Fisheries Research 107 (2011) 1–13

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

A practical guide to statistical methods for comparing means from two-stage sampling

Susan J. Picquelle, Kathryn L. Mier*

Alaska Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 7600 Sand Point Way NE, Seattle, WA 98115, USA

ARTICLE INFO

Article history:
Received 11 February 2010
Received in revised form 14 June 2010
Accepted 9 September 2010

Keywords:
Cluster sampling
Hierarchical design
Mixed model
Nested ANOVA
Pseudoreplication
Two-staged sampling

ABSTRACT

Two-staged sampling is the method of sampling populations that occur naturally in groups and is common in ecological field studies. This sampling method requires special statistical analyses that account for this sample structure. We present and compare several analytical methods for comparing means from two-stage sampling: (1) simple ANOVA ignoring sample structure, (2) unit means ANOVA, (3) Nested Mixed ANOVA, (4) restricted maximum likelihood (REML) Nested Mixed analysis, and (5) REML Nested Mixed analysis with heteroscedasticity.

We consider a fisheries survey example where the independent sampling units are subsampled (i.e., hauls are the sampling unit and fish are subsampled from hauls). To evaluate the five analytical methods, we simulated 1000 samples of fish lengths subsampled from hauls in two regions with various levels of: (1) differences between the region means, (2) unbalance among numbers of hauls within regions and numbers of fish within hauls, and (3) heteroscedasticity. For each simulated sample, we tested for a difference in mean lengths between regions using each of the five methods. The inappropriate, simple ANOVA that ignored the sample structure resulted in grossly inflated Type I errors (rejecting a true null hypothesis of no difference in the means). We labeled this analysis the Pseudoreplication ANOVA based on the term “pseudoreplication” that describes the error of using a statistical analysis that assumes independence among observations when in fact the measurements are correlated. The result of this error is artificially inflated degrees of freedom, giving the illusion of having a more powerful test than the data support.

The other four analyses performed well when the data were balanced and homoscedastic. When there were unequal numbers of fish per haul, the REML Nested Mixed analyses and the Unit Means ANOVA performed best. The Unequal-Variance REML Nested Mixed analysis showed clear benefit in the presence of heteroscedasticity and unbalance in hauls. For the REML Nested Mixed analysis, we compared three software packages, S-PLUS, SAS, and SYSTAT.

A second simulation that compared samples with varying ratios of among-haul to among-fish variance components showed that the Pseudoreplication ANOVA was only appropriate when the haul effect yielded a p-value >0.50.

Published by Elsevier B.V.

1. Introduction

Two-stage sampling is a common practice in fisheries surveys, as well as many other disciplines. In two-stage sampling, the first stage refers to the primary sampling unit which is a cluster of objects, followed by a second stage where individual objects, or sub-units, are subsampled (or completely enumerated) from the cluster (Cochran, 1977). A cluster, or primary sampling unit, is a natural grouping of objects that may have similar attributes. Examples of this include fish (sub-units) caught in a trawl net (cluster), leaves (sub-unit) on a tree (cluster), and seabirds (sub-unit) in a nesting colony (cluster). This sampling scheme is a form of multilevel sampling (Lehtonen and Pahkinen, 2004) and is also referred to as hierarchical (Raudenbush and Bryk, 2002) or nested sampling. This paper focuses on two-stage sampling; however, the methods discussed here also apply to three-stage or higher levels of hierarchical sampling. The example of two-stage sampling that we use throughout this paper is the comparison of mean length of fish in two geographic regions, where fish are caught in a trawl net at several locations within each region, and then the catch at each location is subsampled for individual length measurements.

Sampling structure must be accounted for in the statistical analysis of the resulting data, and the primary goal of this paper is to compare alternate analytical methods. A two-stage sampling structure requires special handling because the way the data are collected often leads to the lack of independence among the obser-

* Corresponding author. Tel.: +1 206 526 4276; fax: +1 206 526 6723.
E-mail address: Kathy.Mier@noaa.gov (K.L. Mier).

0165-7836/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.fishres.2010.09.009

vations. In fisheries research, one frequently wants to measure attributes of individual fish in the field, for example, length, weight, age, stomach content, or otolith measurements. However, for most fish populations, it is impractical to collect a random sample of individual fish in one sampling stage. Instead, volumes of water are sampled with nets catching a cluster of many fish, and some or all of the fish in this cluster are measured. Because fish tend to occur in proximity of other similar fish of the same species, either because of their schooling behavior or because the local habitat attracts similar fish, sampling a cluster of fish by a net usually results in a sample of fish that are similar to each other. So, simply by virtue of being caught in the same net, the fish are not independent observations, but rather correlated. For example, in measuring fish length, fish tend to aggregate with other fish of similar size, so measuring one fish from a net provides information on the lengths of the other fish caught in the same net.

A secondary goal of this paper is to show that ignoring the sampling structure in the analysis results in erroneous conclusions. Specifically, in two-stage sampling, the error of ignoring that objects were sampled in clusters and treating the objects as a simple random sample is a form of pseudoreplication (Hurlbert, 1984), specifically sacrificial pseudoreplication, which has been well documented, and has been a common mistake made by researchers (e.g., Hairston, 1989; Heffner et al., 1996; Millar and Anderson, 2004). Pseudoreplication has also been described in social and medical sciences, and in education, although they may use a different term. In clustered randomized trials, it is called a “mismatch” problem where the units of assignment do not match the units of analysis (Hedges, 2007; Institute of Educational Sciences, 2007).

The analytical methods examined in this paper avoid pseudoreplication by using either aggregated data (using cluster means) or using hierarchical (or nested) analysis. In contrast, the approach that is often used in the *mismatch* literature referenced above corrects pseudoreplication by modifying the test statistics and degrees of freedom, based on the intra-class correlation coefficient within clusters (Hedges, 2007). We advocate methods that avoid rather than correct for pseudoreplication in the statistical analysis.

Before proceeding, we present some necessary statistical terminology. A “primary sampling unit” is defined as an element within a “sampling frame” that is sampled and is statistically independent of other sampling units within the frame. The sampling frame is defined as the collection of all elements (primary sampling units) accessible for sampling in the population of interest. In ecological sampling, the primary unit may be a quadrat from all possible quadrats within a geographic area (frame). Examples are a parcel of land (quadrat) in a watershed (frame), or a column of water (quadrat) in a body of water (frame). In two-stage sampling, the primary sampling units contain sub-units and measurements are taken on individual sub-units. In experimental contexts, the sub-units within an experimental unit are sometimes referred to as evaluation units (Hurlbert and White, 1993; Kozlov and Hurlbert, 2006; Urquhart, 1981). One such example is the diameters (measurement) of trees (sub-unit) in a land parcel (primary sampling unit or quadrat) where the watershed (frame) is separated into subregions (factor of interest). The example in fisheries that we employ in this paper is the lengths (measurement) of fish (sub-unit) in the water column (primary sampling unit or quadrat) where the body of water (frame) is separated into sub-regions (factor). The individual fish are not independent observations from the population. The lack of independence inherent in hierarchical sampling design precludes the use of simple statistics to estimate means and variances, and to test hypotheses.

Hierarchical measurements occur in both field sampling and in laboratory experiments. Manipulative experiments, as opposed to field sampling, are analogous to two-stage sampling if the experimental unit contains evaluation units receiving the same treatment; here, the experimental unit is analogous to the primary sampling unit, the evaluation units are the sub-units, and the treatment is analogous to the factor of interest. For example, a treatment is applied to a tank of fish (experimental unit, the independent observations) and measurements are made on the individual fish in the tank (sub-units, the correlated observations). Of course more than two stages of sampling may be applied. For example, a third level would be where multiple observations are then made on the individual fish, such as multiple samples of muscle tissue taken from each fish for measurement of pollutant levels. For the sake of simplicity, we focus on just two levels; however, our conclusions can be extended to higher level structures.

There are several other special cases of hierarchical sampling protocols that are sometimes confused with the multi-stage sampling that we examine in this study. To avoid confusion, we list some of these in Appendix H.

In this paper, we focus on two-stage sampling as it applies to surveys of fish populations. We simulated samples based on observed population parameters from a well-studied fish population (i.e., a population of juvenile walleye pollock, *Theragra chalcogramma*) in the Gulf of Alaska (Kendall et al., 1996). Based on these simulated samples, we examined the probability of rejecting a true null hypothesis (commonly labeled as a Type I error rate = α) and the probability of not rejecting a false null hypothesis (commonly labeled as a Type II error rate = β) to evaluate the performance of five different analyses for two-stage sampled data. These analyses included two standard non-hierarchical analysis of variance (ANOVA) (one using a measurement per fish and one using the mean measurement per haul) and three analyses that correctly incorporate a nested structure in the data using both ordinary least squares (OLS) and restricted maximum likelihood (REML) (West et al., 2007) methods.

