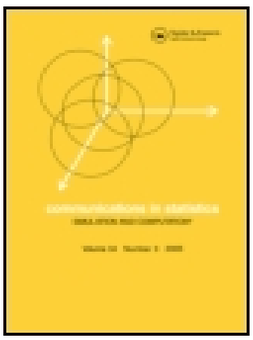

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

The gamma generalized linear model, log transformation, and the robust Yuen-Welch test for analyzing group means with skewed and heteroscedastic data

Victoria K. Y. Ng & Robert A. Cribbie

To cite this article: Victoria K. Y. Ng & Robert A. Cribbie (2018): The gamma generalized linear model, log transformation, and the robust Yuen-Welch test for analyzing group means with skewed and heteroscedastic data, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2018.1440301

To link to this article: https://doi.org/10.1080/03610918.2018.1440301

Published online: 28 Feb 2018.

COMMUNICATIONS IN STATISTICS—SIMULATION AND COMPUTATION®
2018, VOL. 0, NO. 0, 1–18
https://doi.org/10.1080/03610918.2018.1440301

The gamma generalized linear model, log transformation, and the robust Yuen-Welch test for analyzing group means with skewed and heteroscedastic data

Victoria K. Y. Ng and Robert A. Cribbie

Department of Psychology, York University, Toronto, Ontario

ABSTRACT

Alternatives for positively skewed and heteroscedastic data include the Yuen-Welch (YW) test, data transformations, and the generalized linear model (GzLM). Because the GzLM is rarely considered in psychology compared to the other two, we compared these strategies conceptually and empirically. The YW test generally has satisfactory power, but its trimmed mean can deviate substantially from the arithmetic mean, which is often the desired parameter. The gamma GzLM can be used as a substitute for the log transformation and addresses the limitations in inference for the YW and data transformations.

ARTICLE HISTORY
Received 24 August 2016
Accepted 6 February 2018

KEYWORDS
Arithmetic mean; ANOVA; Box-Cox; Gamma; Generalized linear model; Geometric mean; Robust statistics; Transformations; Trimmed mean

MATHEMATICS SUBJECT CLASSIFICATION
62-07 (Data analysis); 00A06 (Mathematics for nonmathematicians); 62F03 (Hypothesis testing); 62G35 (Robustness); 62J12 (Generalized linear models)

1. Introduction

The one-way analysis of variance (ANOVA) is commonly used in psychological research for detecting group mean differences. Under conventional applications, estimation efficiency comes at the cost of strong assumptions about the error structure — homoscedasticity and normality (Fox 2008). When these assumptions are even slightly violated, power suffers (Tukey 1960). When assumptions are grossly violated, statistical inference is rendered invalid (Fox 2008). It is very often the case that assumptions are violated, thereby making the availability of alternatives necessary (Micceri 1989; Golinski and Cribbie 2009) .

For analyzing skewed and heteroscedastic data, two approaches include the Yuen-Welch test (Wilcox 2005) from the robust statistical framework and data transformations. However, neither can make inference to the arithmetic mean. Popularized by Nelder and Wedderburn (1972), though less applied in the analysis of continuous data in psychology, is the generalized linear model (GzLM).

We focus on the analysis of continuous outcomes whose data are non-negative, positively skewed, and heteroscedastic, specifically when the variance is proportional to the mean. These characteristics have been reported to be common with psychological data (Grissom 2000; Micceri 1989). We begin by describing the Yuen-Welch test, power transformations, and their

CONTACT Victoria Ng ng.victoria.ky@gmail.com Department of Psychology, York University, Toronto, Ontario.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

limitations that motivate our consideration of the GzLM. We introduce the GzLM structure, focusing on the gamma model. The simulation then compares these estimators for examining group differences.

1.1. Alternative approaches

1.1.1. Yuen-Welch test

The Yuen-Welch (Wilcox 2005) omnibus test for group differences generalizes Yuen's (1974) proposal to use trimmed means alongside Winsorized variances. The impact of nonnormality is minimized by trimmed means and the associated Winsorized variances; the impact of heteroscedasticity is minimized by a nonpooled standard error and an adjustment to the degrees of freedom. Details on computed terms can be found in Wilcox's text (2005, p. 267).

The trimmed mean is based on the removal of some proportion of cases from distribution tails. Trimming may be done symmetrically or asymmetrically in varying magnitudes, but 20% symmetric trimming has been recommended to applied researchers for maintaining Type I error and power rates under normal and contaminated distributions (Keselman et al. 2002). However, the trimmed mean estimates the population trimmed mean — the location for a distribution whose tails are trimmed away, and therefore unaccounted for. In asymmetric distributions, its confidence intervals generalize to only 60% of the sampled population (Bonett and Price 2002), and the 20% trimmed mean is representative of only the 'typical' responses. If one is indeed interested in only the typical response and power, then the loss of estimator sufficiency may be justified, and recommendations to "bypass classical parametric statistics" altogether (Erceg-Hurn and Mirosevich 2008) may be wholly supported.

However, generalizability is important in some areas. Some variables are typically distributed with positive skewness at the population level, such as response times, clinical dysfunction, and financial costs (Ratcliff 1993; Neal and Simons 2007; Manning 1998; Kilian et al. 2002). Consequently, a researcher may be more satisfied with inferences that do include distribution tails. Consider, for example, the cost of mental health services. Most costs incurred by individuals may fall in some moderate range, but it would not be uncommon to also observe the few who incur much greater costs. Group trimmed mean estimates would provide cost estimates that reflect the typical cases, but any extrapolation would not capture the reality that there are those who consume more resources. Similarly, in reaction time analyses, data in the tails may reflect true processes, and so the ideal analysis should eliminate as few of the meaningful data of interest (Ratcliff 1993).

1.1.2. Raw data transformations

Power transformations can reduce skewness and stabilize variance. When the variance changes with the mean by some power relationship for a strictly positive variable, the Box-Cox class of power transformations is suitable (Box and Cox 1964). The Box-Cox transformation is defined by

$$Y' = \begin{cases} \ln(Y) & \text{if} \quad \lambda = 0 \\ \frac{Y^{\lambda}-1}{\lambda} & \text{if} \quad \lambda \neq 0 \end{cases} \tag{1}$$

where Y is the response variable in original scale, Y' is the corresponding variable in transformed scale, and λ is the power parameter.

The natural log transformation ($\lambda = 0$) is commonly used in psychology; the newly transformed Y' would be submitted to an ANOVA/regression, whereupon the least squares estimation advantage in efficiency could still be harboured (Fox 2008). The model would be

$$\ln(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon, \tag{2}$$

where the log-transformed response, $\ln(Y)$, is expressed as a linear function of p predictor variables, x_1 through x_p, and ϵ is the normally-distributed and homoscedastic error term.

Transformations are most useful when both skewness and heteroscedasticity are simultaneously and fully corrected, but this result is not guaranteed. An inspection of log-transformed residuals may very well indicate the presence of heteroscedasticity in the log error term, thereby invalidating inference (Manning, Mullahy, and Manning 2001). Furthermore, in the case of a single categorical predictor, distributions would need to be made symmetric for each group simultaneously, thereby generally requiring that groups have similar distributions at the outset.

