbol{Y_{ij}} - \bar{\boldsymbol{y}}_{i.})^{\boldsymbol{T}} + \sum_{i=1}^{k}\boldsymbol{n_i}(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})^{\boldsymbol{T}}$$

H_0 is rejected if $\boldsymbol{V}$ is high. Wilk's Lambda $\boldsymbol{\Lambda} = |\mathbf{E}|/|\mathbf{H} + \mathbf{E}|$. H_0 is rejected if Wilk's lambda is near zero. Analyses were done using the IBM SPSS 23.0.

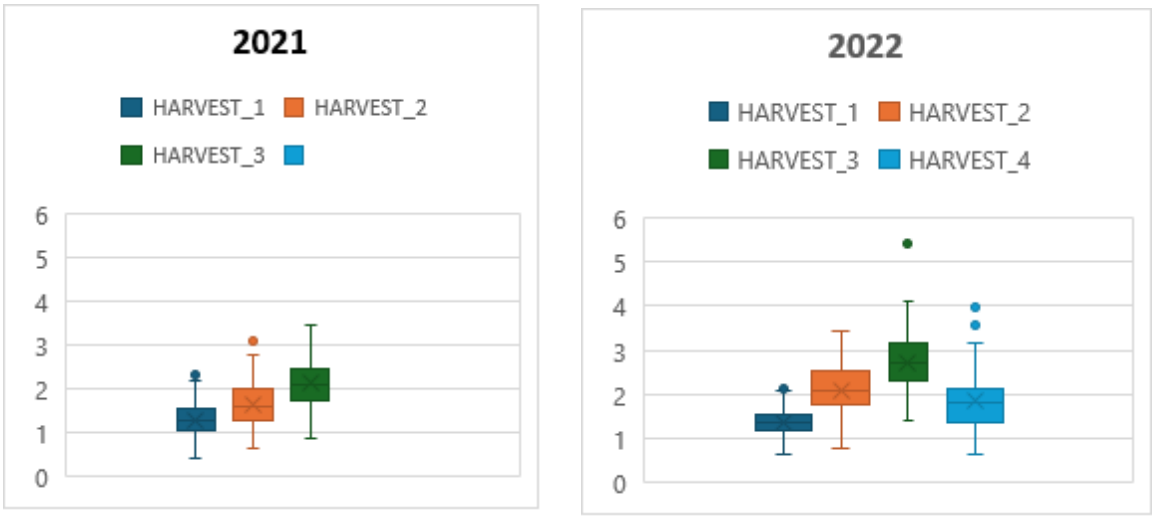
Results and Discussion

Between-Factors Model

Boxplot used to inspect outliers (Williamson et al., 1989) for dry forage yield for harvests in the 2 years (Figs. 1 & 2). Figure 1 shows 2 outliers in Year 1 and 4 in Year 2. According to Wainer (1976), a fringelier lies ≥ 3 s. Draper & Smith (1981, pp.152-153) considered any $|e_i|$-$\bar{e}_i$, $\geq 3s - 4s$ an outlier. Another way to check outliers is to estimate IQR=(Q_3-Q_1) (Whaley, 2005). IQR ranged ~ 0.50-0.75 and ~ 0.40-0.80 in the 2 yr. The Q-Q plot (Huang, 2007; Rousseeuw & Hubert, 2011; Hawkins, 2023) (Fig. 2) indicated that z scores lie off the normal 45^0 line.

Outliers in this study do not seem ingenuine; so, we did not correct for. Kozak et al. (2015) concluded that both means and C.Is. are affected by faulty numbers. We think that even if outliers result from mistakes, their discarding results in unbalanced data. Kozak et al. (2015) indicated that outlier in 'large' n is less likely to occur. Their claim though is theoretically true; still, how 'small' is small? Wu & Zuo (2009) suggested robust estimates such as trimmed and winsorized means. We assert, however, that uncontrolled complex spatial/temporal variations may cause outliers.

Fig. 1. Box plot of total forage yield per harvest in 2021 and 2022.



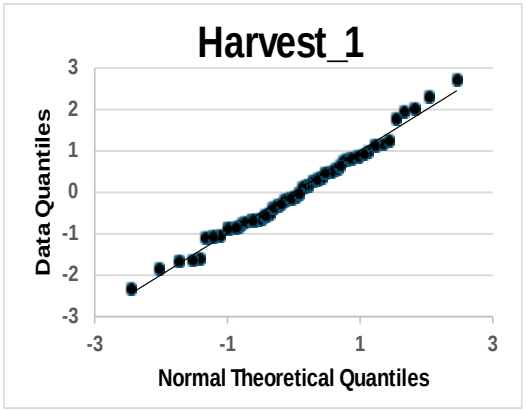
(Rawlings et al., 1998) stated that normal residuals are not necessarily required for parameter estimation and partition of total variation; yet nonnormality affects significance testing and C.Is. Normality is checked by normal Q-Q (Fig. 2) and by HT of skewness γ_1 and kurtosis γ_2, and by Shapiro-Wilk's test (Table 2). Residual errors ($\mu = 0, \sigma^2 = 1$) are represented by the fitted line intercept=0 and slope=1 (Fig. 2). Majority of std. $\hat{e}_{ijl}$ in both years lie on the fitted lines, implying that nonnormality was not a problem. Yet, there exist 'heavy' tails in Year 2. Checking skewness γ_1 and kurtosis γ_2 HT (Table 2), nearly all z scores for the 2 parameters in the 2 years were $<Z_{0.025} = \pm 1.96$. Overall, H_0 failed to be rejected. Surprisingly, in Year 2, Harvest 4, there lies a point with an estimated residual near $3.5\,s$; however, HT for skewness had a $p > 0.05$. This concurs with what Kozak & Piepho (2018) reasoned that some variation might be expected when using residuals plots and HT for checking assumptions. The Shapiro-Wilk's test (Table 2) generally showed p>0.05.

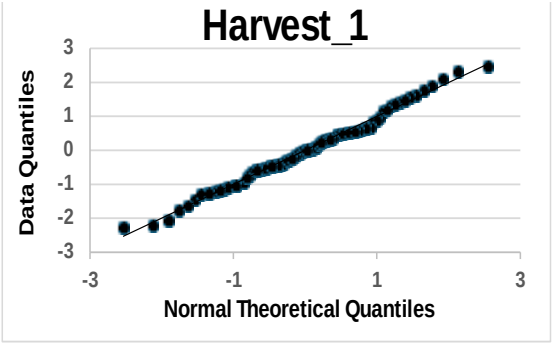
Table 2. Skewness, kurtosis and Shapiro-wilk test for normality of forage yield of harvest in 2021 and 2022.