Our goals were to determine: (1) the most accurate and powerful analyses, (2) how different degrees of unbalance in both the number of sub-units within sampling units and number of sampling units within factor levels affect the results, (3) how unequal variances among the factor levels affect the results, (4) what statistical software packages are currently the most appropriate for this problem, and (5) under what conditions two-stage sampling may be ignored in the analysis, if at all. We also wanted to emphasize the dangers of pseudoreplication, particularly in fisheries science, even though these have already been well publicized, especially in the context of manipulative experiments (Hurlbert, 1984; Hurlbert and White, 1993).

2. Methods

We compared analytical methods for testing the hypothesis of equal means for a hierarchical design, specifically two-stage sampled data, where the first stage consists of primary sampling units within a factor level, each containing many sub-units, and the second stage consists of subsampling the sub-units. In addition to the complexity of a hierarchical design due to the lack of independence among the data, another layer of complexity is added when the sampling (or experimental) units contain different numbers of sub-units. In the laboratory, equal sample sizes are routinely attempted, but rarely attained. For the example of fish subsampled from tanks, fish invariably die during the experiment or some measurements are lost, resulting in unequal numbers of sub-units per experimental unit. In addition, entire tanks might fail, resulting in unequal numbers of experimental units per treatment level. In the field, equal number of sub-units per sampling unit (e.g., numbers of fish per haul where measurements are taken on individual fish) are not expected, and can vary by several orders of magnitude, even after adjusting for the varying amounts of water sampled (i.e., sam-

Table 1
List of conditions that formed the 120 scenarios used to simulate sampled populations: 3 regional effects × 5 levels of haul unbalance × 4 levels of fish unbalance × 2 levels of heteroscedasticity = 120 scenarios.

Level	Region effects	Mean in Region 1	Mean in Region 2
1	Equal	71.6	71.6
2	Differ by 1 SD	71.6	78.2
3	Differ by 2 SD	71.6	84.8
	Haul unbalance (%)	# Hauls in Region 1	# Hauls in Region 2
1	0	15	15
2	20	12	18
3	40	9	21
4	60	6	24
5	80	3	27
	Fish unbalance	# Hauls with # fish per haul	# Hauls with # fish per haul
1	Equal	30 hauls with 25 fish each	
2	Some hauls with 1 fish	24 hauls with 31 fish each	6 hauls with 1 fish each
3	Most hauls with few fish	20 hauls with 5 fish each	10 hauls with 65 fish each
4	Most hauls with many fish	20 hauls with 35 fish each	10 hauls with 5 fish each
	Variability within regions	SD in Region 1	SD in Region 2
1	Equal	6.59	6.59
2	Unequal	4.39	8.78

pling effort). Not only do the numbers of fish per haul vary, but the numbers of hauls per factor level (e.g., year or region or gear) also often vary. Frequently, in retrospective studies, the sampling was not originally designed for the new purpose or analysis. This might result in the numbers of hauls per factor level being unequal and a random variable because they were not fixed for each factor level *a priori*; this is comparable to post-stratification, which has its own issues (Cochran, 1977). Another case where the numbers of hauls per factor level might differ is stratified sampling; frequently the numbers of hauls are proportional to the among-haul standard deviations or mean abundances within each stratum (e.g., Weinberg et al., 2002). This degree of unbalance in both numbers of hauls per region and numbers of fish per haul requires the use of approximations for the variance components in a hierarchical ANOVA (Satterthwaite, 1946; Sokal and Rohlf, 1995) or restricted maximum likelihood (REML) estimation of the variance components (Laird and Ware, 1982; Pinheiro and Bates, 2000).

From here forward, we will follow the example of a fisheries survey where fish are subsampled from hauls; hence, instead of referring to sub-units and units we will refer to fish and hauls. Our example is a comparison of mean lengths of fish from two regions and the null hypothesis is that there is no difference in mean length. Hauls were sampled from the sampling frame of all possible hauls within each region using simple random sampling. Individual fish were then subsampled from the catch of fish in each haul and their lengths were measured, again using simple random sampling.

The analytical methods to test the null hypothesis were assessed by comparing their Type I (rejecting a true hypothesis) and Type II (failing to reject a false hypothesis) error rates, and their accuracy in estimating variance components, i.e., partitioning the variability in the data into haul variance and fish variance. We used simulated data representing a wide range of sampling and population conditions to perform the comparisons (see below).

2.1. Simulation

We based our simulations on observed lengths of juvenile walleye pollock collected by midwater trawling in the Gulf of Alaska (GOA) (Wilson et al., 2006). Samples of fish lengths within each region were simulated from normal probability distributions using the random normal generator within the function “rnorm” in SPLUS, version 7.0. The GOA data yielded the parameters for the normal distributions (Table 1). Rather than simulating a population and then sampling from it, we simulated the samples directly (see details in Appendix A).

We investigated 120 different scenarios (i.e., combinations of several characteristics of populations and sampling designs) simulating 1000 samples for each scenario (Table 1). Each of the 1000 samples from all 120 scenarios consisted of two regions, 30 hauls, and a total of 750 fish lengths. To examine Type I error rates (α) and the power ($1 - \beta$) of the tests, we simulated samples for scenarios with three values for region effect; equal mean lengths in each region (null hypothesis is true and rejecting the null hypothesis is a Type I error), region means differing by one standard deviation and differing by two standard deviations (null hypothesis is false and failing to reject the null hypothesis is a Type II error).

Of primary interest is how the various analytical methods were affected by unequal numbers of fish per haul and unequal numbers of hauls per region (unbalanced design), and unequal among-haul variance between regions (heteroscedasticity). To investigate the effect of unbalance in numbers of hauls among regions in the sample design, we created scenarios where the number of hauls in one region differed from the other by 0% (balanced), 20% (40% of hauls in one region vs. 60% in the other region), 40% (30% vs. 70%), 60% (20% vs. 80%), and 80% (10% vs. 90%). In all haul allocation schemes, the total number of hauls sampled remained the same. For unbalance in numbers of fish among hauls, we looked at four allocation schemes, (1) all hauls had an equal number of fish, (2) some hauls had a single fish, (3) most hauls had few fish and (4) most hauls had many fish (Table 1). In all fish allocation schemes, the total number of fish sampled remained the same.

To examine heteroscedasticity, we examined all scenarios described above with both equal and unequal among-haul variances for each region. For the unequal variance scenarios, among-haul variance in one region was four times that of the other region. One valid reason for an unbalanced design is to base the number of samples on the variability within a factor level. Hence, for the unequal variance, we put the greater number of hauls in the region with the higher among-haul variance. Forming all possible combinations of these factors led to 120 unique scenarios, i.e., 3 (region effects) × 5 (unbalance in hauls) × 4 (unbalance in fish) × 2 (heteroscedasticity) (Table 1).

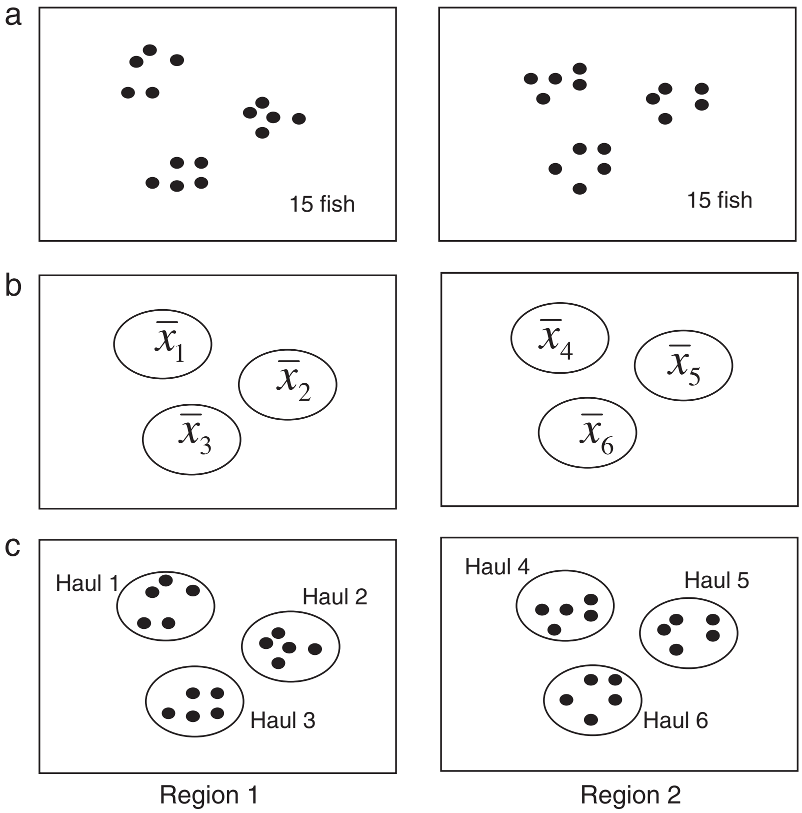

Fig. 1. Schematic of three approaches to the same survey design. Dots represent fish; in (b) $\bar{x}_i$ represents the mean measurement averaged over fish within haul. Ovals represent hauls. Rectangles represent regions. Positions of dots, ovals, and rectangles are not meant to suggest a particular survey design, e.g., random or systematic. (a) Fish from several hauls were pooled as if independent. (b) A single mean measurement was calculated for each haul. (c) Fish were nested within each haul.

