HUDM 5123 - Linear Models and Experimental Design
13 - Robust Methods for One-Way Designs and Contrast Comparisons

1 Robust Methods for One-Way Designs

When the assumptions of homogeneity of variance (HoV) and/or normally distributed residuals are violated, the nominal Type I error rate of the one-way ANOVA omnibus test, typically set to $\alpha = .05$, may differ from the actual Type I error rate when the null hypothesis is true. Violations of assumptions also lead to deviations in power when the null hypothesis is false.

1.1 Two Methods Robust to Violation of HoV

Two alternatives involve parametric modifications to the test statistic.

1. **The Brown Forsythe F^*:**

$$F^* = \frac{\sum_{j=1}^{a} n_j \left(\bar{Y}_j - \bar{Y}\right)^2}{\sum_{j=1}^{a} \left[1 - (n_j/N)\right] s_j^2}.$$

   Under the null hypothesis, F^* has an approximate F distribution with $a-1$ numerator and f denominator degrees of freedom where

$$f = \frac{1}{\sum_{j=1}^{a} \frac{g_j^2}{(n_j - 1)}},$$

   where

$$g = \frac{[1 - (n_j/N)]s_j^2}{\sum_{j=1}^{a}[1 - (n_j/N)]s_j^2}.$$

2. **The Welch W:**

$$W = \frac{\sum_{j=1}^{a} w_j \left(\bar{Y}_j - \tilde{Y}\right)^2 \Big/ (a-1)}{[1 + (2/3)(a-2)\Lambda]},$$

where

$$
\begin{aligned}
w_j &= n_j/s_j^2, \\
\tilde{Y} &= \sum_{j=1}^{a} w_j \bar{Y}_j \Bigg/ \sum_{j=1}^{a} w_j \ , \\
\Lambda &= \frac{3\sum_{j=1}^{a}\left\{\left[1-\left(w_j \Big/ \sum_{j=1}^{a} w_j\right)\right]^2 \Big/ (n_j-1)\right\}}{a^2-1}.
\end{aligned}
$$

Under the null hypothesis, W has an approximate F distribution with $a-1$ numerator and $1/\Lambda$ denominator degrees of freedom.

1.2 A Method that Does not Assume Normality

When the assumption of normality is violated, the nonparametric Kruskal-Wallis test is an alternative that may be used to test the same null hypothesis as the omnibus F test, i.e., $H_0 : \mu_1 = \ldots = \mu_a$, so long as the assumption of equal shaped distributions across groups holds. The test statistic, H, is based on the ranked scores of the outcome variable, where ties receive a score equal to the mean of the ranks of the scores involved in the tie. Once the scores have been ranked, the test statistic is calculated as follows.

$$
H = \frac{12}{N(N+1)} \sum_{j=1}^{a} n_j \left\{ \bar{R}_j - \frac{N+1}{2} \right\}^2,
$$

where $\bar{R}_j$ is the mean rank for group j. When the null hypothesis is true and there are no ties in the data, H is approximately χ^2 distributed with $a-1$ degrees of freedom. When ties occur in the data, the H statistic should be divided by a correction factor T so that $H' = H/T$ where

$$
T = 1 - \frac{\sum_{i=1}^{G}(t_i^3 - t_i)}{N^3 - N},
$$

where t_i is the number of observations tied at a particular value and G is the number of distinct values for which there are ties.

2 Tomarken and Serlin Simulation Study

The simulation study by Tomarken and Serlin (1986; *Psychological Bulletin*) examines the Type I error rate and power of ANOVA alternatives when homogeneity of variance is violated.

With three, four, or more groups, there are many ways that HoV could be violated. Another factor that makes a difference is the group sample sizes and whether the group with the smaller sample size has the largest variance, etc. In their study, Tomarken and Serlin simulated 32 scenarios; 16 with three groups and 16 with four groups. In each case (groups = 3 and groups = 4) they simulated four sample size scenarios and five variance ratio scenarios.

When studying Type I error rate, all group means were necessarily the same (cf. the definition of Type I error). When studying power, they simulated four or five different mean structures. For example, ES = equally spaced means; EX = power concentrated in one extreme with other groups having equal population means, etc. Their Table 1 shows the simulation data generation scenarios.

Design of the Monte Carlo Investigation

n	Variances				
	6/6/6	12/4/1	6/2/1	1/4/12	1/2/6
$K = 3$					
20/20/20	A	B	B		
12/12/12	A	B	B		
30/20/10	A	C	C	D	D
18/12/6	A	C	C	D	D
	6/6/6/6	12/6/4/1	6/3/2/1	1/4/6/12	1/2/3/6
$K = 4$					
20/20/20/20	A	B	B		
12/12/12/12	A	B	B		
30/24/16/10	A	C	C	D	D
18/14/10/6	A	C	C	D	D

Note. The order for $K = 3$ is $n_1 n_2 n_3$; the order for $K = 4$ is $n_1 n_2 n_3 n_4$. A = homogenous variance cases. B = equal sample size heterogeneous variance cases. C = unequal sample size heterogeneous variance, direct-pairing cases. D = unequal sample size heterogeneous variance, inverse-pairing cases.