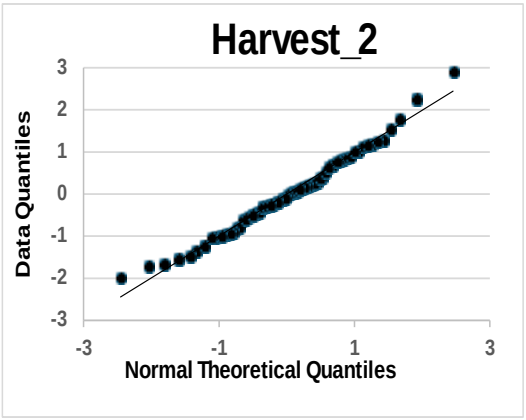
	N	Skewness		Kurtosis		Shapiro Wilk test	
	Statistic	Statistic	Std. Error	Statistic	Std. Error	Statistic	Sig.
				2021			
Harvest 1	72	0.314	0.283	0.255	0.559	0.986	0.613
Harvest 2	72	0.425	0.283	0.216	0.559	0.983	0.463
Harvest 3	72	0.124	0.283	0.169	0.559	0.994	0.984
				2022			
Harvest 1	90	0.440	0.254	- 0.281	0.503	0.990	0.707
Harvest 2	90	- 0.167	0.254	- 0.204	0.503	0.990	0.753
Harvest 3	90	0.513	0.254	0.450	0.503	0.967	0.021
Harvest 4	90	0.372	0.254	0.401	0.503	0.972	0.048

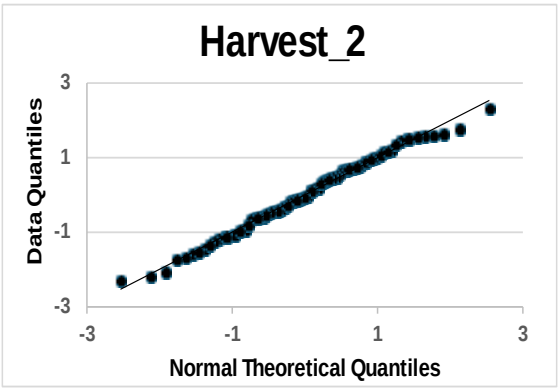
Fig. 2. Normal q-q plot of total forage yield per harvest in 2021 and 2022.

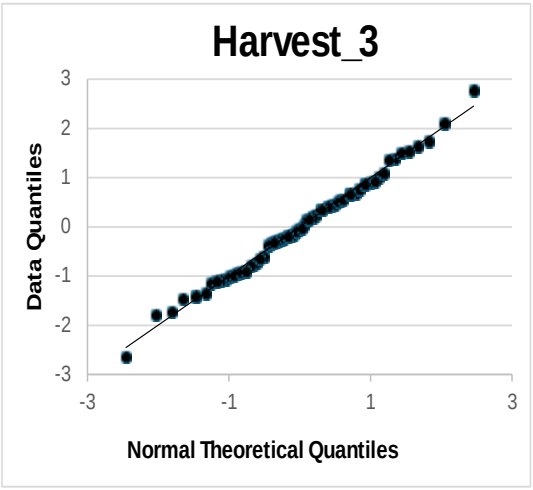
2021 (a)	2022 (b)