Models for a log transformed outcome yield inferences to the arithmetic mean in log scale, which is equivalent to the geometric mean in original scale. With other Box-Cox transformations, back-transformations yield medians in original scale. An attempt to interpret arithmetic means in the original scale using a naive back-transformation would yield discrepant estimates, as the geometric mean is either equal or lower than the arithmetic mean. Certainly, there are cases in which interpretation for the mean in transformed scale is useful, such as when processes are expected to be multiplicative (e.g., a change in X predicts a 10% change in Y). However, when processes are additive or when sum totals are meaningful, the arithmetic mean is more useful; multiplying the mean by the number of observations yields the total sum, but the same cannot be said for the geometric mean (or the trimmed mean). Econometricians have used re-transformation corrections to obtain arithmetic means (Duan 1983; Manning 1998), but one may also leverage the concept of transformations within the GzLM framework to directly make inferences in original scale.

1.2. The generalized linear model

The generalized linear model is typically estimated by maximum likelihood and relaxes the assumption that residuals are Gaussian-distributed. The distribution could be specified as any of the exponential family of probability distributions, such as the gamma, Poisson, binomial, and inverse-Gaussian. Assumptions are then made about the error distribution and the mean-variance relation (Nelder and Wedderburn 1972). For example, the gamma distribution assumes a specific pattern of heteroscedasticity in which the variance increases proportionally with the mean — specifically, the square of the mean. These assumptions can be assessed using deviance residuals (which are analogous to residual sum of squares in ordinary least squares [OLS]). Residuals should tend towards normality and homoscedasticity for continuous responses.

The GzLM also has the link function, which provides the transformation of the expected values of the outcome:

$$\eta = g(\mu), \tag{3}$$

in which η is the linear predictor of p predictor variables

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{4}$$

When the link function is correctly specified, the relation between the expected values of the outcome and the set of predictors, η, is linearized.

The transformation of the expected value, and not of the outcome itself, provides the additional advantage of interpreting the model's estimates in the outcome's original scale (Blough et al. 1999), which is given by the inverse of the link function.

Unlike raw data transformations, the GzLM allows for the transformation to be specified separately from the outcome's distribution (Fox 2008), highlighting the flexibility of exploring combinations of various link functions and probability distributions. We currently consider a particular GzLM suitable for continuous data, with the specifications for the gamma distribution and the log link function, $\ln(\mu)$. The gamma distribution models strictly positive continuous responses and is characterized by two parameters, shape (α) and scale (s). Expected values are given by the following:

$$E(Y) = \alpha s \tag{5}$$

$$Var(Y) = (\alpha s)^2 . \tag{6}$$

The α parameter controls location, while the scale parameter controls dispersion. A gamma specification implies a mean-variance relation of heteroscedasticity where the variance is proportional to the square of the mean (similar to models using the log transformation).

The log transformation is applied to the expected values, giving the model

$$\ln [E\,(Y)] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{7}$$

or equivalently, in original scale,

$$E\,(Y) = \exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right). \tag{8}$$

The natural log transformation is applied to the expected values (Equation 7), which is different from $E[\ln(y)]$ (Equation 2). From Equation 8, it is readily seen that the inverse function for the log link function (ϵ^{η}) yields a back-transformed inference to the arithmetic mean in original scale.

1.2.1. Evaluating model assumptions

While results from the Yuen-Welch test require no strict assumptions regarding probability distributions, the tenability of results from either the GzLM or log transformed model depends on the degree to which assumptions about normality and heteroscedasticity are met. Methods for evaluating assumptions about the residuals include graphical inspection or formal assumptions tests, such as the Shapiro-Wilk test (1965) for normality or the Brown-Forsythe test (1974) for homoscedasticity. The Shapiro-Wilk test behaves favourably in a variety of contaminated distributions (Chen 1971) and has more power compared to alternatives (Razali and Wah 2011). The Brown-Forsythe is a robust form of the original Levene test (1960) and has been used across many disciplines for exploring trends in variances (Gastwirth et al. 2009). The preliminary use of formal assumptions tests for choosing among statistical methods, however, generally has deleterious effects on error rates of the final hypothesis of interest because error rates compound and interact across stages of tests (Zimmerman 2004; Hayes and Cai 2007; García-Pérez 2012). Graphical inspection may be considered a suitable alternative (Schucany and Ng 2006) but is still considered a preliminary analysis if it serves as a condition upon which statistical decisions are made (García-Pérez 2012). However, the use of *preliminary* assumptions checks for determining subsequent analysis steps is different from the use of assumptions checks for evaluating the tenability of statistical conclusions.

1.2.2. GzLMs for determining group mean differences

We focus on GzLMs for the one-way independent groups design. While misspecification of the family distribution inflates standard errors (Jones 2012), the link function becomes less relevant for model fit, with a single categorical predictor. With k groups, there would be only k unique expected values. For two models with only one categorical predictor that share the same family but not the same link function, equivalent model fit would be obtained because the total deviance would be the same. For example, a gamma GzLM with log link yields the same fit as a gamma GzLM with inverse link when there is a single categorical predictor. Similarly, a Gaussian GzLM with log link yields the same fit as a regular ANOVA. In this setting, then, the link function serves only to provide different model interpretations but not statistical fit. Beyond this setting, however, different family and link specifications require model comparison and evaluation, which may involve inspection of deviance residuals, goodness-of-fit tests, and comparison of information criteria.

2. Research objective

Previous studies have compared log transformations with the gamma GzLM with log link (Manning, Mullahy, and Manning 2001; Neal and Simons 2007; Nevill and Copas 1991), but to our knowledge, GzLMs have not been compared to robust estimators like the Yuen-Welch's. We compare the Yuen-Welch test, the log transformation, and the GzLM (gamma family with log link; Gaussian family with log link) for the one-way independent groups design. With the context that some researchers may prefer both power and inference about the arithmetic mean in original scale, we compared these four methods for Type I error and power, as well as the discrepancy of the alternative estimators (the geometric and trimmed means) from the expected arithmetic mean. Assumptions tests are used to inspect residuals within the simulation and to document residual behaviour across data conditions.

3. Method

Monte Carlo simulations were conducted in R Software (R Development Core Team 2015). Data with variances increasing proportionately with the mean were generated from gamma and Box-Cox transformation processes.

3.1. Data generation

Two sets of expected value configurations were used for power, one set of locations being low and another being high, such that there was a shift in location (one unit or three units) for group two. Thus, the expected mean configurations for power settings were the following: [2, 3, 2], [10, 11, 10], [2, 5, 2], and [10, 13, 10]. For Type I error conditions, mean configurations were [2, 2, 2] and [10, 10, 10].

