SMALL SAMPLE PROPERTIES OF THE TWO INDEPENDENT SAMPLE TEST FOR MEANS FROM BETA DISTRIBUTIONS

Edward E. Gbur
Agricultural Statistics Laboratory, University of Arkansas, Fayetteville, AR 72701, egbur@uark.edu

Kevin Thompson
Agricultural Statistics Laboratory, University of Arkansas, Fayetteville, AR 72701, kthompsn@uark.edu

SMALL SAMPLE PROPERTIES OF THE TWO INDEPENDENT SAMPLE TEST FOR MEANS FROM BETA DISTRIBUTIONS

Kevin Thompson and Edward Gbur
Agricultural Statistics Laboratory
Arkansas Agricultural Experiment Station
University of Arkansas, Fayetteville AR 72701

Abstract

Researchers often collect proportion data that cannot be interpreted as arising from a set of Bernoulli trials. Analyses based on the beta distribution may be appropriate for such data. The SAS® GLIMMIX procedure provides a tool for these analyses using a likelihood based approach within the larger context of generalized linear mixed models (GLMM). The small sample behavior of likelihood based tests to compare the means from two independently sampled beta distributions were studied via simulation when the null hypothesis of equal means holds. The numerical techniques used were pseudo-likelihood and Laplace. Two simulation scenarios were defined by equal and unequal sample sizes and equal scale parameters. A third scenario was defined by equal sample sizes and unequal scale parameters. For all three scenarios the values of the common mean μ ranged from 0 and 0.5 and values of the scale parameter ϕ ranged from 0 to 100.

Keywords: beta distribution, two sample problem, GLIMMIX

1. Introduction

Proportions measured on a continuum are often modeled by a beta distribution because of the wide range of possible shapes for its pdf. It also serves as a model for any continuous distribution defined on a finite interval. Our objective here is to study the behavior of the type I error rate (α) for the two independent sample test for the equality of means from a beta distribution using PROC GLIMMIX in SAS®. This study follows the authors' study on small sample estimation in the one sample beta problem (http://newprairiepress.org/agstatconference/2013/).

2. Beta distribution

The standard textbook form of the pdf of a beta distribution for a random variable Y is given by

$$f(y \mid \alpha, \beta) = y^{(\alpha - 1)}(1 - y)^{(\beta - 1)}/B \quad \text{for } 0 < y < 1,$$

where B is the beta function, $\alpha > 0$ and $\beta > 0$. For generalized linear mixed models (GLMMs), the alternative parameterization

$$f(y \mid \mu, \phi) = y^{(\mu\phi - 1)}(1 - y)^{(\phi(\mu - 1) - 1)}/B,$$

where $\mu = \alpha/(\alpha + \beta)$ and $\phi = \alpha + \beta$, is often used. For this parameterization, $0 < \mu < 1$ and $\phi > 0$. The mean and variance of Y are given by

$$E(Y) = \mu \text{ and } Var(Y) = \mu(1 - \mu)/(1 + \phi).$$

The parameter space for the beta distribution can be divided into regions defined by α and β (or equivalently by μ and ϕ) that determine the general shape of the pdf as illustrated in Figure 1 for $0 < \mu \leq 0.5$ and $0 < \phi \leq 10$. The shapes of pdfs for $\mu > 0.5$ are mirror images of the corresponding distributions having $\mu < 0.5$. Distributions with $\mu = 0.5$ are symmetric regardless of value of ϕ and distributions with $\phi > 10$ have the same general shape as the corresponding distributions shown in Figure 1 for $\phi = 10$.