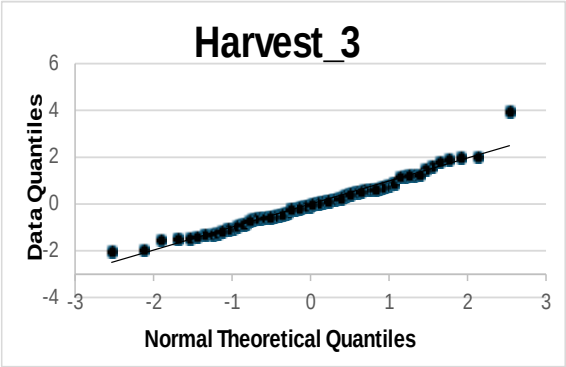


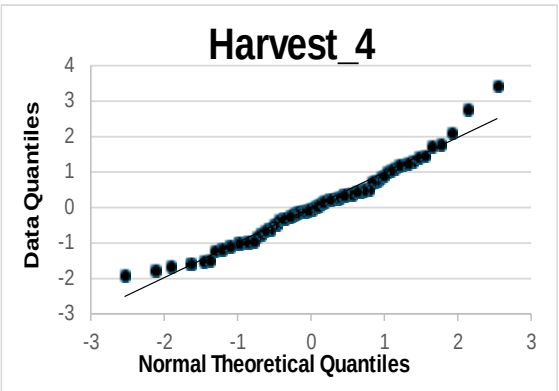












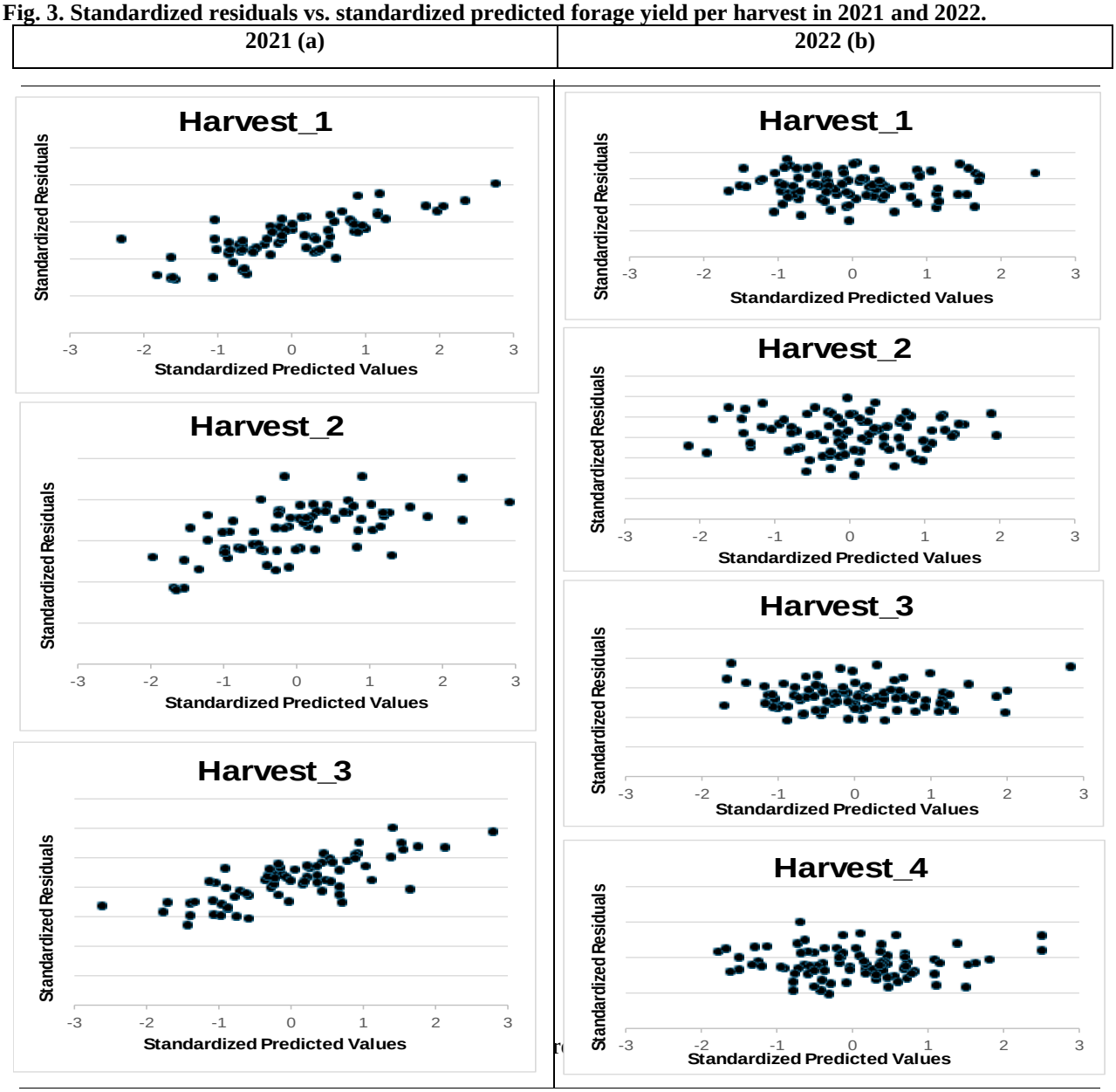
Variance homogeneity is checked by Levene's fitted observations (Fig. 3), vs. clover component (Fig. 4), and vs. forage mixtures (Fig. 5). Levene's declared a heterogeneous variance ($p<0.05$) in Year 1 but homogeneous one ($p > 0.05$) in Year 2. Heterogeneity in Year 1 was disappointing enough to call for increasing blocks to 5 and harvests to 4 in Year 2. Levene's test for harvests in Year 1 had a p

value >0.05 which all contradicted that of total yield. Each harvest's variance homogeneity is quite more informative to relate this to total yield's variance homogeneity. In Year 1 (Fig. 3), the pattern of bands indicated that the residuals linearly increased to declare inconsistent variance, whereas in Year 2 the pattern (Fig. 3) indicated consistent variance.

Table 3. Levene's test of variance homogeneity of harvest and of total yield in 2021 and 2022.

Subjects Effect	F	df 1	df 2	Sig.
		2021		
Harvest 1	1.721	17	54	.067
Harvest 2	1.229	17	54	.275
Harvest 3	.780	17	54	.707
Total Harvest	1.900	17	54	.038
		2022		
Harvest 1	.881	17	72	.597
Harvest 2	.762	17	72	.729
Harvest 3	1.171	17	72	.310
Harvest 4	2.062	17	72	.018
Total Harvest	1.463	17	72	.134

Fig. 3. Standardized residuals vs. standardized predicted forage yield per harvest in 2021 and 2022.



5a), and had outliers (Fig. 4b &5b). More increasing spread means increasing variance with the funnel shape. Kim & Cribbie (2018) emphasized that if H_0 is rejected for variance test, a robust test should be adopted. HT might inflate type-1 error vs nominal α (Zimmerman, 2004)Rasch et al., 2011). HT usually results in poor power, since the aim is to fail to reject H_0. The power of identifying variance homogeneity is negatively related to group sizes even if differences are still the same Kim & Cribbie (2018). Our results (Table 3), quite the opposite, did not support their claim regarding group sizes since increasing replications led us to entertain H_0, but we can't admit what the power was. The authors used 'hedging' statements regarding sample size. Unfortunately, based on their claim, unethical misconduct may manipulate HT by managing group sizes. Altman & Bland (1995) hold that failing to reject H_0 does not most likely verify what it hypothesizes but lack enough proof for rejecting. Ruscio & Roche (2012) assured that β might be committed even under normality. Moreover, HT of variance testing seldom maintains reasonable $1-\beta$ to spot variations even raising nominal α to most likely lead to unsatisfactory power.

Neither of the aforementioned studies has suggested graphics to check variance heterogeneity. Kozak & Piepho (2018) supported graphics to explain what HT cannot, regarding outliers. In ours, graphics indicated that violation in Year 1 cannot be overlooked. We relied more on what graphic tools indicated. Therefore, we abandoned discussing the ANOVA's F test for Year 1 (Table 4). Some studies (Zimmerman, 2006; Ruscio & Roche, 2012; Zhou et al., 2023) adopted exploring factors which might affect variance homogeneity, violation on type-1 error and power and suggested remedies.

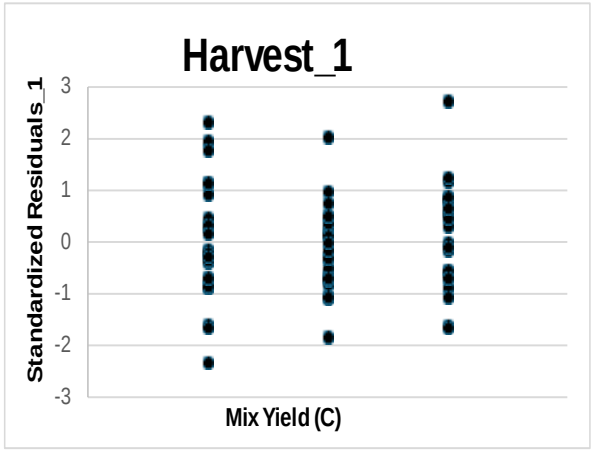
The ANOVAs of the dry forage yield are shown (Table 4). To test the adequacy of the fitted model, R_d^2, were 0.4630 in Year 1 and 0.5105 in Year 2, both were below being acceptable. The residual error not contributed to variation in the response is $\eta^2(eta^2)$ which was 53.69% in Year 1 and 48.94 % in Year 2. Partial eta^2, η_p^2 explains how any effect is ruled out (Levine & Hullett, 2002); Richardson, 2011; Norouzian & Plonsky, 2018).