For the Box-Cox data, the power parameter of λ varied (λ: 0, 0.2, 0.4, or 0.6). In the transformed scale, we compared an equal variances condition ($\sigma^2 = 0.10$, for all groups) to unequal variances ($\sigma_1^2 = \sigma_3^2 = 0.10$ and $\sigma_2^2 = 0.20$).

For the gamma-distributed data, the parameter α varied ($\alpha = 0.5, 1, 1.5, 2$, or 2.5). Increasing α corresponds to decreasing skewness. The scale s parameter was held constant. When α is less or equal to 1, the distribution is monotonically declining; for αs above 1, the distribution is bell-shaped and positively skewed.

We used equal sample sizes ($n = 40$ per group) and two sets of unequal sample sizes ($n_1 = 30$, $n_2 = 60$, $n_3 = 30$ and $n_1 = 50$, $n_2 = 20$, $n_3 = 50$). Because the higher location in group two corresponds to greater variance, the first set of unequal sample sizes is a positively paired unequal sample size condition (i.e., higher variance is weighted by a larger group sample size), while the second set is a negatively paired unequal sample size condition.

3.2. Evaluation

Each estimator was evaluated on 5000 replicates.The gamma with log link (GammaLog), the Gaussian with log link (GaussianLog), log transformation on the raw response (LogDV), and the Yuen-Welch test (YW) were applied to each of the replicates. The models (i.e., GammaLog, GaussianLog, and LogDV) treated Group 1 as the referent group. For example, in raw scale via the inverse link function, the expected response given by the GzLMs with log link is:

$$E(Y) = \exp(\beta_0 + \beta_1(x_1 = Group\,2) + \beta_2(x_2 = Group\,3)). \tag{9}$$

The models' deviance residuals were evaluated by two diagnostic tests, the Shapiro-Wilk (1965) and the Brown-Forsythe (1974), for descriptive purposes; these tests did not inform method selection, as all estimators were each evaluated on 5000 replicates. Power rates were reported unconditionally (significant F-test) and conditionally (significant on F-test and nonsignificant on diagnostic tests). A conditional power rate is therefore the probability of detecting a true effect when there is no evidence of assumption violations. The significance level was .05, and empirical Type I error rates were considered acceptable with liberal bounds of .025 and .075 (Bradley 1978). Type I error rates were also reported unconditionally (false positive without diagnostic tests) and conditionally (false positive and nonsignificant diagnostic tests).

4. Results

Tables 1 and 2 show descriptive statistics for the response in original scale. Standard deviations, skewness, and kurtosis values increased with decreases in α for gamma-distributed data and with decreases in λ for the Box-Cox data.

Table 1. Descriptive statistics for gamma-distributed data.

		SD			Skew			Kurtosis		
Locations	α	G1	G2	G3	G1	G2	G3	G1	G2	G3
Low	0.5	2.64	3.92	2.63	1.97	2.00	1.98	6.65	6.77	6.74
	1	1.91	2.89	1.91	1.48	1.46	1.48	4.86	4.80	4.85
	1.5	1.59	2.37	1.58	1.24	1.22	1.22	4.19	4.15	4.14
	2	1.38	2.06	1.38	1.07	1.09	1.09	3.79	3.80	3.81
	2.5	1.23	1.85	1.23	0.96	0.97	0.98	3.56	3.56	3.59
High	0.5	13.20	14.43	13.15	1.98	1.98	1.98	6.67	6.69	6.66
	1	9.61	10.52	9.58	1.47	1.48	1.48	4.86	4.87	4.89
	1.5	7.93	8.65	7.89	1.22	1.21	1.23	4.11	4.09	4.18
	2	6.86	7.58	6.86	1.08	1.09	1.07	3.77	3.81	3.78
	2.5	6.18	6.77	6.17	0.97	0.96	0.98	3.54	3.54	3.59

Note. G1-G3 refers to the three groups. α = shape parameter of the gamma distribution. *SD* = standard deviation. For brevity, descriptives are shown for the effect size of 1 and for equal sample sizes.

Table 2. Descriptive statistics for Box-Cox data.

			SD			Skew			Kurtosis		
Locations	Variances	λ	G1	G2	G3	G1	G2	G3	G1	G2	G3
Low	Equal	0	0.20	0.30	0.20	0.24	0.24	0.24	2.76	2.77	2.80
		0.2	0.17	0.24	0.17	0.16	0.16	0.16	2.74	2.75	2.74
		0.4	0.15	0.19	0.15	0.10	0.09	0.10	2.73	2.73	2.74
		0.6	0.13	0.15	0.13	0.06	0.04	0.06	2.73	2.72	2.75
	Unequal	0	0.20	0.61	0.20	0.23	0.47	0.24	2.77	2.93	2.78
		0.2	0.17	0.48	0.17	0.17	0.30	0.17	2.75	2.79	2.74
		0.4	0.15	0.38	0.15	0.11	0.19	0.11	2.74	2.74	2.73
		0.6	0.13	0.31	0.13	0.05	0.10	0.06	2.74	2.72	2.73
High	Equal	0	0.99	1.10	1.00	0.24	0.24	0.25	2.78	2.78	2.79
		0.2	0.63	0.68	0.63	0.13	0.10	0.12	2.74	2.74	2.73
		0.4	0.39	0.42	0.39	0.05	0.05	0.07	2.73	2.74	2.72
		0.6	0.25	0.26	0.25	0.03	0.03	0.02	2.73	2.73	2.73
	Unequal	0	1.00	2.23	1.00	0.24	0.47	0.23	2.76	2.92	2.78
		0.2	0.63	1.36	0.63	0.12	0.24	0.12	2.75	2.78	2.74
		0.4	0.39	0.84	0.40	0.05	0.11	0.05	2.73	2.74	2.73
		0.6	0.25	0.52	0.25	0.01	0.05	0.01	2.74	2.73	2.73

Note. G1-G3 refers to the three groups. λ = power parameter. *SD* = standard deviation. The (un)equal variances are in transformed scale. For brevity, descriptives are shown for the effect size of 1 and for equal sample sizes.

Power and Type I error trends were similar across sample size conditions, except for Type I errors on the Box-Cox data. The GammaLog and GaussianLog models yielded unbiased estimates of population means. The trimmed means and unadjusted, back-transformed values were systematically lower than the population arithmetic mean. The discrepancy from the arithmetic mean for the YW was generally less than that of LogDV.

4.1. Gamma-distributed data

4.1.1. Type I error and power

All unconditional (without diagnostics) Type I error rates were nearly nominal. Conditional (with diagnostics) Type I error rates were overly conservative for the LogDV and GaussianLog but still nearly nominal for the GammaLog (Table 3).

The GammaLog had more power than the LogDV. At low locations, the GammaLog advantage ranged 24 to 89 percentage points (pp); at high locations, the advantage ranged 7 to 31

Table 3. Type I error rates for gamma-distributed data.