2.2. Analysis comparisons

For each scenario, 1000 samples were created, and for each sample, five analyses were applied to the data. These included: (1) one-way ANOVA using OLS where the subsampling was ignored; that is, all data were pooled as if independent (Pseudoreplication ANOVA), (2) one-way ANOVA using OLS where haul means were used as observations (Unit Means ANOVA), (3) nested ANOVA using OLS where haul was treated as a random factor and was nested within region, which was fixed (OLS Nested Mixed ANOVA), (4) nested mixed analysis using REML where haul was a random factor nested within region, which was fixed (REML Nested Mixed analysis), and (5) nested mixed analysis using REML that allowed for unequal among-haul variances between regions (Unequal-Variance REML Nested Mixed analysis). For the two nested analyses where we used REML, we applied an ANOVA-type *F*-test to the REML estimates to get approximate *p*-values. REML is not a true ANOVA in the classical sense of partitioning the sum of squares. Fig. 1 illustrates these different approaches to the hypothesis test, where all three nested analyses are represented in Fig. 1c because they differed only in computations, not in their structure.

The choice of analyses to compare was determined by different statistical software programs, specifically S-PLUS, SAS, and SYSTAT.[1] Most software packages have equivalent algorithms for standard, balanced, equal variance ANOVAs, but methods diverge when data are nested and unbalanced or heteroscedastic.

Using OLS for nested mixed ANOVAs is only appropriate for balanced data and requires further post-analysis to test the treatment factor effect appropriately; this is the OLS Nested Mixed ANOVA in our simulation. In contrast, the restricted maximum likelihood estimation compute approximate *F*-tests for the treatment effects without post-analysis manipulation (although SPLUS requires the subsequent use of the “anova” function on the created lme object). We included the OLS Nested Mixed ANOVA of unbalanced data to explore the impact of data unbalance and to emphasize the advantages of using REML for unbalanced data.

Different software packages have different defaults and yield different results for the OLS Nested Mixed ANOVA, depending on which sum of squares is used. Therefore, we compared two computational methods for the OLS Nested Mixed ANOVA, Type 1 and Type 3 sum of squares. [Unfortunately, the accepted nomenclature in statistical literature is Types I and II errors and Types I and III (and other) sums of squares; to avoid confusion, we distinguish between these by using Roman numerals when referring to errors and Arabic numerals when referring to sum of squares.] Type 1 sum of squares is “sequential”, in that each term is added to the model sequentially (order of terms in the model matters). This is recommended for model fitting for prediction (Milliken and Johnson, 1992). Type 3 sum of squares (Yates' weighted squares of means method, Milliken and Johnson, 1992), or “partial” sum of squares, shows the contribution of that term given that all other terms are in the model (order does not matter). This is referred to as a “marginal” or conditional significance test and is recommended for hypothesis testing (Milliken and Johnson, 1992).

Source code for comparing analyses is available upon request from Kathy.Mier@noaa.gov.

[1] Reference to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the United States Government. The views and opinions of authors expressed in this paper do not necessarily state or reflect those of the United States Government, and shall not be used for advertising or product endorsement purposes.

2.3. Type I and Type II errors

The primary criterion that we used to determine the best analysis was to compare the rate that the analysis rejected a true hypothesis (Type I error) to the specified α for the hypothesis test of no difference between region means. Type I errors were counted from all 1000 samples from each of the 40 scenarios where region means were equal and compared among the five analyses. Similarly, Type II errors (failing to reject a false hypothesis) were counted from all 1000 samples from each of the scenarios where region means differed. The tabulated Type I and Type II error rates were then compared among the five analyses. For these comparisons, we set $\alpha = 0.05$.

Note that $\alpha = 0.05$ is an arbitrary test criterion and basing Type I and Type II error rates on arbitrary significance levels are directly relevant only to hypothesis testing. In practice, we highly recommend reporting p-values as evidence in support of a hypothesis instead of accepting or rejecting a hypothesis based on arbitrary significance levels. We use Type I and Type II error rates here for a different purpose—to evaluate the accuracy of different analyses, i.e., how close is the computed p-value associated with the test statistic from the analysis to the actual probability of observing that value of the statistic. Type I error rate compares the nominal probability of $\alpha = 0.05$ with the actual probability, which we estimated by the observed proportion of 1000 simulations that resulted in a probability of 0.05 or smaller. If the observed Type I error rate, i.e., the actual probability, is close to the nominal probability, then the test is accurate. A more rigorous measure of the accuracy of the tests would be to compute the observed p-values for a variety of nominal probabilities and then compare these to an F-distribution. However, the probabilities that are of greatest interest are the small values, so we restricted our comparisons to a single nominal probability, the α value of 0.05. We chose this criterion because, traditionally in ecology, as opposed to quality control in manufacturing, hypothesis testing has usually emphasized controlling and minimizing the probability (α) of incorrectly rejecting a true null hypothesis (Type I error), rather than the probability (β) of failing to reject a false null hypothesis (Type II error). However, the frequency of Type II errors was included in this evaluation as this is indicative of the power of the test (i.e., the ability to correctly reject a false null hypothesis, or $1 - \beta$).

2.4. Variance components

Another measure of performance of the five analyses was to compare the estimated variance components from each of the methods to those used to simulate the sample data (Appendices A and B). That is, we specified a variance about the haul means when simulating them and consequently, a component of the mean square for haul is an estimate of this specified variance. Similarly, we specified a variance for the individual fish lengths when simulating them and the mean square error (MSE) is an estimate of that variance. We focused on the variance among primary sampling units within factor level (i.e., variance among haul means within region) because this is critical for testing for differences among region means. See Appendix B for details.

2.5. Software comparisons

The three software packages (S-PLUS, SAS, and SYSTAT) differ in subtle computational details of the nested analyses. The software code for the REML Nested Mixed analyses is challenging, hence we show it in Appendix C. The details of the software comparisons are in Appendix D.

2.6. Pooling

We conducted a second simulation study to elucidate the conditions that might allow pooling the two sampling stages (and corresponding sums of squares) without compromising the validity of the hypothesis tests. Hurlbert (2004) labeled this "test-qualified pseudoreplication". In other words, when is it acceptable, if ever, to ignore the nested structure of the sampling design and treat all the subsampled individuals (sub-units, in our example, fish) as independent observations? Details of this pooling simulation and brief review of the literature are in Appendix E.

3. Results

3.1. Type I errors

The Pseudoreplication ANOVA rejected the null hypothesis 40–75% of the time when in fact it was true (Type I error) (Fig. 2). This is 8–15 times the nominal 5% error rate. This was true regardless of degree of unbalance in numbers of hauls (primary units) or fish (sub-units), or unequal variances among hauls within region (factor level). The simulations of the equal-variance scenarios resulted in a frequency of Type I errors for the Pseudoreplication ANOVA that was almost always at least 60% (Fig. 2a–d). When numbers of fish were equal among hauls, all non-pseudoreplication analyses were almost identical and produced the desired Type I error rate of 5% (Fig. 2a). A single exception was a slightly inflated Type I error rate for the Unequal-Variance REML Nested Mixed analysis when haul unbalance was >40%.