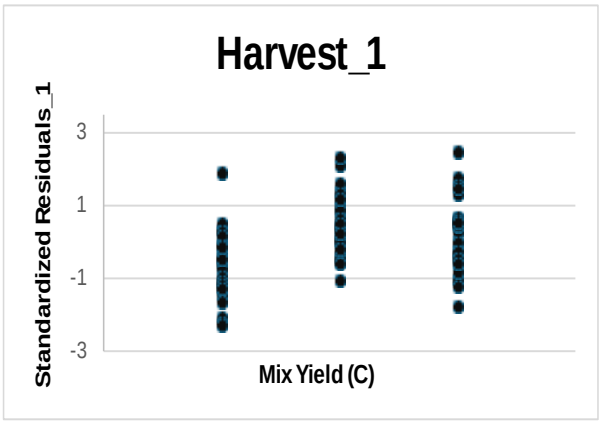
Table 4. Anova of clover cultivar component in forage yield, percentage mixture and interaction in 2021 and 2022.

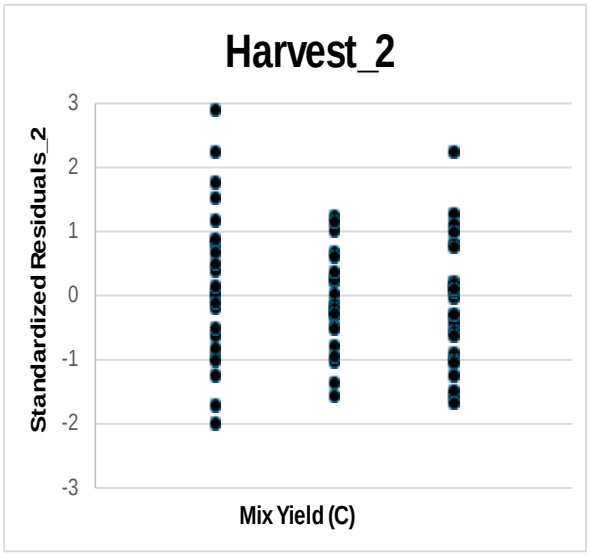
Tests Between Subjects Effects						
Source	**Type lll Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**	**Partial Eta Squared**
			2021			
Block	6.202	3	2.067	9.724	0.000	0.364
Cultivar	0.256	2	0.128	0.601	0.552	0.023
Mix	1.017	5	0.203	0.956	0.453	0.086
Cultivar * Mix	1.877	10	0.188	0.883	0.555	0.148
Error	10.843	51	0.213			
			2022			
Block	4.520	4	1.130	2.783	0.033	0.141
Cultivar	10.031	2	5.015	12.349	0.000	0.266
Mix	5.705	5	1.141	2.809	0.023	0.171
Cultivar * Mix	8.547	10	0.855	2.103	0.036	0.236
Error	27.617	68	0.406			

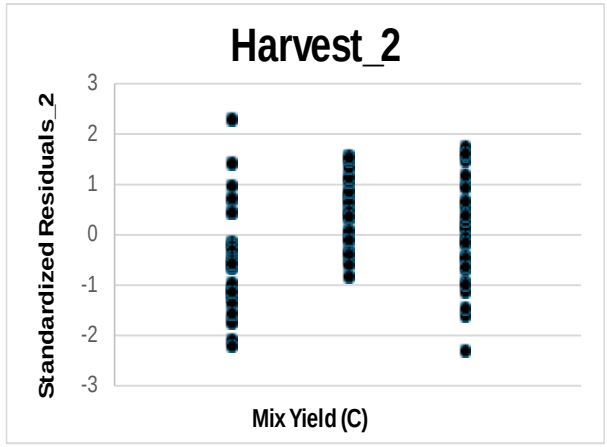
Fig. 4. Standardized residuals per clover cultivar in 2021 and 2022.

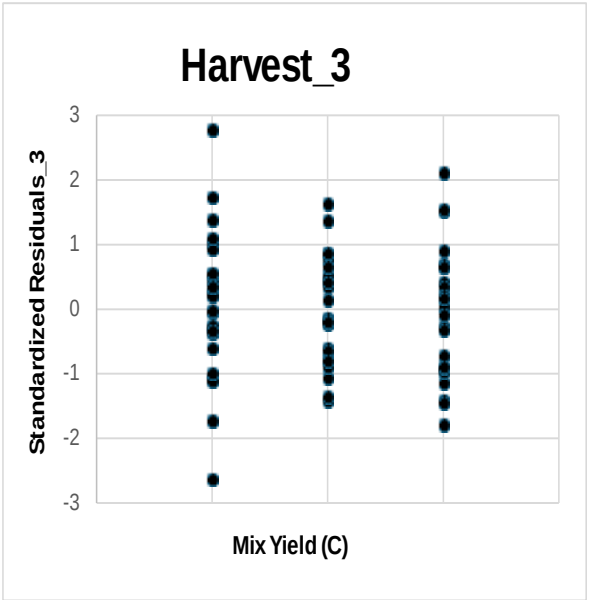
2021 (a)	2022 (b)