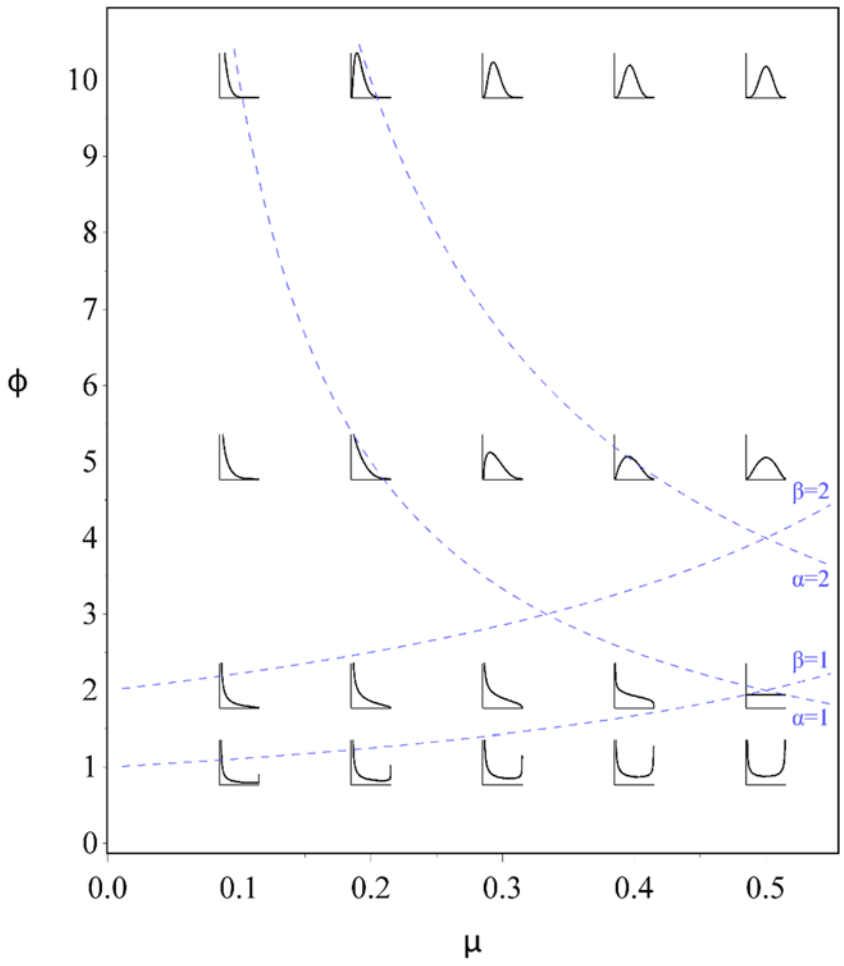

Figure 1. General shapes of the pdf for beta distributions when $0 < \mu \leq 0.5$ and $0 < \phi \leq 10$.

3. Simulation details

The simulation was conducted using SAS® 64 bit version 9.4 (TS1M2, Analytics version 13.2).

Pairs of samples were generated independently under the null hypothesis H_0: $\mu_1 = \mu_2$ for parameters from a rectangular grid on (μ, ϕ) parameter space with

$\mu = 0.1, 0.2, 0.3, 0.4, 0.5$

and

$\phi = 1, 2, 5, 10, 50, 100$.

The sample sizes included in the simulation for each population were N = 5, 10, 15, 20, 25, 50, and 100. For each (μ, ϕ, N) combination 1000 samples were generated using the beta function RAND('BETA') in SAS®.

Pairs of samples from the two populations with the same mean were analyzed as a generalized linear model (GLM) using PROC GLIMMIX with a logit link function and beta distribution having the (μ, ϕ) parameterization. The MODEL statement contained one factor with two levels (treatments) and the SOLUTION, LINK and DIST options. The LSMEANS statement with the ILINK option was used to obtain estimates of μ_1 and μ_2 and their estimated standard errors. The numerical techniques used were pseudo-likelihood (RSPL) and Laplace.

The p-values for the type 3 F-test of H_0: $\mu_1 = \mu_2$ were used to estimate type I error rates from the simulation for $\alpha = 0.01$, 0.05 and 0.10. For each α, the estimate $\hat{\alpha}$ was the proportion of tests with p-values less than or equal to the nominal α. If we assume that $\hat{\alpha}$ follows a binomial distribution with $E(\hat{\alpha}) = \alpha = 0.05$, then the standard error of $\hat{\alpha}$ would be approximately 0.007 and an approximate Wald 95% confidence interval for α under H_0 would be (0.036, 0.064).

Only results for $\alpha = 0.05$ and pseudo-likelihood are reported here. Conclusions for $\alpha = 0.01$ and 0.10 and for Laplace were generally similar.