Details of how the Type I error rates were impacted by unbalance in numbers of haul and fish, and by unequal variances among hauls within the two regions are presented in Appendix F and are summarized here: (1) the Pseudoreplication ANOVA, i.e., the ANOVA based on individual fish measurements, rejected a true null hypothesis 8–15 times more often than the Unit Means ANOVA and REML Nested Mixed analyses; (2) the Unit Means ANOVA and the REML Nested Mixed analysis performed well and nearly identically in all scenarios; and (3) the OLS Nested Mixed ANOVA was equivalent to the Unit Means ANOVA and the REML Nested Mixed analysis when there were equal numbers of fish sampled in each haul, but was inaccurate when fish were unbalanced. For the homoscedastic scenarios (equal among-haul variances): (1) when the numbers of fish sampled from each haul were the same, then all non-pseudoreplication analyses (except for the Unequal-Variance REML Nested Mixed analysis) produced the desired 5% Type I error rate, regardless of unbalance in haul; and (2) when there were unequal numbers of fish sampled, the REML Nested Mixed analysis and the Unit Means ANOVA performed best, maintaining a 5% error rate. For the heteroscedastic scenarios (unequal among-haul variances): (1) for all analyses, except the Unequal-Variance REML Nested Mixed analysis, the rate of rejecting a true null hypothesis actually decreased as the level of unbalance in the numbers of hauls increased, but only if the number of hauls was higher for the region with higher variance; and (2) the Unequal-Variance REML Nested Mixed analysis maintained the desired 5% error rate until haul unbalance was >60%.

3.2. Type II errors

As expected, the rate of failing to reject a false null hypothesis (Type II error) decreased when the true difference between region means increased (Figs. 3 vs. 4). Also expected, the Pseudoreplication

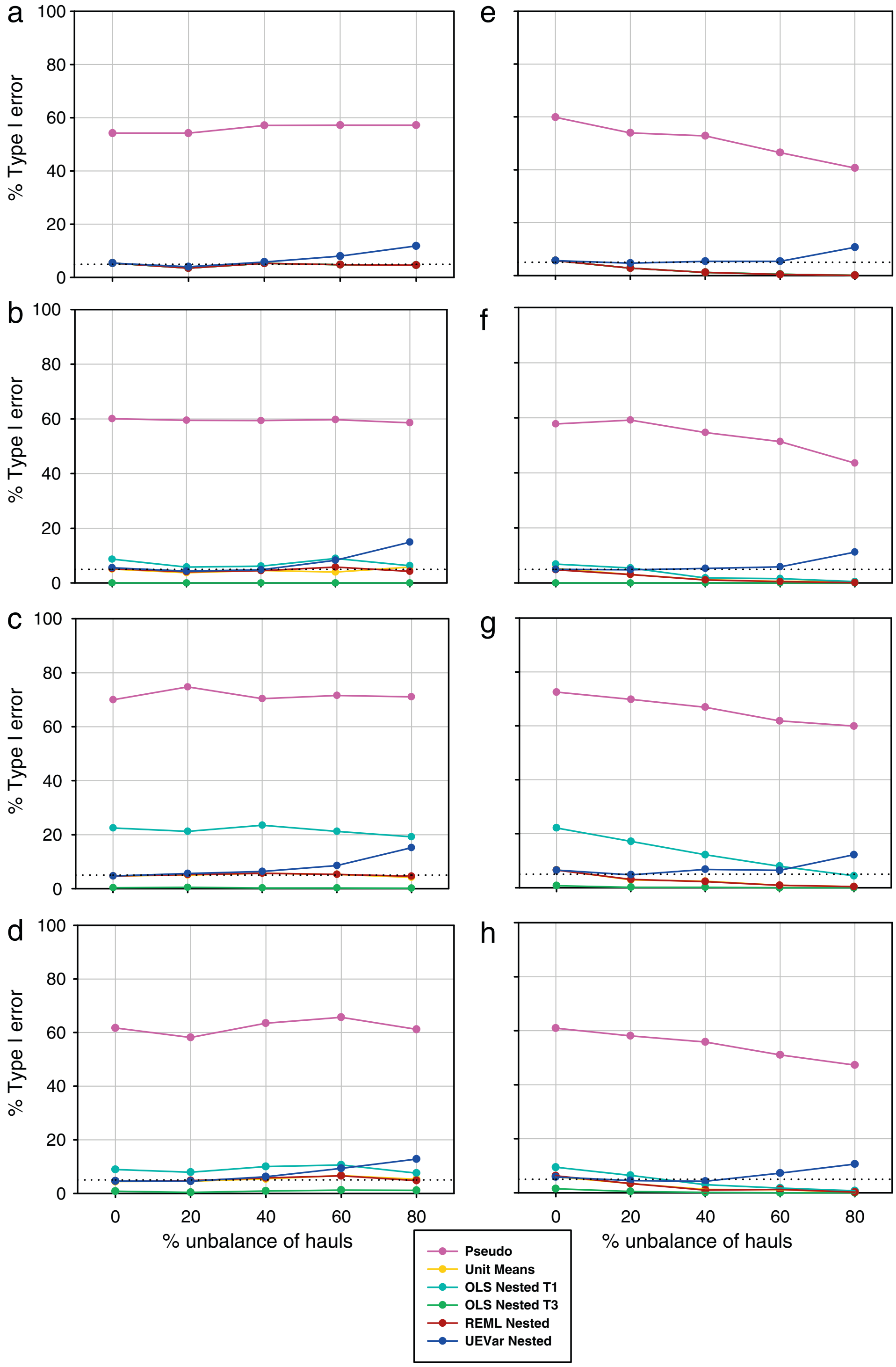

Fig. 2. Observed Type I error rates (%) for 1000 simulations with equal (a–d) and unequal (e–h) among-haul variances. Panel titles for Figs. 2–5 are (a and e) equal number fish per haul, (b and f) some hauls with just one fish, (c and g) most hauls with few fish, (d and h) most hauls with many fish. The dotted line corresponds to $\alpha = 0.05$. The line labels in Figs. 2–4 are abbreviated as follows: Pseudo = Pseudoreplication ANOVA, unit mean = Unit Means ANOVA, OLS Nested T1 = OLS Nested Mixed ANOVA using Type 1 sum of squares (default for SPLUS), OLS Nested T3 = OLS Nested Mixed ANOVA using Type 3 sum of squares (default for SYSTAT), REML Nested = REML Nested Mixed analysis assuming equal variances, and UEVar Nested = REML Nested Mixed analysis accounting for unequal variances. Missing lines on graphs are obscured by the REML or UEVar nested lines.

ANOVA had the lowest Type II error rate (5–15%, which is extremely low) and thus had the greatest power (Figs. 3 and 4). However, this is a consequence of an extremely high error rate of rejecting a true null hypothesis (Type I error) (Fig. 2). This apparent increase in power is spurious and will mislead the researcher into thinking they have a more precise estimate than they really do. Pseudoreplication almost always biases *p*-values downwards which leads to a very high rate of rejecting the null hypothesis whether it is true or false.

Details of how the Type II error rates were impacted by unbalance in numbers of hauls and fish, and by unequal variances among hauls within the two regions are presented in Appendix G and