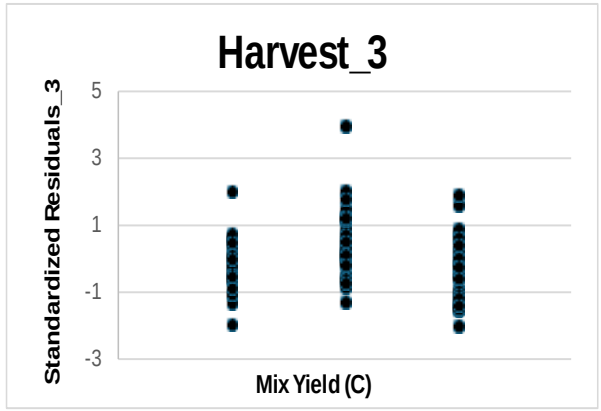












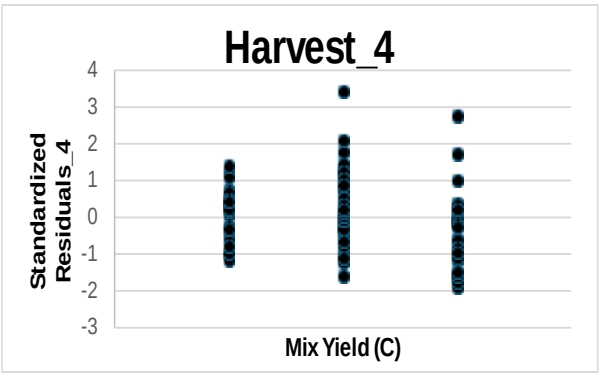
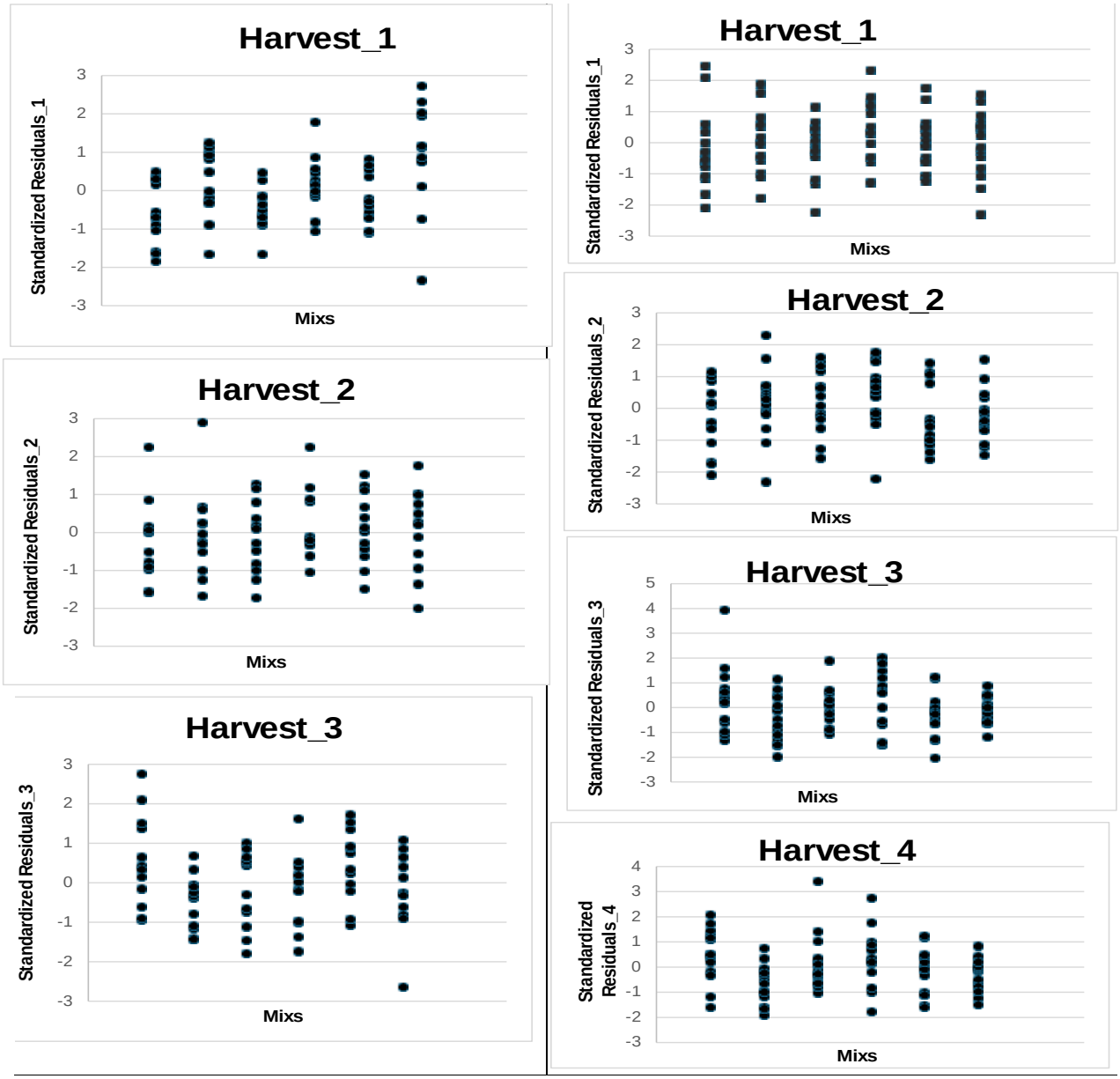


Fig. 5. Standardized residuals per mixture in 2021 and 2022.

2021 (a)	2022 (b)



Repeated-Measures Model

Variance homogeneity of repeated harvests need to be checked in the 2 yr. Mauchly's W test criterion of sphericity (Table 5) is applied, given that normality holds. Repeated measures ANOVA were robust with nonnormality and valid sphericity, despite extreme skewness and kurtosis. Moreover, power did not decline with nonnormality (Blanca, Arnau, García-castro, et al., 2023).

Moulton (2012) stressed that sphericity characterizes a state of both covariance and between-levels variance homogeneity. For small sample size ($n < (k + 10)$), however, W test lacks enough power for detecting deviation from sphericity. Moulton (2012) did not indicate whether the above inequality was mathematically or empirically based. Our samples were much $> (k + 10)$ --72 and 90. Mauchly's W declared minor violation of sphericity in the 2 yr ($p \leq 0.048$) (Table 5). The closer W is to 1.0, the less violation. This dictated adjusting df, using either GG's $\hat{\varepsilon}$ or HF's $\tilde{\varepsilon}$ --$\hat{\varepsilon}$ is less biased for low ε, whereas $\tilde{\varepsilon}$ is less biased for $\varepsilon > 0.75$ (Huynh & Feldt, 1976). Thereby, HF is recommended: $\tilde{\varepsilon}$ = 1.0 with LB=$(k\text{-}1)^{-1}$=0.50&0.33 in both yr. Blanca et al. (2023a) recommend using GG for $\epsilon < 0.6$, and HF for $\varepsilon \geq 0.60$. Furthermore, the more violated sphericity, the more liberal F test but both corrections control inflated type-1 error.

Table 5. Mauchly's test for sphericity, greenhouse-geisser and huynh-feldt for forage yield within harvest in 2021 and 2022.

Within subjects' effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
			2021				
Harvest	0.866	7.184	2	0.028	0.882	1.000	0.500
			2022				
Harvest	0.846	11.188	5	0.048	0.915	1.000	0.333

Repeated-harvests main and interactions effects (Tables 6a & 7a), sphericity's adjusted df (Tables 6b & 7b). Pillai's Trace and Wilks' Lambda –one was closer to 1.0 and the other to 0.0-- for harvests, indicating it contributed relatively more, this was supported by high η_p^2 (85% in Year 1 and 91% in Year 2) (Tables 6a & 7a). With minor violated sphericity, total dry forage yield substantially varied among harvests in both years ($p<0.05$) with HF's η_p^2 of 69% and 73% in the 2 yr (Tables 6b & 7b).