4. Convergence issues

Convergence was assumed when GLIMMIX indicated convergence and either all parameter estimates and standard errors were given or all estimates except SE($\hat{\phi}$) were given. At least 99% of the 1000 sample pairs when $\phi_1 \geq 5$ regardless of the values of ϕ_2, μ, n_1, and n_2 converged under this definition. In addition, convergence was assumed for at least 90% of the sample pairs when $\phi_1 \geq 1$ or when $\mu \geq 0.2$ regardless of the values of the remaining parameters in either case. For $\mu = 0.1$ and $\phi_1 = 0.5$, there were 68 of the 1,715 parameter combinations (4%) where less than 90% of the sample pairs were assumed to converge. The worst case scenario was for $\phi_1 = 0.5$, $\phi_2 = 1$, $\mu = 0.1$, and $n_1 = n_2 = 5$ where the convergence percentage was 76%.

Among samples that converged under the above definition there were pairs of samples for which either

the F value was infinite (and the p-value exactly zero)

or

the Fit Statistics table contained large values such as 2×10^{20} for -2 log likelihood, AIC, AICC, and the other information criteria

or
both.

These problematic samples were included in the results that follow.

5. Equal N and equal ϕ

For equal sample sizes N and a common value of $\phi \leq 5$ for both populations, Table 1 contains the estimated type I error rates $\hat{\alpha}$ under H_0: $\mu_1 = \mu_2$ when $\alpha = 0.05$. Values of $\hat{\alpha}$ in the table that are outside the approximate 95% confidence interval (0.036, 0.064) for α are highlighted in red. In general the estimates for $\phi = 10, 50, 100$ were similar to those for $\phi = 5$.

Table 1. Estimated type I error rates $\hat{\alpha}$ for selected values of the common mean and the scale parameter ϕ.

		$\mu_1 = \mu_2$				
ϕ	N	0.1	0.2	0.3	0.4	0.5
5	5	0.065	0.057	0.066	0.052	0.069
	10	0.040	0.057	0.053	0.047	0.041
	15	0.044	0.052	0.045	0.066	0.059
	20	0.050	0.072	0.060	0.060	0.058
	25	0.057	0.059	0.054	0.053	0.053
	50	0.053	0.062	0.044	0.059	0.061
	100	0.058	0.057	0.062	0.045	0.054
2	5	0.061	0.068	0.081	0.062	0.071
	10	0.059	0.065	0.050	0.054	0.072
	15	0.058	0.068	0.047	0.051	0.063
	20	0.080	0.053	0.060	0.039	0.061
	25	0.072	0.068	0.056	0.050	0.054
	50	0.098	0.050	0.050	0.061	0.040
	100	0.114	0.055	0.052	0.062	0.046
1	5	0.146	0.057	0.055	0.072	0.071
	10	0.184	0.063	0.066	0.047	0.063

Table 1. Estimated type I error rates $\hat{\alpha}$ for selected values of the common mean and the scale parameter ϕ.

		$\mu_1 = \mu_2$				
ϕ	N	0.1	0.2	0.3	0.4	0.5
	15	0.219	0.051	0.050	0.061	0.061
	20	0.221	0.062	0.047	0.050	0.064
	25	0.231	0.061	0.046	0.053	0.049
	50	0.231	0.073	0.063	0.043	0.047
	100	0.267	0.066	0.046	0.048	0.055

For small μ and ϕ (close to the parameter space boundary), the estimated type I error rate becomes much greater than the nominal 0.05 level unexpectedly as the sample size increases. This corresponds to the region of the parameter space where the number of problematic samples described previously in Section 4 becomes more common with increasing sample size. When these problematic samples are removed from the analysis the estimated type I error rates become dramatically smaller. For example, for $\phi_1 = 1$ and $\mu = 0.1$, $\hat{\alpha}$ is approximately 0.083 when $n = 5$ and is never larger than 0.068 for any $n > 5$. For $\phi_1 = 1$ and $\mu = 0.1$, $\hat{\alpha}$ ranges from approximately 0.042 to 0.065. Unfortunately the number of convergent samples is reduced and the number of problematic samples with that subset increases. making the results less clear cut. As both ϕ and μ move away from the boundary of the parameter space, the estimated type I error rate becomes relatively close to the nominal level as the sample size increases.

6. Unequal N and equal ϕ

Type I error rates were estimated for all combinations of N_1 and N_2 and each combination of common mean μ and scale parameter ϕ. For the common value of μ equal to 0.1 and each $\phi \leq 5$ for both populations, Table 2 contains the estimated type I error rates $\hat{\alpha}$ under H_0: $\mu_1 = \mu_2$ when $\alpha = 0.05$ for all combinations of sample sizes N_1 and N_2. For $\phi > 5$, in general the nominal α level falls within the confidence interval based on the estimated type I error rate. For a given $\phi > 5$ there is no apparent pattern to the sample size combinations where the error rate is over-estimated.