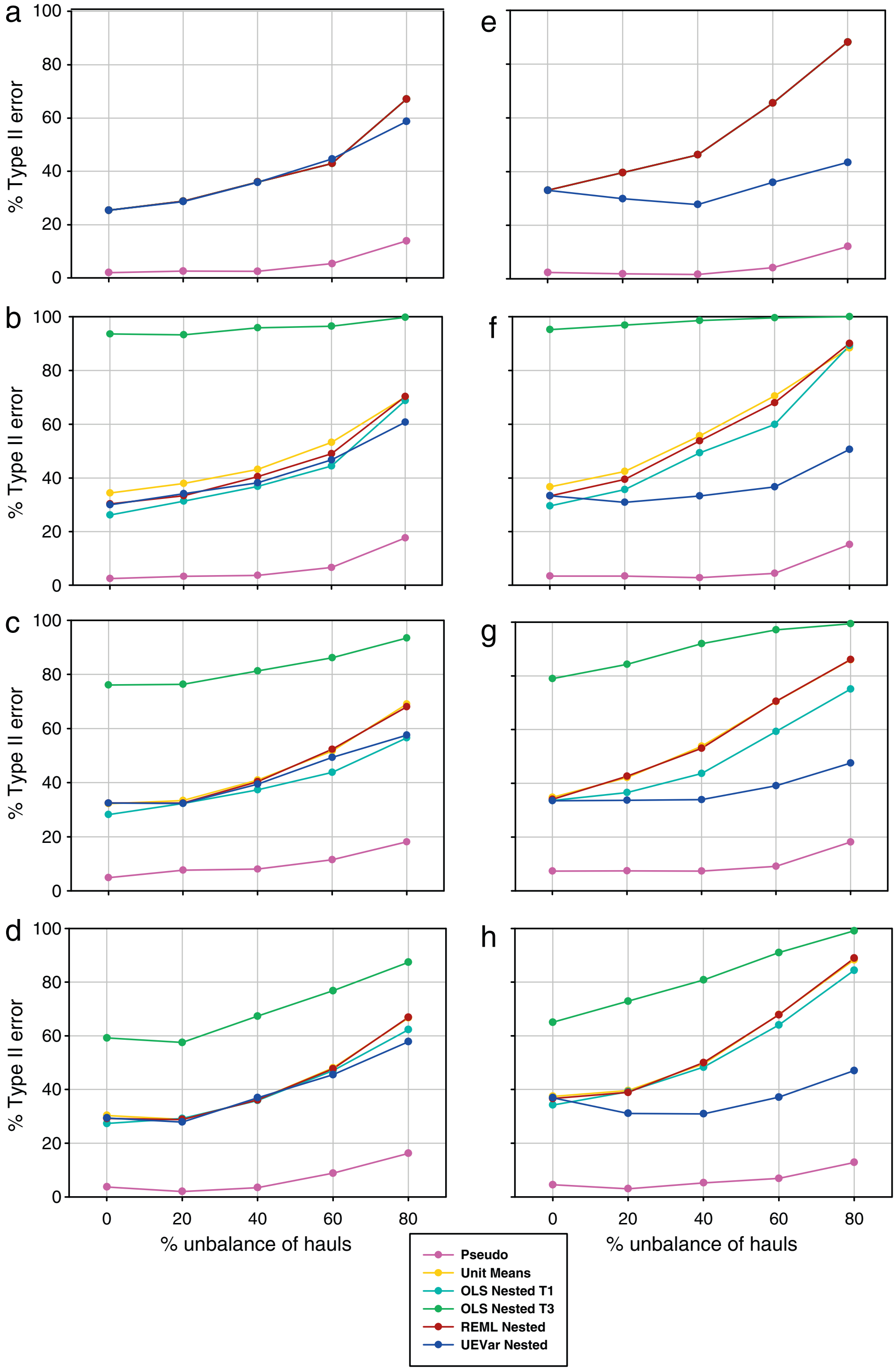

Fig. 3. Observed Type II error rates (%) for 1000 simulations with equal (a–d) and unequal (e–h) among-haul variances; region means differing by one standard deviation. Panel titles and line labels as in Fig. 2. Missing lines on graphs are obscured by the REML or UEVar nested lines.

are summarized here: (1) the Pseudoreplication ANOVA artificially increased the power of the test (reduced the rate of failing to reject a false null hypothesis); (2) when the numbers of fish sampled from each haul were equal and the among-haul variances were equal, then all non-pseudoreplication analyses performed equally well; (3) when there were unequal numbers of fish sampled, then all non-pseudoreplication analyses performed well except for the OLS Nested Mixed ANOVA using the Type 3 sum of squares; (4) the power tended to decrease with increasing unbalance in numbers of hauls (Type II error increased). For the heteroscedastic scenarios: (1) when the numbers of hauls were unequal, the Unequal-Variance REML Nested Mixed analysis had greater power (smaller Type II error) than the other non-pseudoreplication analyses, and this improvement increased as unbalance in hauls increased; (2) when the numbers of hauls were equal, the Unequal-Variance REML Nested Mixed analysis did not perform appreciably better than the Unit Means ANOVA or the REML Nested Mixed analysis.

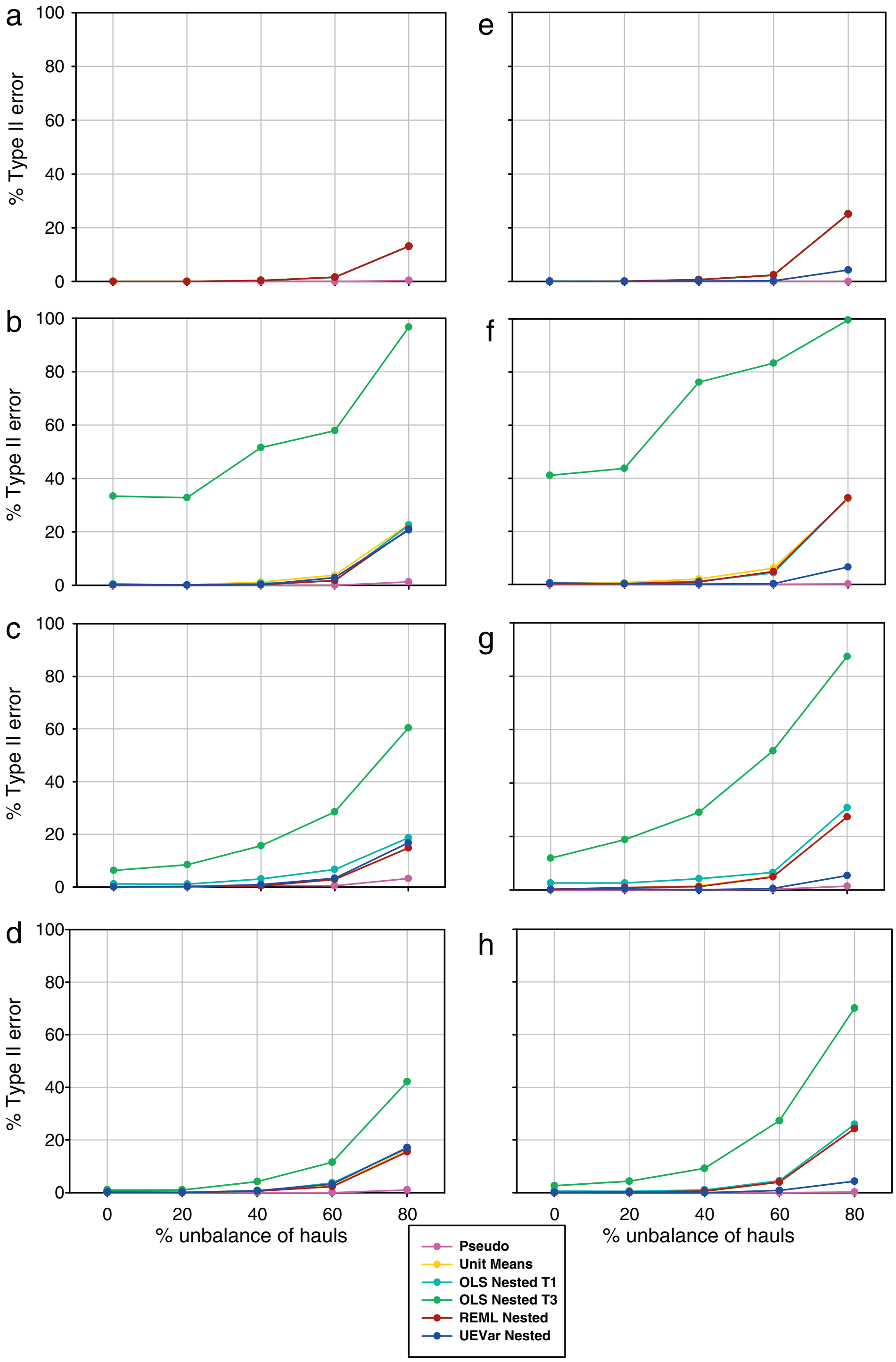

Fig. 4. As in Fig. 3, but for region means differing by two standard deviations.

3.3. Variance components

Another criterion for comparing these analytical methods is how well the estimated variance components, as derived from observed mean squares, match the true variance components that were used to simulate the data (Appendix B, $\sigma_{h(r)}$ vs. Appendix A, $\sigma'_{h(r)}$). Detailed results are in Appendix B. Table 2 summarizes the results of the analytical method comparisons based on the Type I and Type II error rates and on estimation of variance components.

3.4. Software comparisons

Details of the software comparison are in Appendix D, and are summarized here: (1) the p-values from SAS (Kenward–Roger), SYSTAT (containment) and S-PLUS (classical balanced) were equivalent for the REML Nested Mixed analysis that assumed equal variances among hauls; (2) the p-values from SAS were larger than S-PLUS for the Unequal-Variance REML Nested Mixed analysis when hauls were unbalanced; (3) the only situation where differences among the software packages impacted Type I or Type

Table 2
Performance evaluations of five analytical methods. The performance measures are based on Type I error rates (how close to $\alpha = 0.05$), Type II error rates (how small), and accuracy of estimates of variance components (how close to actual variances). Scenarios are reduced to four categories based on balance of the design: completely balanced (=*H*, =*F*), unbalanced in hauls and balanced in fish (≠ *H*, =*F*), balanced in hauls and unbalanced in fish (=*H*, ≠ *F*), and completely unbalanced (≠ *H*, ≠ *F*), where *H* represents the numbers of hauls within regions and *F* represents the numbers of fish within hauls.