		Without diagnostics				With diagnostics					
Locations	Shapes	YW	GammaLog	LogDV	GaussianLog	GammaLog	Pass	LogDV	Pass	GaussianLog	Pass
Low (2, 2, 2)	0.5	.04	.06	.05	.04	.06	4711	**.00**	5	**.00**	0
	1	.05	.06	.05	.05	.05	4663	**.00**	146	**.00**	0
	1.5	.05	.06	.05	.05	.05	4626	**.01**	568	**.00**	0
	2	.05	.05	.05	.05	.05	4602	**.01**	982	**.00**	1
	2.5	.05	.06	.06	.05	.05	4611	**.02**	1493	**.00**	1
High (10, 10, 10)	0.5	.05	.06	.05	.04	.06	4694	**.00**	4	**.00**	0
	1	.05	.06	.05	.05	.05	4649	**.00**	166	**.00**	0
	1.5	.05	.05	.04	.04	.05	4637	**.01**	576	**.00**	0
	2	.05	.05	.05	.05	.05	4597	**.01**	1041	**.00**	0
	2.5	.06	.06	.05	.05	.05	4592	**.02**	1469	**.00**	4

Note. *YW* = Yuen-Welch test. *GammaLog*= gamma model with log link. *LogDV* = log transformed model. *GaussianLog*= Gaussian model with log link. Unconditional rates are obtained without diagnostics tests. Conditional rates are obtained with nonsignificant results on diagnostic tests. 'Pass' columns indicate the number of replicates (out of 5000) that passed both diagnostics tests. Trends were similar for the unequal sample sizes.

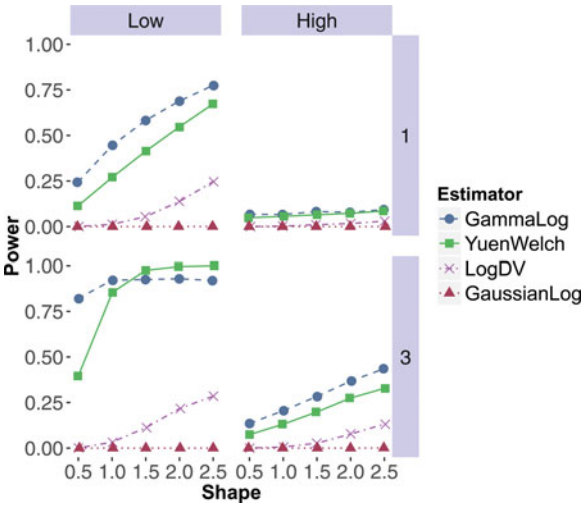

Figure 1. Power rates accounting for diagnostic tests, for gamma-distributed data. 'Low' and 'High' refers to location conditions. *GammaLog* = gamma model with log link. *LogDV* = log transformed model. *GaussianLog* = Gaussian model with log link. *Shape* refers to the gamma distribution parameter (α). '1' and '3' refers to effect size conditions. The unequal sample size conditions are excluded because trends were similar.

pp (Figure 1). Across conditions, median decrease in power due to diagnostics was 3 pp for the GammaLog, 22 pp for LogDV, and 38 pp for GaussianLog. Table 4 shows that the rates at which both diagnostic tests were simultaneously passed by the GammaLog and LogDV did not differ between the effect sizes. Without diagnostics, however, the models generally had advantage over the YW, especially the GammaLog at lower αs.

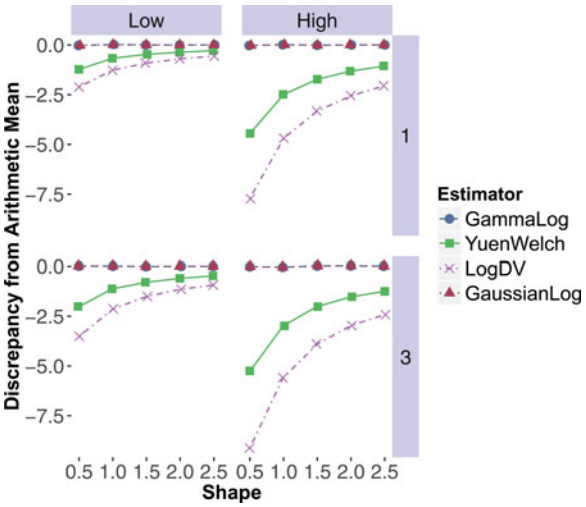

Figure 2. Discrepancy from the arithmetic mean, for gamma-distributed data. *GammaLog* = gamma model with log link. *LogDV* = log transformed model. *GaussianLog* = Gaussian model with log link. *Shape* refers to the gamma distribution parameter (α). 'Low' and 'High' refers to location conditions. '1' and '3' refers to effect size conditions. The geometric and trimmed means are more discrepant from the arithmetic mean for lower shapes, higher locations, and larger effect sizes. For brevity, only discrepancies for Group 2 are shown.

4.1.2. Discrepancy from the arithmetic mean

Discrepancies for the LogDV and YW were stronger with lower α and with higher locations (Figure 2). Across conditions, the LogDV discrepancy ranged from −0.56 to −9.13, and the trimmed mean discrepancy ranged from −0.29 to −5.25.

4.2. Box-Cox data

4.2.1. Type I error and power

Type I error rates (equal variances in transformed scale)

The models had adequate unconditional and conditional Type I error rates across all sample size conditions.

Type I error rates (unequal variances in transformed scale)

Without diagnostics, the model-based estimators were slightly liberal in equal sample size conditions (Table 5), slightly conservative in the positively paired sample sizes condition (Table 6), and overly liberal in the negatively paired sample sizes condition (Table 7). With diagnostics, conditional Type I error rates were near zero across all sample size conditions.

Conditional power (equal variances in transformed scale)

Tables 8 and 9 show diagnostic test behaviour for effect sizes 1 and 3, respectively. With the lower effect size, the LogDV had a trivial advantage over the GammaLog at $\lambda = 0$. The GaussianLog increased in power with increased λ and surpassed the GammaLog and LogDV above $\lambda \sim 0.50$; this corresponded with the increasing rates for the diagnostic tests, as well as distributions that were increasingly normalized. As λ increased, power for the LogDV and GammaLog decreased; the extent of these decreases was more pronounced with the larger effect size and can be explained by the failure to pass the Brown-Forsythe test.

Conditional power (unequal variances in transformed scale)

With the larger effect size, however, power for LogDV and GammaLog increased . At $\lambda = 0$, for which the log transformation would have been most suitable, power was essentially nil when variances were unequal in transformed space (Figure 3), which corresponded to the low rates at which the Brown-Forsythe test was passed. Similarly, even with more normal distributions at higher λ, the GaussianLog almost always failed the Brown-Forsythe.

Unconditional power

Without diagnostics, unconditional power for the GammaLog was superior to the YW on all conditions. Across the four conditions where YW's power was below 80% (unequal variances, high location, λ near zero), the GammaLog and LogDV had 12 to 33 pp more power than the YW. The reduction in power due to diagnostic tests was greater in unequal variance conditions; while median reduction in power was 11 pp for both GammaLog and LogDV and 29 pp for GaussianLog in equal variance conditions, median reduction was 91 pp for both GammaLog and LogDV and 100 pp for GaussianLog in unequal variance conditions.