For $\mu = 0.1$ as ϕ decreases toward its lower limit, the estimated type I error rate tends to exceed the nominal level for fixed N_1 and N_2. For fixed N_1, as the difference between N_1 and N_2 increases the estimated type I error rate tends to increase for small values of ϕ. As μ moves away from the boundary, the estimated error rate tends to be within sampling variation of the nominal level α regardless of the value of ϕ (data not shown).

Table 2. Estimated type I error rates $\hat{\alpha}$ for the common mean μ equal to 0.1 and selected values of ϕ when the sample sizes differ.

		ϕ		
N1	N2	5	2	1
5	5	0.065	0.061	0.146
5	10	0.068	0.051	0.179
5	15	0.065	0.063	0.188
5	20	0.049	0.055	0.244
5	25	0.068	0.057	0.270
5	50	0.043	0.099	0.347
5	100	0.061	0.087	0.419
10	10	0.040	0.059	0.184
10	15	0.065	0.060	0.179
10	20	0.044	0.064	0.195
10	25	0.057	0.054	0.209
10	50	0.059	0.085	0.288
10	100	0.060	0.101	0.380
15	15	0.044	0.058	0.219
15	20	0.052	0.074	0.185
15	25	0.042	0.072	0.234
15	50	0.057	0.092	0.261
15	100	0.066	0.122	0.327
20	20	0.050	0.080	0.221
20	25	0.055	0.079	0.225
20	50	0.054	0.079	0.241
20	100	0.046	0.106	0.287
25	25	0.057	0.072	0.231
25	50	0.054	0.069	0.217

Table 2. Estimated type I error rates $\hat{\alpha}$ for the common mean μ equal to 0.1 and selected values of ϕ when the sample sizes differ.

		ϕ		
N1	**N2**	**5**	**2**	**1**
25	**100**	**0.054**	**0.123**	**0.281**
50	**50**	**0.053**	**0.098**	**0.231**
50	**100**	**0.057**	**0.109**	**0.223**
100	**100**	**0.058**	**0.114**	**0.267**

7. Equal N and unequal ϕ

The current version of PROC GLIMMIX assumes that the scale parameter ϕ is the same for all treatments. To examine the impact of this assumption on the type I error rate, each value of ϕ_1 was paired with each possible value of ϕ_2 and the simulation was rerun for each common value of the mean μ and common sample size N. The results are shown in Figure 2 for $\phi_1 = 5$.

The patterns displayed in the figure are typical of the relationships between the estimated type I error rate and sample size for all (ϕ_1, ϕ_2) combinations. For a fixed value of the common mean μ, the estimated error rate increases as a function of the common sample size N but at a decreasing rate of increase as μ approaches 0.5; i.e., as the pdf approaches symmetry about 0.5. For "small" values of μ, the type I error rate is grossly overestimated regardless of the values of the ϕ's. The magnitude of the overestimation tends to decrease as μ approaches 0.5 which is the point at which beta distributions are symmetric.

8. Conclusions

The behavior of the type I error rate in the two sample problem for the means of a beta distribution under the null hypothesis H_0: $\mu_1 = \mu_2$ has been studied herein via simulation using PROC GLIMMIX in SAS®. The most important finding is that under the assumption of a common scale parameter ϕ that is near the zero boundary of the parameter space, the estimated type I error rate becomes significantly larger than the nominal α level as the value of the common mean μ approaches zero. This behavior is exhibited for both equal and unequal sample sizes even for individual sample sizes as large as 100. Given the symmetry of the beta pdf about $\mu = 0.5$, the same behavior would be expected as μ approaches one.