	Type I errors				Type II errors				Variance estimate			
Balance	=*H*,=*F*	≠ *H*,=*F*	=*H*, ≠ *F*	≠ *H*, ≠ *F*	=*H*,=*F*	≠ *H*,=*F*	=*H*, ≠ *F*	≠ *H*, ≠ *F*	=*H*,=*F*	≠ *H*,=*F*	=*H*, ≠ *F*	≠ *H*, ≠ *F*
Analyses												
Pseudoreplication ANOVA	–	–	–	–	–[a]	–[a]	–[a]	–[a]	NA	NA	NA	NA
Unit Means ANOVA	**+**	**+**/–	**+**	**+**/–	**+**	**+**/–	**+**	**+**/–	**+**/–	**+**/–	–	–
OLS Nested Mixed ANOVA, Type 1 SS	**+**	**+**/–	–	–	**+**	**+**/–	**+**	**+**/–	**+**/–	**+**/–	**+**/–	**+**/–
OLS Nested Mixed ANOVA, Type 3 SS	**+**	**+**/–	–	–	**+**	**+**/–	–	–	NA	NA	NA	NA
REML Nested Mixed Analysis	**+**	**+**/–	**+**	**+**/–	**+**	**+**/–	**+**	**+**/–	**+**/–	**+**/–	**+**/–	**+**/–
Unequal-Variance REML Nested Mixed Analysis	**+**	+/**+**	**+**	+/**+**	**+**	**+**	**+**	**+**	**+**	–/**+**	**+**	–/**+**

"**+**" = superior; "+" = acceptable; "–" = unacceptable; "NA" = not available; "/" separates different scores for equal and unequal variance data where needed.

[a] Even though the Pseudoreplication ANOVA gave the smallest Type II error rates, we graded it as unacceptable because these rates were artificially reduced.

II errors was when the Unequal-Variance REML Nested Mixed analysis was used with extreme unbalance among hauls (>40%); (4) when hauls are unbalanced, SAS resulted in smaller Type I errors closer to the nominal 5%, but with a slightly larger Type II error rate compared to S-PLUS. Table 3 outlines these results.

3.5. Pooling

The simulation that examined pooling the among-haul and within-haul terms in the ANOVA, thus ignoring the nested structure, resulted in a well-defined guideline for pooling. Theoretically, if there is no haul effect, then the fish within a haul are not correlated and can be treated as if they are independent observations, even though the fish from the pooled hauls do not comprise a simple random sample of fish. Hence, if the collective samples of fish adequately mimic a simple random sample, then the Pseudoreplication ANOVA will not result in any of the erroneous results described in the previous sections.

The guideline for pooling is a specified value of α for testing the significance of the haul effect that essentially guarantees that there truly is no haul effect. Satisfying this stringent test of the haul effect ensures that pooling will result in the actual rate of rejecting a true null hypothesis (Type I error) being close to the nominal Type I error rate for testing the region effect. In all scenarios, a 0.50 *p*-value for the test of no haul effect guaranteed a 5% Type I error rate for testing the region effect (Fig. 5). In other words, for the variety of scenarios tested, if one wishes the protection of an actual α of 0.05 for testing region, then one may pool hauls only when the test for a difference among hauls yields a *p*-value >0.50. The trends were the same no matter what level of unbalance in hauls for the equal-variance scenarios, and showed only slight differences for the unequal-variance scenarios. Interestingly, the more unbalanced the numbers of hauls for

Table 3
(a) Software procedure characteristics for the REML Nested Mixed analysis.

	SAS	S-PLUS	SYSTAT
Name of procedure	PROC MIXED	lme	MIXED
Sum of squares used	Type 1 Type 3	Type 1	Type 3
Unequal variance option	Yes	Yes	Possible with some manipulation and only if balanced
Approximation method(s) for denominator degrees of freedom	• Containment • Between/Within • Residual • Satterthwaite • Kenward–Roger	• Classical decomposition for balanced multi-level ANOVA	• Containment (factor levels < 300) • Residual otherwise

(b) Software procedure performances for the REML Nested Mixed analysis. EV refers to the analysis that assumes equal variances and UEV refers to the analysis that assumes unequal variances.

	SAS		S-PLUS		SYSTAT
	EV	UEV	EV	UEV	EV
Type I error, equal variance data	Superior	Acceptable[a]	Superior	Acceptable[b]	Superior
Type I error, unequal variance data	Acceptable[c]	Superior	Acceptable[c]	Acceptable[a]	Acceptable[c]
Recommendations	• Best Type I error • Acceptable Type II error for unequal variance data • Most versatile and well documented		• Acceptable Type I error • Best Type II error for unequal variance data		• Unequal-Variance REML Nested Mixed analysis unavailable for unbalanced designs

[a] If haul unbalance <80%.
[b] If haul unbalance <60%.
[c] But high Type II error.

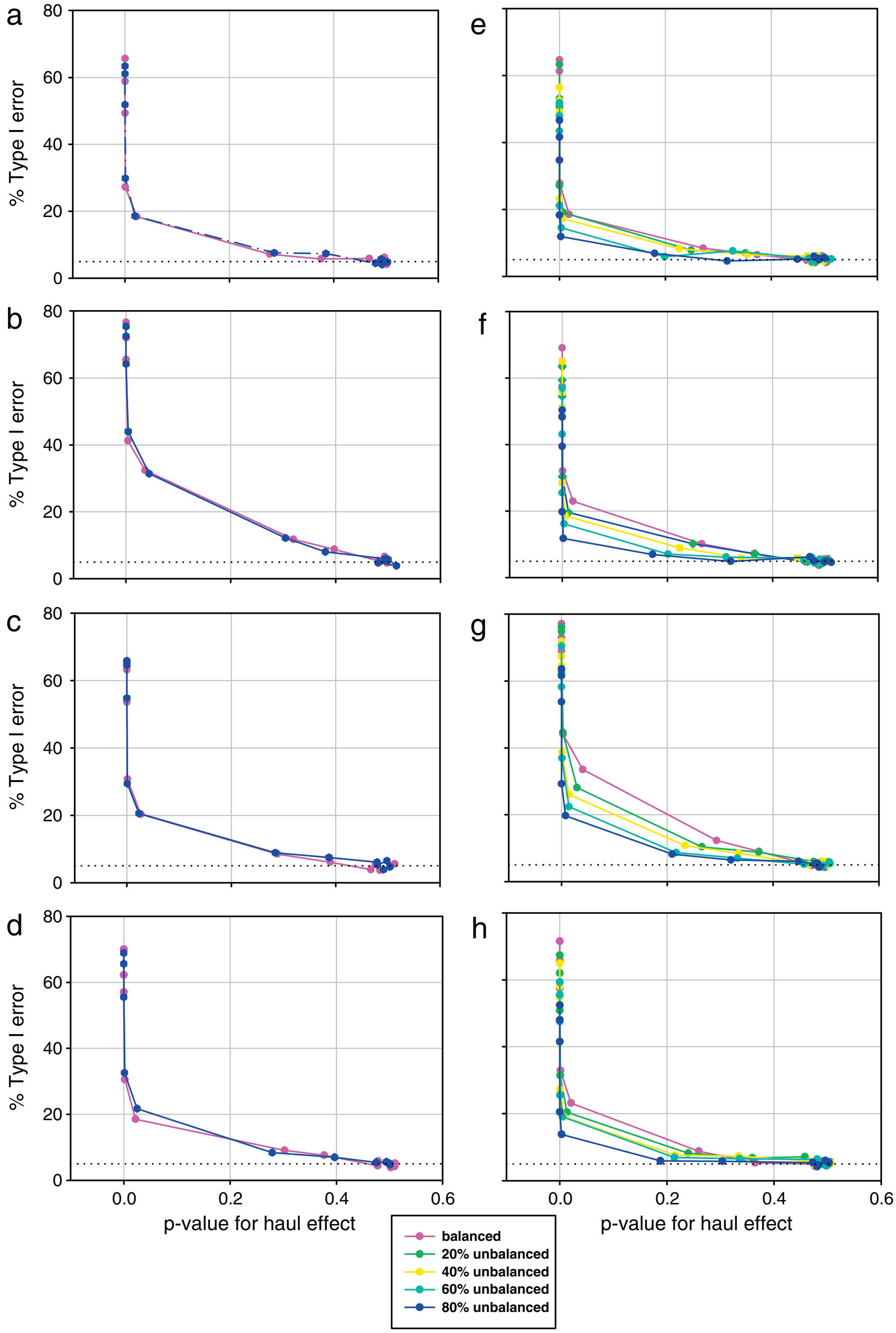

Fig. 5. Observed Type I error rates (%) from Pseudoreplication ANOVA plotted against the mean *p*-value from the *F*-test for the haul effect, averaged over 1000 simulations with equal (a–d) and unequal (e–h) among-haul variances. Panel titles are as in Fig. 2. Results for equal among-haul variances are shown only for scenarios that were balanced in hauls and 80% unbalanced because all intermediate levels of unbalance were indistinguishable. The points on the lines correspond to the discrete ratios of the among-haul variance to the within-haul variance used in the simulations; these are not labeled, but fall in order from largest (2.0) to smallest (0.00001) going from left to right along the plotted lines. The dotted line corresponds to $\alpha = 0.05$. Missing lines on graphs are obscured by the 80% unbalanced line.

the unequal-variance scenarios, the smaller the *p*-values for haul effect that led to Type I error rates approaching 5%. This was undoubtedly due to our haul allocation where more samples were taken in the region with the larger variance in the unbalanced designs.