4.2.2. Discrepancy from the arithmetic mean

Discrepancy was consistently near zero for the GammaLog and GaussianLog but increased slightly for LogDV and YW with higher skewness (i.e., lower λ), higher locations, and

Table 4. Rates of passing diagnostics tests for gamma-distributed data, in power conditions.

Effect size 1

		Pass Shapiro-Wilk only			Pass Brown-Forsythe only			Pass both		
Locations	Shapes	GammaLog	LogDV	GaussianLog	GammaLog	LogDV	GaussianLog	GammaLog	LogDV	GaussianLog
Low (2, 3, 2)	0.5	.95	.00	.00	.99	.96	.79	.94	.00	.00
	1	.96	.04	.00	.98	.96	.64	.94	.03	.00
	1.5	.96	.12	.00	.97	.96	.57	.93	.11	.00
	2	.95	.23	.00	.97	.95	.53	.92	.21	.00
	2.5	.95	.33	.01	.97	.96	.50	.92	.31	.00
High (10, 11, 10)	0.5	.95	.00	.00	.99	.95	.95	.94	.00	.00
	1	.96	.03	.00	.97	.95	.95	.93	.03	.00
	1.5	.96	.11	.00	.97	.95	.93	.92	.10	.00
	2	.96	.22	.00	.97	.96	.94	.92	.20	.00
	2.5	.95	.32	.00	.97	.96	.94	.92	.30	.00

Effect size 3

		Pass Shapiro-Wilk only			Pass Brown-Forsythe only			Pass both		
Locations	Shapes	GammaLog	LogDV	GaussianLog	GammaLog	LogDV	GaussianLog	GammaLog	LogDV	GaussianLog
Low (2, 5, 2)	0.5	.95	.00	.00	.99	.95	.18	.94	.00	.00
	1	.96	.04	.00	.97	.95	.04	.93	.03	.00
	1.5	.96	.12	.00	.97	.96	.01	.93	.11	.00
	2	.96	.23	.00	.97	.96	.01	.93	.22	.00
	2.5	.95	.31	.01	.96	.95	.01	.92	.29	.00
High (10, 13, 10)	0.5	.95	.00	.00	.99	.95	.89	.94	.00	.00
	1	.96	.04	.00	.98	.96	.84	.93	.03	.00
	1.5	.95	.12	.00	.97	.96	.80	.92	.11	.00
	2	.95	.22	.00	.97	.96	.78	.93	.21	.00
	2.5	.95	.33	.00	.96	.95	.77	.92	.31	.00

Note. *GammaLog* = gamma model with log link. *LogDV* = log transformed model. *GaussianLog* = Gaussian model with log link. The third set of columns, 'Pass both', indicates the proportion of replicates that passed both diagnostics tests. Trends were similar for the unequal sample sizes.

Table 5. Type I error rates for Box-Cox data, with equal sample sizes.

			Without diagnostics				With diagnostics					
Locations	Variances	λ	YW	GammaLog	LogDV	GaussianLog	GammaLog	Pass	LogDV	Pass	GaussianLog	Pass
Low (2, 2, 2)	Equal	0	.04	.05	.05	.05	.04	4669	.04	4730	.04	4042
		0.2	.05	.05	.05	.05	.05	4756	.05	4730	.04	4408
		0.4	.05	.05	.05	.05	.05	4743	.04	4673	.05	4594
		0.6	.05	.05	.05	.05	.05	4714	.04	4608	.04	4684
	Unequal	0	.05	**.08**	.06	**.08**	**.00**	2546	**.00**	2563	**.00**	1619
		0.2	.05	**.08**	.06	.07	**.00**	2593	**.00**	2503	**.00**	2087
		0.4	.05	.06	.06	.06	**.00**	2548	**.00**	2430	**.00**	2367
		0.6	.05	.07	.07	.07	**.00**	2423	**.00**	2273	**.00**	2471
High (10, 10, 10)	Equal	0	.05	.05	.05	.05	.05	4662	.05	4755	.04	3977
		0.2	.05	.05	.05	.05	.04	4723	.04	4735	.04	4579
		0.4	.05	.05	.05	.06	.05	4762	.05	4744	.05	4723
		0.6	.05	.06	.05	.06	.05	4726	.05	4719	.05	4729
	Unequal	0	.05	**.08**	.06	**.08**	**.00**	2529	**.00**	2587	**.00**	1660
		0.2	.05	.06	.06	.06	**.00**	2591	**.00**	2549	**.00**	2295
		0.4	.05	.06	.06	.06	**.00**	2559	**.00**	2523	**.00**	2522
		0.6	.05	.06	.06	.06	**.00**	2526	**.00**	2501	**.00**	2554

Note. *YW* = Yuen-Welch test. *GammaLog* = gamma model with log link. *LogDV* = log transformed model. *GaussianLog* = Gaussian model with log link. λ = power parameter. Unconditional rates are obtained without diagnostics tests. Conditional rates are obtained with nonsignificant results on diagnostic tests. 'Pass' columns indicate the number of replicates (out of 5000) that passed both diagnostics tests.

higher effect sizes. Discrepancies were also slightly more pronounced with unequal variances (Figure 4), though magnitudes were not as large as those in the gamma setting.

5. Discussion

We compared the GzLM, the log transformation, and the Yuen-Welch test with gamma-distributed and Box-Cox data for their abilities to detect group mean differences when data exhibit non-normality and heteroscedasticity and for their discrepancies from the arithmetic mean.

Table 6. Type I error rates for Box-Cox data, with positively paired unequal sample sizes.

			Without diagnostics				With diagnostics					
Locations	Variances	λ	YW	GammaLog	LogDV	GaussianLog	GammaLog	Pass	LogDV	Pass	GaussianLog	Pass
Low (2, 2, 2)	Equal	0	.05	.05	.05	.05	.05	4512	.05	4562	.04	4053
		0.2	.05	.05	.05	.05	.05	4595	.05	4582	.04	4357
		0.4	.05	.05	.05	.05	.04	4551	.04	4518	.04	4439
		0.6	.06	.05	.05	.05	.05	4566	.05	4515	.05	4542
	Unequal	0	.05	.03	**.02**	**.02**	**.00**	247	**.00**	249	**.00**	190
		0.2	.06	.03	**.02**	.03	**.00**	282	**.00**	270	**.00**	245
		0.4	.05	**.02**	**.02**	**.02**	**.00**	287	**.00**	269	**.00**	274
		0.6	.05	**.02**	**.02**	**.02**	**.00**	261	**.00**	254	**.00**	253
High (10, 10, 10)	Equal	0	.05	.05	.05	.05	.05	4529	.04	4555	.04	4106
		0.2	.06	.05	.05	.05	.05	4543	.05	4547	.05	4430
		0.4	.06	.06	.06	.06	.05	4590	.05	4584	.05	4584
		0.6	.05	.05	.05	.05	.05	4589	.05	4572	.05	4592
	Unequal	0	.06	.03	**.02**	.03	**.00**	261	**.00**	265	**.00**	198
		0.2	.05	.03	**.02**	**.02**	**.00**	253	**.00**	243	**.00**	246
		0.4	.05	**.02**	**.02**	**.02**	**.00**	246	**.00**	241	**.00**	239
		0.6	.05	**.02**	**.02**	**.02**	**.00**	252	**.00**	252	**.00**	252

Note. *YW* = Yuen-Welch test. *GammaLog* = gamma model with log link. *Log DV* = log transformed model. *GaussianLog* = Gaussian model with log link. λ = power parameter. Unconditional rates are obtained without diagnostics tests. Conditional rates are obtained with nonsignificant results on diagnostic tests. 'Pass' columns indicate the number of replicates (out of 5000) that passed both diagnostics tests.