For values of μ and ϕ "close" to zero, the beta distribution is U-shaped with only a very small probability in the upper tail. For beta distributions with μ "close" to zero the median is much that

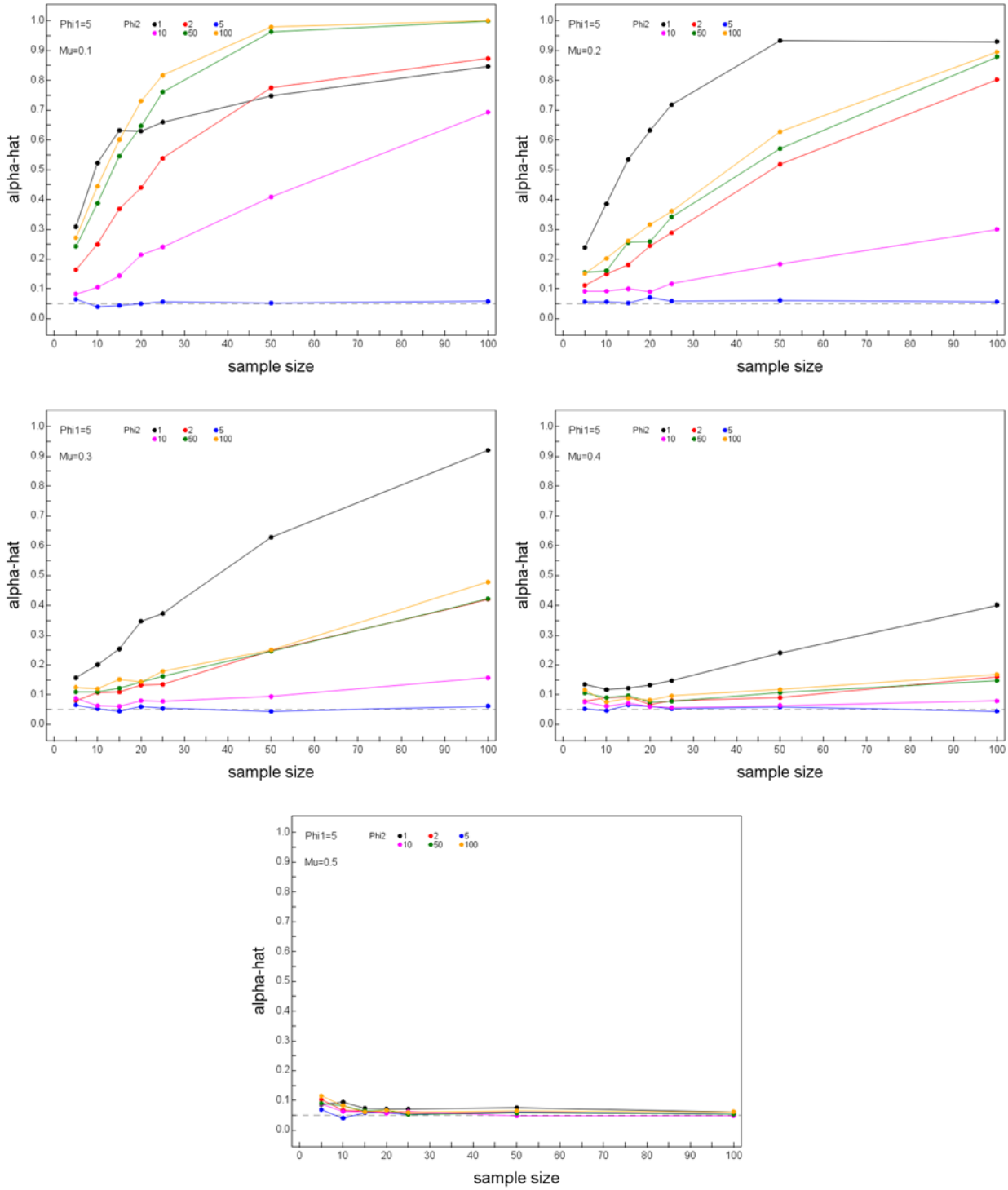

Figure 2. Estimated type I error rate $\hat{\alpha}$ as a function of the common sample size N for $\phi_1 = 5$ and each (μ, ϕ_2) combination.

smaller than the mean leading to the potential for samples to contain individual observations that are very small. In addition, the convergence/computational issues described in Section 4 are most common in this region of the parameter space. From our simulation as presented here we were not able to disentangle the role of each of these two factors in the over-estimation of the type I error rates.

The GLIMMIX procedure assumes that the scale parameter for beta distributions (and all other distributions that the procedure allows) is the same for all groups in the model. The somewhat unexpected gross over-estimation of the type I error rate that was found in Section 7 when the scale parameters were not equal should serve as a warning to GLIMMIX users in situations where this assumption may not hold based on histograms or boxplots of the data or on subject matter considerations.