Two values of α, .05 and .01, are used to study Type I error rates. Results of the simulation are displayed in their Table 2. In the note the explain that each rate reported in the table is actually an average of four rates (the letter types from Table 1). The "Violations" is the number of those four that exceeded the nominal alpha level by about 50%.

Table 2
Empirical Type I Error Probabilities

Condition	α = .05					α = .01				
	F	W	BF	KW	NS	*F*	W	BF	KW	NS
				K = 3						
Variance homogeneity										
M[a]	.053	.050	.051	.046	.045	.008	.010	.008	.007	.008
Violations[b]	0	0	0	0	0	0	0	0	0	0
Variance heterogeneity										
Equal *n*										
M	.069	.048	.062	.060	.049	.019	.008	.016	.012	.012
Violations	3	0	0	0	0	4	0	1	0	0
Unequal *n*/direct pairing[c]										
M	.022	.049	.061	.031	.023	.009	.009	.017	.006	.004
Violations	0	0	0	0	0	0	0	2	0	0
Unequal *n*/inverse pairing[d]										
M	.167	.057	.064	.106	.112	.078	.014	.019	.032	.039
Violations	4	0	1	4	4	4	1	2	4	4
				K = 4						
Variance homogeneity										
M	.053	.055	.052	.048	.046	.009	.010	.009	.008	.007
Violations	0	0	0	0	0	0	0	0	0	0
Variance heterogeneity										
Equal *n*										
M	.064	.048	.059	.055	.049	.018	.008	.015	.009	.011
Violations	1	0	0	0	0	2	0	1	0	0
Unequal *n*/direct pairing										
M	.025	.050	.059	.028	.022	.005	.008	.014	.005	.003
Violations	0	0	0	0	0	0	0	1	0	0
Unequal *n*/inverse pairing										
M	.144	.056	.059	.092	.099	.061	.014	.019	.024	.029
Violations	4	0	0	4	4	4	1	3	3	4

Note: F = ANOVA *F*. W = Welch V_W. BF = Brown and Forsythe F^*. KW = Kruskal-Wallis H. NS = inverse normal scores *W*.
[a] All means are averaged over the four cases of each condition as denoted in Table 1. [b] The number of criterion violations is the number of cases of each type (maximum possible = 4) that yielded empirical rejection rates > .07 at nominal α = .05 and empirical rejection rates > .015 at nominal α = .01. [c] Larger sample sizes associated with larger variances. [d] Larger sample sizes associated with smaller variances.

There are two simulations looking at power. The first uses data that were generated such that homogeneity of variance is met; the second uses data that were generated such that homogeneity of variance is violated. As expected, when HoV is met, the ANOVA F is most powerful, though only slightly so.

Table 3

Empirical Power When the Homogeneity of Variance Assumption Was Met (Across-Case Means)

Mean structure	No. of cases	$\alpha = .05$					$\alpha = .01$				
		F	W	BF	KW	NS	F	W	BF	KW	NS
					$K = 3$						
ES[a]	4	.697	.665	.684	.663	.675	.449	.395	.426	.382	.386
EX[b]	2	.712	.693	.709	.686	.696	.462	.424	.454	.398	.403
EX1[c]	2	.700	.662	.674	.680	.683	.439	.368	.416	.396	.398
EX3[d]	2	.683	.639	.664	.630	.659	.441	.395	.408	.341	.364
					$K = 4$						
ES	4	.698	.664	.684	.658	.678	.449	.396	.427	.388	.396
EX	2	.704	.672	.702	.666	.681	.459	.424	.453	.396	.398
EX1	2	.682	.628	.666	.642	.653	.426	.351	.404	.380	.379
EX4[e]	2	.684	.637	.667	.621	.659	.445	.380	.410	.306	.358
2M[f]	4	.699	.660	.685	.660	.669	.445	.389	.423	.378	.389

Note. For all cases, mean structures were specified corresponding to an estimated ANOVA power of .70 at $\alpha = .05$. F = ANOVA F. W = Welch V_W. BF = Brown and Forsythe F^*. KW = Kruskal-Wallis H. NS = inverse normal scores W.

[a] ES = equally spaced means. [b] EX = power concentrated in one extreme group with other groups having equal population means (applicable in equal sample size homogeneous variance cases only). [c] EX1 = EX pattern with μ_1 the extreme mean (applicable in unequal sample size homogeneous variance cases and in all heterogeneous variance cases). [d] EX3 = EX pattern with μ_3 the extreme mean (applicable in $K = 3$ unequal sample size homogeneous variance cases and in all $K = 3$ heterogeneous variance cases; K indicates the number of groups studied). [e] EX4 = EX pattern with μ_4 the extreme mean (applicable in $K = 4$ unequal sample size homogeneous variance cases and in all $K = 4$ heterogeneous variance cases). [f] 2M = "two in the middle" pattern, with $\mu_1 > \mu_2 = \mu_3 > \mu_4$ (applicable in all $K = 4$ cases).