Table 6. Pillai's trace, wilks' lambda, ANOVA for within harvests and interactions for forage yield in 2021.

a- **Multivariate Tests**

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Harvest	Pillai's Trace	0.855	147.936	2	50	0.000	0.855
	Wilks' Lambda	0.145	147.936	2	50	0.000	0.855
Harvest*Block	Pillai's Trace	0.494	5.570	6	102	0.00	0.247
	Wilks' Lambda	0.515	6.566	6	100	0.00	0.283
Harvest * Cultivar	Pillai's Trace	0.088	1.167	4	102	0.330	0.044
	Wilks' Lambda	0.913	1.164	4	100	0.331	0.045
Harvest * Mix	Pillai's Trace	0.540	3.772	10	102	0.000	0.270
	Wilks' Lambda	0.483	4.385	10	100	0.000	0.305
Harvest * Cultivar * Mix	Pillai's Trace	0.445	1.458	20	102	0.114	0.222
	Wilks' Lambda	0.602	1.446	20	100	0.119	0.224

b- **Tests of Within-Subjects Effects**

Source		Type lll Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Harvest	Sphericity Assumed	24.921	2	12.461	111.574	0.000	0.686
	Huynh-Feldt	24.921	2	12.461	111.574	0.000	0.686
Harvest*Block	Sphericity Assumed	4.595	6	0.766	6.858	0.000	0.287
	Huynh-Feldt	4.595	6	0.766	6.858	0.000	0.287
Harvest * Cultivar	Sphericity Assumed	0.397	4	0.099	0.888	0.474	0.034
	Huynh-Feldt	0.397	4	0.099	0.888	0.474	0.034
Harvest * Mix	Sphericity Assumed	4.450	10	0.445	3.985	0.000	0.281
	Huynh-Feldt	4.450	10	0.445	3.985	0.000	0.281
Harvest * Cultivar * Mix	Sphericity Assumed	3.413	20	0.171	1.528	0.088	0.231
	Huynh-Feldt	3.413	20	0.171	1.528	0.088	0.231
Error (Harvest)	Sphericity Assumed	11.391	102	0.112			
	Huynh-Feldt	11.391	102	0.112			

Table 7. Pillai's trace, wilks' lambda, ANOVA for within harvests and interactions for forage yield in 2022.

a- **Multivariate Tests**

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Harvest	Pillai's Trace	0.907	213.7	3	66	0.000	0.907
	Wilks' Lambda	0.093	213.7	3	66	0.000	0.907
Harvest*Block	Pillai's Trace	0.473	3.185	12	204	0.000	0.158
	Wilks' Lambda	0.577	3.370	12	174	0.000	0.168
Harvest * Cultivar	Pillai's Trace	0.364	4.977	6	134	0.000	0.182
	Wilks' Lambda	0.659	5.103	6	132	0.000	0.188
Harvest * Mix	Pillai's Trace	0.466	2.500	15	204	0.002	0.155
	Wilks' Lambda	0.581	2.644	15	182.598	0.001	0.165
Harvest * Cultivar * Mix	Pillai's Trace	0.417	1.098	30	204	0.341	0.139
	Wilks' Lambda	0.637	1.078	30	194.399	0.367	0.140

b- **Tests of Within-Subjects Effects**

Source		Type lll Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Harvest	Sphericity Assumed	87.228	3	29.076	184.953	0.000	0.731
	Huynh-Feldt	87.228	3	29.076	184.953	0.000	0.731
Harvest*Block	Sphericity Assumed	7.630	12	0.636	4.045	0.000	0.192
	Huynh-Feldt	7.630	12	0.636	4.045	0.000	0.192
Harvest * Cultivar	Sphericity Assumed	5.193	6	0.866	5.506	0.000	0.139
	Huynh-Feldt	5.193	6	0.866	5.506	0.000	0.139
Harvest * Mix	Sphericity Assumed	6.660	15	0.444	2.824	0.001	0.172
	Huynh-Feldt	6.660	15	0.444	2.824	0.001	0.172
Harvest * Cultivar * Mix	Sphericity Assumed	5.125	30	0.171	1.087	0.355	0.138
	Huynh-Feldt	5.125	30	0.171	1.087	0.355	0.138
Error (Harvest)	Sphericity Assumed	32.070	204	0.157			
	Huynh-Feldt	32.070	204	0.157			

Generally, in forage trials, the final mean response is summed over harvests. This is a rational practice, given controllable field management and variation; it might be a real problem otherwise. Inference based on this sum may obscure any variational differences, especially with increasing harvest number, the more repeated measures, the greater the violation of sphericity. Not only did the main effect of harvests show tremendous differences but so did its first-order interactions. It is worth mentioning that variations in forage yield were great due to harvest number given clover cultivars in the mixture, when harvests increased to 4.

Temporal variation in forage yield is most likely to occur especially with varying percentage species in a mixture. This variation may lead to violate sphericity but not necessarily makes repeated measures F test invalid. Surprisingly, contrary to theory, sphericity though violated in this experiment was trivial. This was initially based on HF correction reaching its maximum value, as if sphericity were not violated.

ANOVA's F test should be conservatively applied since sphericity is likely violated. Also, Mauchly's validity has been questioned for it is based upon failing to reject H_o to make type 2 error likely be committed. Kim & Cribbie (2018) warned against Levene's test H_0, causing a reduction of power. Blanca et al. (2023a) reported that Mauchly's W, under non-normality, neither did it maintain type 1 error nor was it sensitive to little violation of sphericity. The authors doubt its validity as a 'gatekeeper'. Moulton (2012) concluded that trivially violated sphericity ($\varepsilon > 0.70$) declared by significant W, might result from large sample size.

There have been innumerable diagnostic and judgmental tools testing violations of assumptions, especially sphericity violation's effects on both type 1 error and power. Unfortunately, studies have not been conducted in agronomy. Hence, this gap calls attention to and to consider other statistical perspectives (e.g. MANOVA, Moulton (2012)). The situation gets more complicated in the case of testing main and interaction effects using univariate repeated measures ANOVA; this is not fully understood by

researchers (Langenberg et al., 2023). They suggested 'Structural Equation Models' to handle sphericity or to relax its assumption.

Acknowledgements

We deeply value the farm crew's help during all field management, harvesting, collecting and weighing samples for 7 harvests. We show gratitude for the students who volunteered in lab work.