Table 7. Type I error rates for Box-Cox data, with negatively paired unequal sample sizes.

			Without diagnostics				With diagnostics					
Locations	Variances	λ	YW	GammaLog	LogDV	GaussianLog	GammaLog	Pass	LogDV	Pass	GaussianLog	Pass
Low (2, 2, 2)	Equal	0	.05	.05	.05	.05	.05	917	.05	913	.04	843
		0.2	.05	.04	.04	.04	.04	909	.04	903	.03	890
		0.4	.04	.05	.05	.04	.04	916	.04	893	.04	915
		0.6	.05	.04	.04	.04	.04	906	.04	915	.04	769
	Unequal	0	.06	**.16**	**.15**	**.16**	**.01**	76	**.01**	72	**.01**	67
		0.2	.05	**.16**	**.14**	**.16**	**.01**	75	**.01**	76	**.01**	79
		0.4	**.08**	**.17**	**.17**	**.17**	**.01**	74	**.01**	73	**.02**	77
		0.6	.06	**.15**	**.16**	**.15**	**.02**	100	**.02**	95	**.02**	100
High (10, 10, 10)	Equal	0	.06	.05	.05	.05	.05	893	.05	910	.04	778
		0.2	.06	.05	.05	.05	.05	920	.05	919	.04	885
		0.4	.06	.05	.05	.05	.05	926	.05	926	.05	915
		0.6	.05	.05	.05	.05	.04	912	.04	911	.04	916
	Unequal	0	.06	**.17**	**.14**	**.16**	**.02**	87	**.02**	86	**.01**	66
		0.2	.05	**.14**	**.13**	**.14**	**.01**	97	**.01**	92	**.01**	92
		0.4	.06	**.12**	**.12**	**.13**	**.01**	93	**.01**	93	**.01**	85
		0.6	.04	**.13**	**.13**	**.13**	**.02**	83	**.02**	83	**.01**	77

Note. *YW* = Yuen-Welch test. *GammaLog* = gamma model with log link. *LogDV* = log transformed model. *GaussianLog* = Gaussian model with log link. λ = power parameter. Unconditional rates are obtained without diagnostics tests. Conditional rates are obtained with nonsignificant results on diagnostic tests. 'Pass' columns indicate the number of replicates (out of 5000) that passed both diagnostics tests.

For positive skewness and variances that are proportional to means, two estimators that retain all relevant data are the LogDV model and the gamma model. For the Box-Cox data, both had similar rates for power and diagnostic tests, but the LogDV did not show clear superiority in power compared to the GammaLog. For the gamma-distributed data, the LogDV tended to have significant departures from normality. Overall, although both models generally follow the same trends, there is a greater power disadvantage when applying the LogDV to gamma-distributed data than there is when applying the GammaLog to the Box-Cox data. Further, when the GammaLog Even when $\lambda = 0$ for the tested conditions, for which the log raw data transformation would be ideal, the LogDV advantage was only slight when error variances were equal in transformed scale. For situations in which the natural logarithmic transformation is viable, the gamma GzLM may also be considered. Certainly, this depends on the data situation; previous simulations have shown that the gamma model suffers from inefficiency, more so than the log-transformed model, when there is kurtosis in transformed scale (Manning, Mullahy, and Manning 2001).

For both these models, the state of error variances in transformed space is important to note, along with the mean-variance relation concept. These were demonstrated by the condition of unequal variances in transformed space for the Box-Cox data. Though results were stated as relations to increasing and decreasing λ values, it is not exactly the λ parameter that matters. Consider $\lambda = 0$. When variances were equal, models with either the log-transformed response (LogDV) or the gamma family (GammaLog) worked well because linearity could be achieved. When variances were unequal, both failed. Transformations do not always guarantee linearization, and GzLMs do not always fit if the mean-variance relation is not approximated. It is not simply the presence or magnitude of skewness and heteroscedasticity that determines whether one should use the GzLM but rather the presence of some mean-variance relation.

Our simulation used null hypothesis tests as proxies for the inspection of model residuals, leading to the contrast between unconditional and conditional error rates. The reduction in power due to assumptions tests was drastic when assumptions were not approximated well,

Table 8. Rates of passing diagnostics tests for Box-Cox data, in power condition (effect size 1).

			Pass Shapiro-Wilk only			Pass Brown-Forsythe only			Pass both		
Locations	Variances	λ	GammaLog	LogDV	GaussianLog	GammaLog	LogDV	GaussianLog	GammaLog	LogDV	GaussianLog
Low (2, 3, 2)	Equal	0	.93	.95	.72	.95	.95	.35	.89	.90	.26
		0.2	.95	.94	.85	.94	.94	.54	.89	.89	.46
		0.4	.95	.94	.91	.87	.88	.73	.83	.82	.66
		0.6	.93	.92	.94	.76	.76	.87	.71	.70	.82
	Unequal	0	.48	.48	.03	.02	.02	.00	.01	.01	.00
		0.2	.63	.61	.06	.05	.05	.00	.03	.03	.00
		0.4	.74	.71	.14	.12	.12	.00	.10	.10	.00
		0.6	.83	.79	.25	.25	.25	.00	.21	.20	.00
High (10, 11, 10)	Equal	0	.93	.95	.80	.95	.95	.93	.89	.91	.74
		0.2	.95	.95	.92	.96	.96	.94	.91	.91	.86
		0.4	.95	.94	.94	.96	.96	.95	.91	.90	.90
		0.6	.95	.95	.95	.95	.95	.95	.90	.90	.90
	Unequal	0	.50	.50	.23	.02	.02	.00	.01	.01	.00
		0.2	.54	.53	.34	.02	.02	.01	.02	.02	.00
		0.4	.58	.57	.41	.03	.03	.01	.02	.02	.01
		0.6	.60	.60	.45	.04	.04	.01	.03	.03	.01

Note. *GammaLog* = gamma model with log link. *LogDV* = log transformed model. *GaussianLog* = Gaussian model with log link. λ = power parameter. The third set of columns, 'Pass both', indicates the proportion of replicates that passed both diagnostics tests. Trends were similar for the unequal sample sizes.

Table 9. Rates of passing diagnostics tests for Box-Cox data, in power condition (effect size 3).