4. Discussion

Statistical assessment of hypotheses is fraught with pitfalls and although statistical packages have made hypothesis testing easy to execute, they have also made it easy to mis-specify the ANOVA and to misinterpret the results. To aid researchers in choosing an appropriate analysis in multi-stage sampling, we presented and compared several analyses (i.e., Pseudoreplication ANOVA, Unit Means ANOVA, Nested Mixed ANOVA using OLS, and Nested Mixed analysis using REML). Other problems beyond mis-specification in hypothesis testing include the arbitrary specification of a value for the significance level α and ignoring the power of a test. These concerns are discussed in greater detail in Balleurka et al. (2005), Boruch (2007), Lombardi and Hurlbert (2009), Nickerson (2000), Parkhurst (2001), and Ziliak and McCloskey (2008). We employ hypothesis testing with the arbitrary significance level of $\alpha = 0.05$ in this study only as a device to compare analytical methods; our focus on Type I error rates (the rate of rejecting a true null hypothesis) is not meant to imply our endorsement of hypothesis testing as a means of assessing hypotheses. Rather than accepting or rejecting hypotheses, we prefer to report the probability associated with a test statistic as evidence in support of a null or alternative hypothesis.

Our simulations comparing analytical methods for two-stage sampling showed that the Pseudoreplication ANOVA performed by far the worst of all analyses; it had the highest probability of indicating a difference when there was none, much greater than the specified rate of 5%. In fact, the frequency of "detecting" a non-existent difference was almost always >50%. In the long run, flipping a coin, thus avoiding the inconvenience and expense of collecting data, would give better results! All of the alternative analyses performed equally well for data that are homoscedastic (equal variance among hauls within each region) and have equal numbers of fish among the hauls, but this balance is frequently impossible to attain.

The consequence of unbalance among numbers of fish is that it invalidates the OLS Nested Mixed ANOVA, i.e., the Type 1 sum of squares for this OLS ANOVA has a high rate of Type I errors (rejecting a true null hypothesis), and the Type 3 sum of squares has a high rate of Type II errors (failing to reject a false null hypothesis). Hence, if there is unbalance among numbers of fish, we recommend the REML Nested Mixed analysis or the Unit Means ANOVA as being both accurate and powerful. This recommendation holds regardless of the level of unbalance in hauls within regions, making these analytical methods relatively robust to many of the problems encountered by field studies. In addition to unbalance in fish and hauls, if the data are also heteroscedastic (unequal variance among hauls within each region), then we recommend the Unequal-Variance REML Nested Mixed analysis because this reduces the Type II error rates.

Comparing the REML Nested Mixed analyses to the Unit Means ANOVA, we see that each has advantages and disadvantages. The REML Nested Mixed analyses require more sophisticated software, while the Unit Means ANOVA is computationally much easier. However, the Unit Means ANOVA produces a slightly biased estimate of the standard deviation among the haul means, and provides no estimate of the standard deviation among the fish. In spite of this bias and the loss of information about the among-fish variability, our simulations show that the hypothesis test from the Unit Means ANOVA was just as accurate and powerful as the more complete and complex REML Nested Mixed analyses, if the data are homoscedastic. This observation is consistent with Hurlbert's assertion that a unit means analysis is just as powerful as a nested analysis (Hurlbert, 1984). The Unit Means ANOVA does not allow predictors or factors at the sub-unit level, and other experts claim that ignoring variation at multiple levels can result in biased or inefficient estimates of between-unit variance components (Raudenbush and Bryk, 2002). Our simulations of the Unit Means ANOVA demonstrated a bias, however it was small.

Comparing the impact of unbalance in hauls to unbalance in fish, we found that unbalance in fish only affects the OLS Nested Mixed ANOVA, invalidating it, and minimally affects the REML Nested Mixed analyses. In contrast, the unbalance in hauls affects all the analyses. However, unbalance in hauls only affects the power of the test, i.e., Type II error (failing to reject a false null hypothesis) increases as the degree of unbalance in hauls increases, and this impact is substantial only at extreme levels of unbalance. Similarly, unbalance in hauls affects the relative performances of the software packages (based on the comparison of p-values and error rates) only at extreme levels of unbalance. Unbalance in numbers of hauls within regions should be of little concern if the unbalance is 20% or less (i.e., 40% of the hauls in one region and 60% of the hauls in the other).

The effect of heteroscedasticity among hauls within region on the rate of rejecting a true hypothesis (Type I error) was minimal in our simulation, even when the among-haul variance in one region was four times that of another. The robustness of all analyses in the presence of extreme heteroscedasticity was reassuring. When the numbers of hauls were balanced, the effect of unequal variances was also minimal on the rate of failing to reject a false hypothesis (Type II error). When the numbers of hauls were unbalanced, the effect of unequal variance was to exaggerate the effect of the unbalance on the Type II error, which was an increase in error with an increase in haul unbalance. Unlike the Type II errors, the Type I errors were reduced with greater haul unbalance for all analyses except the Unequal-Variance REML Nested Mixed analysis, but this is likely an artifact of our allocation of hauls in the two regions (i.e., sampling more hauls in the region with higher variance) and may not hold in general. Also noteworthy is that the Unequal-Variance REML Nested Mixed analysis showed little impact from unbalance in hauls and maintained constant Type I and Type II error rates until the unbalance in hauls was extreme. However, our unbalanced designs may have mitigated the effect of the unequal variances when numbers of hauls were unbalanced because more hauls were sampled in the region with the higher variance.

Both REML Nested Mixed analyses did a good job of estimating the standard deviation among hauls. The Unit Means ANOVA overestimated the standard deviation, but this did not impact the accuracy of its ANOVA results. This bias was reduced when the numbers of fish were equal among hauls. If the data are heteroscedastic then only the Unequal-Variance REML Nested Mixed analysis provided estimates of multiple variances and these estimates were very accurate.

Overall, SAS *Proc Mixed* or S-PLUS *lme* software routines proved to be better than SYSTAT (version 12.0 for Windows 2004) for analyzing hierarchical designs as SYSTAT did not conveniently allow for unequal variances for the REML Nested Mixed analysis.

In our simulations, pooling fish from different hauls within a region did not inflate the Type I error rate if $\alpha = 0.50$ was used to test the significance of the haul effect. Attaining such a high p-value required an extremely small variance among the haul means relative to the variance among the fish within hauls. Even if the variance of the fish is known to be more than 100 times the variance of the hauls, the Type I errors for testing regions were much greater than 5%, that is, pseudoreplication remained a problem until the among-haul variance was extremely small. Hence, we recommend against pooling haul (unit) and fish (sub-unit) variances unless the F-test for the haul effect is not significant at $\alpha = 0.50$ (i.e., $p \geq 0.5$). A word of caution—this value for α was an adequate criterion for our scenarios, but might not apply to smaller sample sizes, greater heteroscedasticity, or larger variances. A large p-value may not be strong evidence that the haul effect is zero or close to it; instead it might just indicate that the power of the test is low due to very high variances or low sample sizes. In addition, some researchers might want to interpret our pooling criterion as a test of independence, but it is not. Our pooling criterion simply detects the point at which the correlation among the sub-units is small enough that it has no consequence on probability statements about the main effect. If a researcher wants to avoid the complexity of using a nested analysis and cannot or does not want to rely on this criterion for pooling, then the Unit Means ANOVA is the only valid option. However, the Unit Means ANOVA might not be sufficient when questions being asked require multi-level analyses that incorporate covariates for units at different levels in the hierarchy (e.g., Bickel, 2007; Hox, 2002; Goldstein, 2003; Raudenbush and Bryk, 2002).

Our results were based on a very simple model – one fixed factor with 2 levels and one nested random factor – but we anticipate that our conclusions will apply to more complicated analyses. Complex analytical methods present challenges to researchers beyond the scope of this paper, however, there are several helpful books to guide the researcher through model specification and analysis (e.g., Milliken and Johnson, 1992; Quinn and Keough, 2002; Raudenbush and Bryk, 2002; West et al., 2007). More important than the complexity of the analysis is the messiness of the data, and our 120 scenarios have covered a wide range of messy data (i.e., unbalanced and heteroscedastic). We included the worst case scenario of having just one fish per haul for some hauls, which we show is clearly problematic. However, biological data can have sample sizes smaller than our 30 hauls, be more unbalanced, and have variances that differ by more than fourfold, as did the extreme cases in our examples. One can only speculate whether our conclusions apply to data with fewer samples or more extreme heteroscedasticity.