References

Abdi, H. (2010). The Greenhose-Geisser correction. In Nril Salkind (ed.), Encyclopedia of Research Design. (pp. 1–10). https://doi.org/10.1017/cbo9780511606663.003

Akritas, M., & Arnold, S. (2000). Asymptotics for analysis of variance when the number of levels is large. Journal of the American Statistical Association, *95*(449), 212–226. https://doi.org/10.1080/01621459.2000.10473915

Akritas, M. G., & Papadatos, N. (2004). Heteroscedastic one-way ANOVA and lack-of-fit tests. Journal of the American Statistical Association, *99*(466), 368–382. https://doi.org/10.1198/016214504000000412

Altman, Douglas G., Bland, J. M. (1995). Statistical notes: Absence of evidence is not evidence of absence. BMJ, *311*, 485. https://doi.org/doi: https://doi.org/10.1136/bmj.311.7003.485

Arnold, S. F. (1980). Asymptotic validity of F tests for the ordinary linear model and the multiple correlation model. Journal of the American Statistical Association, *75*(372), 890–894. https://doi.org/10.1080/01621459.1980.10477568

Ateş, C., Kaymaz, Ö., Kale, H. E., & Tekindal, M. A. (2019). Comparison of test statistics of nonnormal and unbalanced samples for multivariate analysis of variance in terms of type-I error rates. Computational and Mathematical Methods in Medicine. https://doi.org/10.1155/2019/2173638

Bathke, A. (2004). The ANOVA F test can still be used in some balanced designs with unequal variances and nonnormal data. Journal of Statistical Planning and Inference, *126*(2), 413–422. https://doi.org/10.1016/j.jspi.2003.09.010

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? Psicothema, *29*(4), 552–557. https://doi.org/10.7334/psicothema2016.383

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2018). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit? Behavior Research Methods, *50*(3), 937–962. https://doi.org/10.3758/s13428-017-0918-2

Blanca, M. J., Arnau, J., García-castro, F. J., Alarcón, R., & Bono, R. (2023). Non-normal data in repeated measures ANOVA : Impact on type I rrror and power. *Psicothema*, *35*(1), 21–29.

Blanca, M. J., Arnau, J., García-Castro, F. J., Alarcón, R., & Bono, R. (2023). Repeated measures ANOVA and adjusted F-tests when sphericity is violated: which procedure is best? Frontiers in Psychology, 14, 1–11. https://doi.org/10.3389/fpsyg.2023.1192453

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. Methodology, *9*(2), 78–84. https://doi.org/10.1027/1614-2241/a000057

Boos, D. D., & Brownie, C. (1995). ANOVA and rank tests when the number of treatments is large. Statistics and Probability Letters, *23*(2), 183–191. https://doi.org/10.1016/0167-7152(94)00112-L

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. The Annals of Mathematical Statistics, *25*(2), 290–302. https://doi.org/10.1214/aoms/1177728786

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. Journal of the American Statistical Association, *69*(346), 364–367. https://doi.org/10.1080/01621459.1974.10482955

Debashis, P. (2013). Effects of violations of model assumptions. *Statistics Libretexts*, *3*, 2–4. https://stats.libretexts.org/Core/Statistical_Computing/Analysis_of_Variance/Effects_of_violations_of_model assumptions

Draper, N.R. & Smith, H. (1981). Applied regression analysis (second ed.). John Wiley & Sons, Inc.

Driscoll, W. C. (1996). Robustness of the ANOVA and Tukey-Kramer statistical tests. Computers and Industrial Engineering, *31*(1–2), 265–268. https://doi.org/10.1016/0360-8352(96)00127-1

Halldestam, M. (2016). ANOVA - The effect of outliers. http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A935253&dswid=-1723

Haverkamp, N., & Beauducel, A. (2017). Violation of the sphericity assumption and its effect on type-I error rates in repeated measures ANOVA and multi-level linear models (MLM). Frontiers in Psychology, *8*, 1–12. https://doi.org/10.3389/fpsyg.2017.01841

Hawkins, D. M. (2023). Quantile-Quantile methodology-detailed results. *ArXiv:2303.03215v2 [Stat.ME]*, 1–38. https://doi.org/10.48550/arXiv.2303.03215

Huang, M. L. (2007). A quantile-score test. *1*(11),

507–516.
Huynh, H., and Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. Journal Of Educational Statistics, *1*(1), 69–82. https://doi.org/doi: 10.2307/1164736
Jones, P. R. (2019). A note on detecting statistical outliers in psychophysical data. Attention, Perception, and Psychophysics, *81*(5), 1189–1196. https://doi.org/10.3758/s13414-019-01726-3
Kim, H.-Y. (2015). Statistical notes for clinical researchers: A one-way repeated measures ANOVA for data with repeated observations. Restorative Dentistry & Endodontics, *40*(1), 91. https://doi.org/10.5395/rde.2015.40.1.91
Kim, H.-Y. (2019). Statistical notes for clinical researchers: simple linear regression 3 – residual analysis. Restorative Dentistry & Endodontics, *44*(1), 1–8. https://doi.org/10.5395/rde.2019.44.e11
Kim, Y. J., & Cribbie, R. A. (2018). ANOVA and the variance homogeneity assumption: Exploring a better gatekeeper. British Journal of Mathematical and Statistical Psychology, *71*(1), 1–12. https://doi.org/10.1111/bmsp.12103
Kohr, R. L., & Games, P. A. (1974). Robustness of the analysis of variance, the welch procedure and a box procedure to heterogeneous variances. Journal of Experimental Education, *43*(1), 61–69. https://doi.org/10.1080/00220973.1974.10806305
Kozak, M., Krzanowski, W., Cichocka, I., & Hartley, J. (2015). The effects of data input errors on subsequent statistical inference. Journal of Applied Statistics, *42*(9), 2030–2037. https://doi.org/10.1080/02664763.2015.1016410
Kozak, M., & Piepho, H. P. (2018). What's normal anyway? Residual plots are more telling than significance tests when checking ANOVA assumptions. Journal of Agronomy and Crop Science, *204*(1), 86–98. https://doi.org/10.1111/jac.12220
Kvam, P. H., & Vidakovic, B. (2007). Nonparametric statistics with applications to science and engineering. *In* Nonparametric Statistics with Applications to Science and Engineering. https://doi.org/10.1002/9780470168707
Lane, D. M. (2016). The assumption of sphericity in repeated-measures designs: What it means and what to do when it is violated. The Quantitative Methods for Psychology, *12*(2), 114–122. https://doi.org/10.20982/tqmp.12.2.p114
Langenberg, B., Helm, J. L., Günther, T., & Mayer, A. (2023). Understanding, testing, and relaxing sphericity of repeated measures ANOVA with manifest and latent variables using SEM. Methodology, 19(1), 60–95. https://doi.org/10.5964/meth.8415
Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. Human Communication Research, *28*(4), 612–625. https://doi.org/10.1111/j.1468-2958.2002.tb00828.x
Lind, J. C., & Zumbo, B. D. (1993). The continuity principle in psychological research: An introduction to robust statistics. Canadian Psychology / Psychologie Canadienne, *34*(4), 407–414. https://doi.org/10.1037/h0078861
Liu, Y., Zumbo, B. D., & Wu, A. D. (2012). A Demonstration of the impact of outliers on the decisions about the number of factors in exploratory factor analysis. Educational and Psychological Measurement, *72*(2), 181–199. https://doi.org/10.1177/0013164411410878
Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. Review of Educational Research, *66*(4), 579–619. https://doi.org/10.3102/00346543066004579
Marcinko, T. (2014). Consequences of assumption violations regarding one-way anova. The 8th International Days of Statistics and Economics, 974–985. https://msed.vse.cz/static/msed_2014/article/342-Marcinko-Tomas-paper.pdf
Moulton, S. L. (2012). Mauchly test. In Encyclopedia of Research Design (1–5). https://doi.org/DOI:https://doi.org/10.4135/9781412961288
Norouzian, R., & Plonsky, L. (2018). Eta- and partial eta-squared in L2 research: A cautionary review and guide to more appropriate usage. *Second Language Research*, *34*(2), 257–271. https://doi.org/10.1177/0267658316684904
Onoz, B., & Oguz, B. (2003). Assessment of outliers in statistical data analysis. Integrated Technologies for Environmental Monitoring and Information Production, 173–180. https://doi.org/10.1007/978-94-010-0231-8_13
Orr, J. M., Sackett, P. R., & Dubois, C. L. Z. (1991). Outlier detection and treatment in I/O psychology: A survey. Personnel Psychology, *44*, 473–487.
Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should ALWAYS check for them). Practical Assessment, Research and Evaluation, 9(6).
Quintana, S. M., & Maxwell, S. E. (1994). A Monte Carlo comparison of seven ε-adjustment procedures in repeated measures designs with small sample sizes. Journal of Educational Statistics, *19*(1), 57–71.