			Pass Shapiro-Wilk only			Pass Brown-Forsythe only			Pass both		
Locations	Variances	λ	GammaLog	LogDV	GaussianLog	GammaLog	LogDV	GaussianLog	GammaLog	LogDV	GaussianLog
Low (2, 5, 2)	Equal	0	.94	.95	.14	.96	.96	.00	.90	.91	.00
		0.2	.95	.95	.42	.85	.85	.01	.81	.80	.00
		0.4	.92	.91	.71	.50	.50	.10	.45	.45	.08
		0.6	.86	.85	.87	.13	.13	.43	.11	.11	.39
	Unequal	0	.48	.48	.00	.02	.02	.00	.01	.01	.00
		0.2	.77	.75	.00	.14	.14	.00	.12	.11	.00
		0.4	.90	.89	.00	.53	.53	.00	.48	.48	.00
		0.6	.94	.92	.06	.89	.89	.00	.83	.81	.00
High (10, 13, 10)	Equal	0	.94	.95	.78	.96	.96	.70	.90	.91	.54
		0.2	.95	.95	.91	.94	.94	.80	.90	.90	.73
		0.4	.95	.95	.94	.92	.92	.87	.88	.88	.82
		0.6	.95	.94	.95	.89	.89	.92	.84	.84	.87
	Unequal	0	.48	.48	.08	.01	.01	.00	.01	.01	.00
		0.2	.59	.58	.17	.03	.03	.00	.02	.02	.00
		0.4	.68	.68	.26	.07	.07	.00	.06	.06	.00
		0.6	.73	.73	.33	.11	.11	.00	.09	.09	.00

Note. *GammaLog* = gamma model with log link. *LogDV* = log transformed model. *GaussianLog* = Gaussian model with log link. λ = power parameter. The third set of columns, 'Pass both', indicates the proportion of replicates that passed both diagnostics tests. The unequal sample size conditions are excluded because trends were similar.

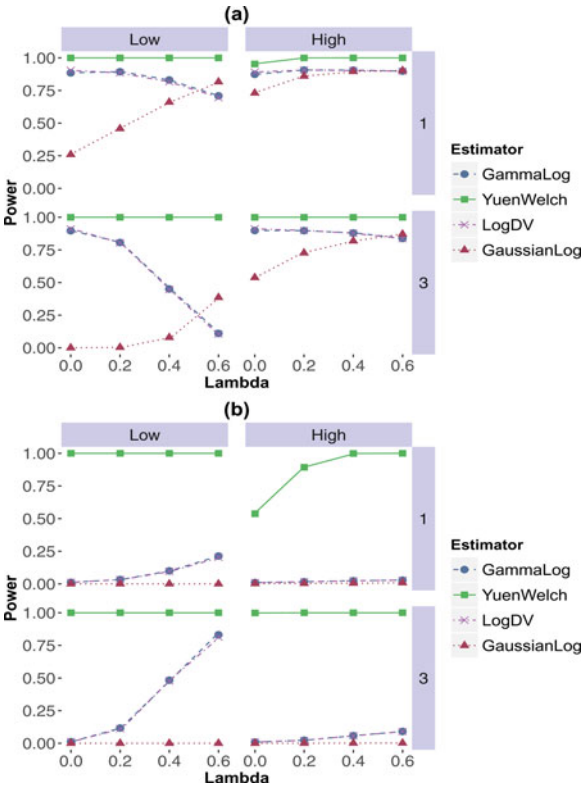

Figure 3. Power accounting for diagnostic tests, for Box-Cox data. *GammaLog* = gamma model with log link. *LogDV* = log transformed model. *GaussianLog* = Gaussian model with log link. *Lambda* is the power parameter (λ). Panel (a) shows the equal variances condition while panel (b) shows the unequal variances condition. 'Low' and 'High' refers to location conditions. '1' and '3' refers to effect size conditions. The unequal sample size conditions are excluded because trends were similar.

and they were smaller when assumptions could be somewhat met (e.g., the GammaLog' s rates were reduced by 3 pp under gamma-generated data and by 11 pp in some Box-Cox data). Still, a logical limitation of interpreting conditional error rates on their own is the assumption of tests' results corresponding to researchers' judgments on the severity of violations. As sample size increases, these tests have increased power for detecting even minute departures, though researchers may consider such departures less consequential. Even if formal tests are used in practice, stringency could be imposed with a more logically-aligned equivalence test, such as one for detecting evidence of homoscedasticity (Kim and Cribbie 2017; Mara and Cribbie 2017) or with conservative alpha levels (Schucany and Ng 2006). In practice, we recommend neither reliance on assumptions tests nor conditional approaches for method selection. We do recommend researchers to consider alternative frameworks and to do due diligence by inspecting assumptions, be it graphically or visually, for critically evaluating conclusions.

The YW with 20% trimmed means is an alternative that requires no assumption evaluation and that has adequate properties across many distributions and sample size configurations. When lower sample sizes are paired with higher variance and when errors are unequal in transformed space, the YW maintains (unconditional) Type I error rates where parametric alternatives are not suitable. Its inference to the trimmed population does, however, reduce generalizability. When distributions are approximately symmetrical, as with higher λ and α

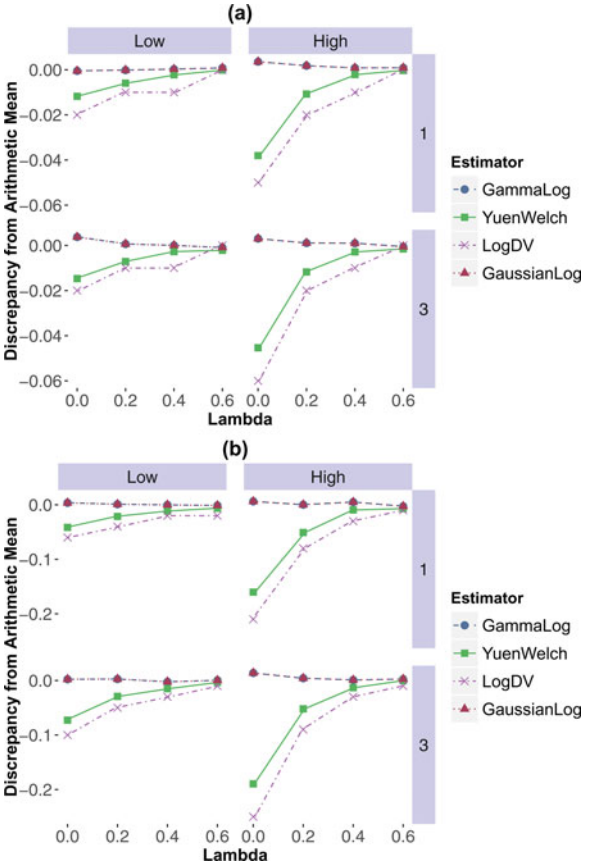

Figure 4. Discrepancy from the arithmetic mean, for Box Cox data. *GammaLog* = gamma model with log link. *LogDV* = log transformed model. *GaussianLog* = Gaussian model with log link. *Lambda* is the power parameter (λ). Panel (a) shows the equal variances condition while panel (b) show the unequal variances condition. Note the differences in scale for the panels' y-axies. 'Low' and 'High' refers to location conditions. '1' and '3' refers to effect size conditions. Geometric and trimmed means are more discrepant from the arithmetic mean with lower power parameters and the higher location. For brevity, only discrepancies for Group 2 are shown.

values, the trimmed mean is only slightly discrepant from the arithmetic mean. When population distributions are more asymmetric, the 20% trimmed mean is more discrepant; the inference to the bulk of the population has no implied representation for potentially meaningful distribution tails, thereby lowering generalizability. The YW is a valuable alternative, but researchers should be cognizant of generalizability and that there may be situations where it is statistically and substantively worthwhile to determine whether a parametric option is viable.