In conclusion, nested structure in a survey or experimental design can rarely be ignored. We presented results from comparisons of three hierarchical analyses that incorporated nesting, a non-hierarchical ANOVA using means of aggregated data, and the non-hierarchical Pseudoreplication ANOVA that ignored the nested structure. Using the inappropriate Pseudoreplication ANOVA produced seriously inflated Type I errors, i.e., it rejected a true null hypothesis more often than not; the level of inflation decreased with the size of the variance component from hauls relative to the variance component from fish. This Pseudoreplication ANOVA yielded accurate Type I errors when the *F*-test for the haul effect produced a *p*-value of at least 0.50. When the numbers of fish (sub-units) per haul (unit) were the same and the data were homoscedastic, all non-pseudoreplication analyses performed equally well. When there are unequal numbers of fish per haul, we recommend one of the two REML Nested Mixed analyses or a Unit Means ANOVA. The Unit Means ANOVA offers simplicity, but at the cost of a slightly biased estimate of the haul variance component and no estimate of the within-haul variance. When the data were heteroscedastic (unequal among-haul variances in the two regions), the Unequal-Variance REML Nested Mixed analysis showed clear benefit over other analyses that assumed equal variances, but only when the number of hauls were unbalanced in the two regions, and SAS is preferred to S-PLUS in this case (Systat does not offer an option for heteroscedastic data). Heteroscedasticity had a minimal effect if the numbers of hauls in each region were equal. Unbalance in numbers of fish greatly impacted the rate of rejecting a true hypothesis (Type I error) and failing to reject a false hypothesis (Type II error) for the OLS Nested Mixed ANOVA, invalidating it. Achieving balance in the hauls is more important than balance in fish for all other analyses with respect to Type II error, in that unbalance in hauls reduces the power of the hypothesis test.

Acknowledgements

This project was inspired by the seminal work on pseudoreplication by Dr. Stuart Hurlbert and we appreciate his generosity in discussing aspects of our research with us. The class “Analysis of Messy Data” that we took from George Milliken has been invaluable in many ways, especially in regards to analyzing nested data. We also appreciate the advice that Dr. Milliken has provided on many occasions over the years on specific analyses. We wish to thank Matthew Wilson, Jennifer Cahalan, Drs. Jeffrey Napp, James Meador, and Stuart Hurlbert for thorough reviews and insightful suggestions. The suggestions given by *Fisheries Research*’s reviewers, Dr. Mikhail Kozlov and an anonymous reviewer, also were helpful and appreciated. We are indebted to those responsible for collecting, measuring, and verifying the data used to design the simulations. We also thank the many researchers who asked for our expertise to help them analyze their nested data, especially those who, after we explained the perils of pseudoreplication, graciously accepted the sometimes disappointing results from the less powerful, but more accurate, correctly applied nested mixed model test. This project was motivated by their need for a practical guide to statistical methods for comparing means from two-stage sampling. This research is contribution EcoFOCI-R746 to NOAA’s Ecosystem & Fisheries-Oceanography Coordinated Investigations.

Appendices A–H. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fishres.2010.09.009.

References

Balleurka, N., Gómez, J., Hidalgo, D., 2005. The controversy over null hypothesis significance testing revisited. Methodology 1 (2), 55–70.

Bickel, R., 2007. Multilevel Analysis for Applied Research: It’s Just Regression. Guilford Press, New York.

Boruch, R., 2007. The null hypothesis is not called that for nothing: statistical tests in randomized trials. Journal of Experimental Criminology 3, 1–20.

Cochran, W.G., 1977. Sampling Techniques, 3rd edition. John Wiley and Sons, New York.

Goldstein, H., 2003. Multilevel Statistical Models, 3rd edition. Oxford University Press, New York.

Hairston Sr., N.G., 1989. Ecological Experiments: Purpose, Design, and Execution. Cambridge University Press, Cambridge.

Hedges, L.V., 2007. Correcting a significance test for clustering. Journal of Educational Behavioral Statistics 32, 151–179.

Heffner, R.A., Butler, M.J., Reilly, C.K., 1996. Pseudoreplication revisited. Ecology 77 (8), 2558–2562.

Hox, J.J., 2002. Multilevel Analysis: Techniques and Applications. Lawrence Erlbaum, Mahwah, NJ, USA.

Hurlbert, S.H., 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54 (2), 187–211.

Hurlbert, S.H., 2004. On misinterpretation of pseudoreplication and related matters: a reply to Oksanen. Oikos 104 (3), 591–597.

Hurlbert, S.H., White, M.D., 1993. Experiments with freshwater invertebrate zooplanktivores: quality of statistical analyses. Bulletin of Marine Science 53 (1), 128–153.

Institute of Educational Sciences, 2007. Technical details of WWC-conducted computations. http://ies.ed.gov/ncee/wwc/references/iDocViewer/Doc.aspx?docId=20&tocId=6 [accessed 20 January 2009].

Kendall, A.W., Schumacher, J.D., Kim, S., 1996. Walleye pollock recruitment in Shelikof Strait: applied fisheries oceanography. Fisheries Oceanography 5 (Suppl. 1), 4–18.

Kozlov, M.V., Hurlbert, S.H., 2006. Pseudoreplication, chatter, and the international nature of science: a response to D.V. Tatarnikov. Journal of Fundamental Biology 67 (2), 145–152.

Laird, N.M., Ware, J.H., 1982. Random-effects models for longitudinal data. Biometrics 38, 962–974.

Lehtonen, R., Pahkinen, E., 2004. Practical Methods for Design and Analysis of Complex Surveys. John Wiley and Sons, New York.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., Schabenberger, O., 2006. SAS for Mixed Models, 2nd edition. SAS Institute Inc., Cary, NC, USA.

Lombardi, C.M., Hurlbert, S.H., 2009. Misprescription and misuse of one-tailed tests. Australian Journal of Ecology 34, 447–468.

Millar, R.B., Anderson, M.J., 2004. Remedies for pseudoreplication. Fisheries Research 70, 397–407.

Milliken, G.A., Johnson, D.E., 1992. Analysis of Messy Data. Chapman and Hall, London.

Nickerson, R.S., 2000. Null hypothesis significance testing: a review of an old and continuing controversy. Psychological Methods 5 (2), 1051–1057.

Parkhurst, D.F., 2001. Statistical significance tests: equivalence and reverse tests should reduce misinterpretation. BioScience 51, 1051–1057.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-effects Models in S and S-PLUS. Springer-Verlag, New York.

Quinn, G.P., Keough, M.J., 2002. Experimental Design and Data Analysis for Biologists. Cambridge University Press, Cambridge.

Raudenbush, S.W., Bryk, A.S., 2002. Hierarchical Linear Models, Applications and Data Analysis Methods, 2nd edition. Sage Publications, Thousand Oaks, CA, USA.

Satterthwaite, F.E., 1946. An approximate distribution of estimates of variance components. Biometrics Bulletin 2, 110–114.

Sokal, R.R., Rohlf, F.J., 1995. Biometry, 3rd edition. W.H. Freeman and Company, New York.

Steel, R.G.D., Torrie, J.H., Dickey, D.A., 1997. Principles and Procedures of Statistics: A Biometrical Approach, 3rd edition. McGraw-Hill Series in Probability and Statistics, McGraw-Hill, New York.

Urquhart, D.J., 1981. The Principles of Librarianship. Scarecrow Press, Metuchen, NJ, USA.

Weinberg, K.L., Wilkins, M.E., Shaw, F.R., Zimmerman, M., 2002. The 2001 Pacific west coast bottom trawl survey of groundfish resources: estimates of distribution, abundance, and length and age composition (Ed. U.S.D.o. Commerce, NOAA Technical Memorandum NMFS-AFSC-128).

West, B.T., Welch, K.B., Galecki, A.T., 2007. Linear Mixed Models: A Practical Guide using Statistical Software. Chapman and Hall/CRC, Boca Raton, FL, USA.

Wilson, M.T., Mazur, M., Buchheister, A., Duffy-Anderson, J.T., 2006. Forage fishes in the western Gulf of Alaska: variation in productivity. North Pacific Research Board Final Report 308.

Ziliak, S.T., McCloskey, D.N., 2008. The Cult of Statistical Significance. University of Michigan Press, Ann Arbor, MI, USA.