https://doi.org/10.3102/10769986019001057
Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample t test: Pre-testing its assumptions does not pay off. Statistical Papers, *52*(1), 219–231. https://doi.org/10.1007/s00362-009-0224-x
Rawlings, J. O., Pantula, S. G., & Dickey, D. a. (1998). Applied regression analysis: A research tool, Second ed. Springer Texts in Statistics.
Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. Educational Research Review, *6*(2), 135–147. https://doi.org/10.1016/j.edurev.2010.12.001
Rogan, J. C., & Keselman, H. J. (1977). Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal?: An investigation via a coefficient of variation. American Educational Research Journal, *14*(4), 493–498. https://doi.org/10.3102/00028312014004493
Roser, B., Jaume, A., Rafael, A., & Maria, B. J. (2020). Bias, precision, and accuracy of skewness and kurtosis estimators for frequently used continuous distributions. Symmetry, *12*(19), 2–17.
Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1), 73–79. https://doi.org/10.1002/widm.2
Ruscio, J., & Roche, B. (2012). Variance heterogeneity in published psychological research: A review and a new index. Methodology, *8*(1), 1–11. https://doi.org/10.1027/1614-2241/a000034
Schützenmeister, A., Jensen, U., & Piepho, H. P. (2012). Checking normality and homoscedasticity in the general linear model using diagnostic plots. Communications in Statistics: Simulation and Computation, *41*(2), 141–154. https://doi.org/10.1080/03610918.2011.582560
Schützenmeister, A., & Piepho, H. P. (2012). Residual analysis of linear mixed models using a simulation approach. Computational Statistics and Data Analysis, *56*(6), 1405–1416. https://doi.org/10.1016/j.csda.2011.11.006
Shahin, S. (2017). Analysis of messy data. The International Encyclopedia of Communication Research Methods, 1–8. https://doi.org/10.1002/9781118901731.iecrm0152
Snell, J., & Sprent, P. (1995). Applied nonparametric statistical methods. In Journal of the Royal Statistical Society. Series A (Statistics in Society) (158: 2). https://doi.org/10.2307/2983315
Steyn, P. (2021). Data Assumptions: Its about the residuals, and not the variables' raw data. IntroSpective Mode. https://www.introspective-mode.org/assumptions-residuals-variables/
Vandierendonck, A. & De Soete, G. (1983). Some robust statistics for psychologists. *Pscchlo.Belg.*, *23*(1), 73–83.
Wainer, H. (1976). Robust Statistics: A survey and some prescriptions. *1*(4), 285–312.
Wang, L., & Akritas, M. G. (2006). Two-way heteroscedastic anova when the number of levels is large. Statistica Sinica, *16*(4), 1387–1408.
Watkins, M. W. (2018). Exploratory Factor Analysis: A guide to best practice. Journal of Black Psychology, *44*(3), 219–246. https://doi.org/10.1177/0095798418771807
Whaley, D. L. (2005). The Interquartile Range: Theory and Estimation. Digital Commons @ East Tennessee State University
Williamson, D. F., Parker, R. A., & Kendrick, J. S. (1989). The box plot: A simple visual method to interpret data. Annals of Internal Medicine, 110(11), 916–921. https://doi.org/10.7326/0003-4819-110-11-916
Wu, M., & Zuo, Y. (2009). Trimmed and winsorized means based on a scaled deviation. Journal of Statistical Planning and Inference, 139(2), 350–365. https://doi.org/10.1016/j.jspi.2008.03.039
Zaiontz, C. (2023). Box's M test basic concepts. Real Statistics. https://real-statistics.com/multivariate-statistics/boxs-test/boxs-test-basic-concepts/
Zhou, Y., Zhu, Y., & Wong, W. K. (2023). Statistical tests for homogeneity of variance for clinical trials and recommendations. Contemporary Clinical Trials Communications, *33*, 101119. https://doi.org/10.1016/j.conctc.2023.101119
Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. British Journal of Mathematical and Statistical Psychology, *57*(1), 173–181. https://doi.org/10.1348/000711004849222
Zimmerman, D. W. (2006). Two separate effects of variance heterogeneity on the validity and power of significance tests of location. Statistical Methodology, *3*(4), 351–374. https://doi.org/10.1016/j.stamet.2005.10.002