6. Final remarks

Overall, the choice of alternative frameworks involves substantive and data considerations asides from power. The growing awareness and application of robust statistics is greatly beneficial, but one should not simply default to robust approaches just because nonnormality and heteroscedasticity are present. Similarly, just because skewness and heteroscedasticity can be modeled using the gamma GzLM (or a model with log transformation) does not

mean that the assumed nature of heteroscedasticity is actually tenable. It is not necessarily the case that the psychological data of interest are truly generated from gamma processes (just as data are likely not truly generated from Gaussian processes), and therefore, it may also not be the case that a gamma model would be ideal across as many scenarios as the YW. The log model and gamma GzLM have similar trends; with the popular use of natural log transformations in psychology, there may be many scenarios in which the gamma GzLM could serve well. If a GzLM is tenable, then one obtains unbiased inferences about the population mean, estimator sufficiency, and potentially, higher power for detecting effects. We encourage researchers who desire not only power but also interpretation about population means to consider and evaluate the tenability of parametric options before or alongside robust alternatives.

Funding

Social Sciences and Humanities Research Council of Canada.

References

Blough, D. K., C. W. Madden, and M. C. Hornbrook. 1999. Modeling risk using generalized linear models. *Journal of Health Economics* 18 (2):153–71.

Bonett, D. G., and R. M. Price. 2002. Statistical inference for a linear function of medians: Confidence intervals, hypothesis testing, and sample size requirements. *Psychological Methods* 7 (3): 370–83.

Box, G., and D. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)* 26 (2):211–252.

Bradley, J. V. 1978. Robustness? *British Journal of Mathematical and Statistical Psychology* 31 (2):144–52.

Brown, M. B., and A. B. Forsythe. 1974. Robust tests for the equality of variances. *Journal of the American Statistical Association* 69 (346):364–7.

Chen, E. H. 1971. The power of the Shapiro-Wilk W test for normality in samples from contaminated normal distributions. *Journal of the American Statistical Association* 66 (336):760–2.

Duan, N. 1983. Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association* 78 (383):605–10.

Erceg-Hurn, D., and V. Mirosevich. 2008. Modern Robust Statistical Methods. *American Psychologist* 63 (7):591–601.

Fox, J. 2008. *Applied Regression Analysis and Generalized Linear Models*. Applied regression analysis and generalized linear models (2nd ed.). Thousand Oaks, CA: Sage Publications.

García-Pérez, M. A. 2012. Statistical conclusion validity: Some common threats and simple remedies. *Frontiers in Psychology* 3:325. Available at: http://doi.org/10.3389/fpsyg.2012.00325

Gastwirth, J. L., Y. R. Gel, and W. Miao. 2009. The impact of Levene's test of equality of variances on statistical theory and practice. *Statistical Science* 24 (3):343–360.

Golinski, C., and R. Cribbie. 2009. The expanding role of quantitative methodologists in advancing psychology. *Canadian Psychology/Psychologie canadienne* 50 (2):83–90.

Grissom, R. J. 2000. Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology* 68 (1):155–65.

Hayes, A. F., and L. Cai. 2007. Further evaluating the conditional decision rule for comparing two independent means. *British Journal of Mathematical and Statistical Psychology* 60 (2):217–44.

Jones, A. 2012. *The Elgar Companion to Health Economics*. (2nd ed.). Edward Elgar.

Keselman, H., R. Wilcox, A. R. Othman, and K. Fradette. 2002. Trimming, transforming statistics, and bootstrapping: Circumventing the biasing effects of heterescedasticity and nonnormality. *Journal of Modern Applied Statistical Methods* 1 (2):288–309.

Kilian, R., H. Matschinger, W. Löeffler, C. Roick, and M. C. Angermeyer. 2002. A comparison of methods to handle skew distributed cost variables in the analysis of the resource consumption in schizophrenia treatment. *The Journal of Mental Health Policy and Economics* 5 (1):21–31.

Kim, Y. J., and R. A. Cribbie. 2017. ANOVA and the variance homogeneity assumption: Exploring a better gatekeeper. *British Journal of Mathematical and Statistical Psychology* 71 (1):1–12.
Levene, H. 1960. Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to Probability and Statistics*, pp. 278–292. Palo Alto, California: Stanford University Press.
Manning, W. 1998. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics* 17 (3):283–95.
Manning, W. G., and J. Mullahy. 2001. Estimating log models: to transform or not to transform? *Journal of Health Economics* 20 (4):461–94.
Mara, C. A., and R. A. Cribbie. 2017. Equivalence of population variances: Synchronizing the objective and analysis. *The Journal of Experimental Education* 1–16.
Micceri, T. 1989. The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin* 105 (1):156–66.
Neal, D. J., and J. S. Simons. 2007. Inference in regression models of heavily skewed alcohol use data: a comparison of ordinary least squares, generalized linear models, and bootstrap resampling. *Psychology of Addictive Behaviors* 21 (4):441–52.
Nelder, J. A., and R. W. M. Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* 135 (3):370–84.
Nevill, A. M., and J. B. Copas. 1991. Using generalized linear models (GLMs) to model errors in motor performance. *Journal of Motor Behavior* 23 (4):241–50.
R Development Core Team. 2015. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
Ratcliff, R. 1993. Methods for dealing with reaction time outliers.*Psychological Bulletin* 114 (3):510–32.
Razali, N., and Y. Wah. 2011. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics* 2 (1):21–33.
Schucany, W. R., and T. Ng. 2006. Preliminary goodness-of-fit tests for normality do not validate the one-sample Student t. *Communications in Statistics - Theory and Methods* 35 (12):2275–86.
Shapiro, S., and M. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52 (3-4):591–611.
Tukey, J. W. 1960. A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, ed. I. Olkin, S. Ghurye, W. Hoeffding, W. Madow, and H. Mann, Redwood City, CA: Stanford University Press. vol. 2, pp. 448–85.
Wilcox, R. 2005. *Introduction to Robust Estimation and Hypothesis Testing*. (2nd ed.). Academic Press.
Yuen, K. K. 1974. The two-sample trimmed t for unequal population variances. *Biometrika* 61 (1):165–70.
Zimmerman, D. W. 2004. A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology* 57 (1):173–81.