When homogeneity of variance is violated there are large differences in performance.

Empirical Power When the Homogeneity of Variance Assumption Was Violated (Across-Case Means)

Case type	No. of cases	α = .05					α = .01				
		F	*W*	BF	KW	NS	*F*	*W*	BF	KW	NS
						$K = 3$					
Equal *n*											
ES	4	.683	.765	.659	.692	.615	.437	.510	.396	.418	.332
EX1	4	.646	.493	.628	.544	.530	.426	.250	.398	.306	.289
(EX1)[a]	(2)	(.803)	(.645)	(.784)	(.694)	(.685)	(.553)	(.376)	(.537)	(.441)	(.420)
EX3	4	.766	.938	.743	.856	.773	.462	.790	.417	.601	.441
(EX3)[b]	(1)	(.554)	(.864)	(.529)	(.729)	(.596)	(.275)	(.643)	(.244)	(.435)	(.254)
Unequal *n*/direct pairing											
ES	4	.665	.909	.855	.741	.635	.386	.741	.646	.441	.336
EX1	4	.634	.680	.804	.622	.563	.394	.421	.613	.368	.299
EX3	4	.760	.992	.940	.907	.798	.392	.948	.734	.567	.398
(EX3)[b]	(3)	(.547)	(.945)	(.803)	(.747)	(.591)	(.205)	(.816)	(.501)	(.318)	(.208)
Unequal *n*/inverse pairing											
ES	4	.691	.538	.387	.617	.597	.489	.265	.179	.366	.347
(ES)[a]	(2)	(.840)	(.701)	(.514)	(.772)	(.758)	(.668)	(.383)	(.250)	(.520)	(.492)
EX1	4	.767	.722	.421	.770	.731	.537	.424	.172	.510	.456
EX3	4	.646	.298	.384	.437	.483	.475	.128	.196	.228	.277
(EX3)[a]	(4)	(.756)	(.404)	(.514)	(.559)	(.605)	(.613)	(.196)	(.290)	(.327)	(.379)
						$K = 4$					
Equal *n*											
ES	4	.681	.779	.660	.688	.605	.432	.535	.403	4.11	.324
EX1	4	.634	.444	.622	.518	.536	.439	.214	.415	.290	.310
(EX1)[a]	(4)	(.756)	(.570)	(.730)	(.632)	(.652)	(.571)	(.310)	(.544)	(.399)	(.414)
EX4	4	.770	.961	.751	.892	.783	.462	.867	.427	.625	.460
(EX4)[b]	(2)	(.584)	(.924)	(.564)	(.756)	(.598)	(.289)	(.754)	(.259)	(.458)	(.274)
2M	4	.688	.816	.667	.709	.632	.442	.578	.414	.440	.350
Unequal *n*/direct pairing											
ES	4	.665	.900	.817	.729	.626	.396	.723	.600	.433	.333
EX1	4	.613	.567	.756	.563	.541	.396	.309	.556	.317	.299
EX4	4	.770	.994	.925	.908	.796	.420	.955	.698	.584	.411
(EX4)[b]	(4)	(.550)	(.964)	(.777)	(.754)	(.593)	(.209)	(.872)	(.456)	(.344)	(.212)
2M	4	.637	.886	.799	.698	.600	.376	.710	.568	.394	.315
Unequal *n*/inverse pairing											
ES	4	.712	.615	.467	.647	.621	.507	.333	.221	.397	.363
EX1	4	.792	.848	.497	.824	.768	.545	.586	.214	.579	.476
EX4	4	.645	.301	.430	.434	.503	.477	.137	.235	.208	.280
(EX4)[a]	(4)	(.754)	(.402)	(.558)	(.337)	(.611)	(.597)	(.199)	(.333)	(.299)	(.379)
2M	4	.728	.678	.471	.686	.646	.514	.387	.221	.426	.386

Note. For all cases, mean structures were specified corresponding to an estimated ANOVA power of .70 of at $\alpha = .05$. F = ANOVA F. W = Welch V_W. BF = Brown and Forsythe F^*. KW = Kruskal-Wallis H. NS = inverse normal scores W. ES = equally spaced means. EX1 is an EX pattern with μ_1 the extreme mean. EX3 is an EX pattern with μ_3 the extreme mean. EX4 is a EX pattern with μ_4 the extreme mean. 2M is a "two in the middle" pattern, with $\mu_1 > \mu_2 = \mu_3 > \mu_4$.