

# A Two-Way Analysis of Covariance Model for Classification of Stability Data\*

Hongshik Ahn,<sup>1</sup> James J. Chen,<sup>2</sup> and Tsae-Yun D. Lin<sup>3</sup>

<sup>1</sup>Department of Applied Mathematics and Statistics  
State University of New York at Stony Brook  
Stony Brook, NY 11794 - 3600  
U.S.A.

<sup>2</sup>Division of Biometry and Risk Assessment  
National Center for Toxicological Research  
Food and Drug Administration  
Jefferson, AR 72079  
U.S.A.

<sup>3</sup>Division of Biometrics IV  
Center for Drug Evaluation and Research  
Food and Drug Administration  
Rockville, MD 20857  
U.S.A.

## Abstract

This paper proposes a procedure for testing and classifying stability data with multiple factors. A two-way analysis of covariance is used to classify the differences among the batches as well as another factor such as package type and/or product strength. In the test procedure, slopes and intercepts of the main effects are tested using a combination of simultaneous and sequential  $F$ -tests. Based on the test procedure results, the data are classified into one of four different groups. For each group, shelf life can be calculated accordingly. We examine if the procedure produces satisfactory control of the probability of a Type I error and the power of detecting the difference of degradation rates and intercepts for different nominal levels. The method is evaluated with a Monte Carlo simulation study. The proposed procedure is compared with the current FDA procedure using real data.

Key Words: batch; classification; expiration date; Monte Carlo; shelf life.

---

\*The views expressed in this paper are those of authors and not necessarily of the Food and Drug Administration.

# 1 Introduction

The U.S. Food and Drug Administration (FDA) requires that the expiration dating period (shelf life) must be indicated on the immediate container label for every drug product and biologic in the market (FDA, 1987). The expiration dating period of a drug product is defined as the time interval that the average drug characteristic over a particular batch of the drug product is expected to remain within the approved specifications after manufacture. The intent of a long-term stability study is to insure that the identity, strength, quality and purity of a drug product are within established specification ranges prior to the expiration date. The FDA and pharmaceutical companies conduct stability analyses to characterize the degradation of the drug product by testing various batches of the product at several time points. Currently, a one-way analysis of covariance (ANCOVA) model is used to analyze stability data. The model recommended is given by

$$y_{ijk} = \alpha_i + \beta_i x_{ij} + \epsilon_{ijk},$$

where the errors  $\epsilon_{ijk}$  are independent  $N(0, \sigma^2)$ ,  $y_{ijk}$  is the test result from the  $k$ th replicate in the  $j$ th time point of the  $i$ th batch,  $x_{ij}$  is the time of the stability sample corresponding to  $y_{ijk}$ ,  $\alpha_i$  is the batch effect at time 0 and  $\beta_i$  is the degradation rate of the  $i$ th batch. The FDA requires that at least three batches should be tested to allow for some estimates of batch-to-batch variability.

The following two hypothesis tests are performed to determine the pooling of batch data.

1. Testing of equality of slopes for a model with separate intercepts and separate slopes.
2. Testing of equality of intercepts given parallel lines.

The level of significance recommended by the FDA for the tests is .25, which is based on Bancroft's (1964) result on preliminary tests (Lin, Lin and Kelly, 1993). Both of the above tests should be accepted to pool the data. If the data are determined to be pooled, then the shelf life is estimated as the time at which the 95% one-sided lower (upper) confidence bound for the mean degradation curve intersects the acceptable lower (upper) specification limit. In this paper, we assume the degradation product decreases with time. The lowest acceptable limit for drug content is usually 90% of the labeled amount. If the first test is accepted and the second test is rejected, then the data may be combined for the purpose of estimating the common slope. If both of the above tests result in rejection, then the data from different batches would be separated. If at least one of the above tests is rejected, the FDA suggests computing the shelf life of each individual

batch and using the minimum of all the shelf lives as the estimated shelf life.

Note that if the hypothesis of equal slope is rejected, the FDA procedure uses different variance estimates for different batch models. Rejection of the hypothesis of equal slopes does not imply unequal variances among batches. The procedure described in this paper focuses on the extension of the current FDA procedure. The procedure can be modified under the model of a equal batch variance.

Chow and Shao (1989) proposed several test procedures for batch-to-batch variation of a drug product using a random effects model. Chow and Shao (1991) proposed a linear random effects model to estimate a confidence limit or shelf life on the predicted response of a drug for marketing stability analysis using the weighted least squares method. Chen et al. (1995) presented a mixed effects model using an EM algorithm procedure to obtain the maximum likelihood estimates of the fixed effects and random effects regression coefficients and variance components. Ruberg and Stegeman (1991) discussed the problem of the significance level suggested by the FDA for hypothesis testing to determine the pooling of batch data.

The current FDA guideline (1987) considers analysis of stability data for a single drug product within a single type of package. The FDA currently does not have an official procedure for multi-factor analysis of stability data. For estimation of drug shelf life with different types of packages or different strengths, batch-to-batch variation is currently tested separately at each package and strength. This method may require many unnecessary tests for packages or strengths with similar degradation rates or intercepts. In this paper, we propose a procedure which can test the difference of the factors as well as the batch-to-batch variation. This can be done by conducting a combination of simultaneous and sequential  $F$ -tests in a two-way ANCOVA. This procedure is a generalization of the current FDA procedure which considers only a single type of package or strength. The proposed procedure classifies stability data into one of four different groups.

In this paper, we assume drug products with several types of packages in various batches. In this case, two-way ANCOVA is more appropriate than one-way ANCOVA for analyzing the stability data. In Section 2, we propose a test procedure for the two-way ANCOVA model. In Section 3, we conduct a simulation study to evaluate the proposed procedure. Section 4 contains an example with real data. Conclusions of the study are given in Section 5.

## 2 Test Procedure

### 2.1 Test procedure for a two-way model

Assuming that  $I$  batches and  $J$  packages are considered in the stability analysis, we consider the following analysis of covariance model:

$$y_{ijkl} = \alpha_{ij} + \beta_{ij}x_{ijk} + \epsilon_{ijkl},$$

where

$y_{ijkl}$  = response from the  $l$ th replicate in the  $k$ th time point of batch  $i$  and package  $j$ ,

$\alpha_{ij}$  = intercept of the  $i$ th batch and  $j$ th package,

$\beta_{ij}$  = degradation rate of the  $i$ th batch and  $j$ th package,

$x_{ijk}$  = time of the stability sample corresponding to  $y_{ijkl}$ ,

$\epsilon_{ijkl}$  = random error corresponding to  $y_{ijkl}$

for  $i = 1, \dots, I$ ;  $j = 1, \dots, J$ ;  $k = 1, \dots, K$  and  $l = 1, \dots, n$ . The random error  $\epsilon_{ijkl}$  is assumed to be independently and normally distributed with mean zero and common variance  $\sigma^2$ . In our simulation study, we assume no replication, i.e.,  $n = 1$ .

In the proposed test procedure, the data are classified into four groups. The classification is according to whether the data are pooled or separated in estimating shelf life. In a given group, calculation of the shelf life may be slightly different among the final models. The test procedure is as follows:

**Step 1:** We test both of the following hypotheses:

(a) Test for equality of slopes for batches:

$$H_{01} : y_{ijk} = \alpha_{ij} + \beta_j x_{ijk} + \epsilon_{ijk} \tag{M1}$$

versus

$$H_{a1} : y_{ijk} = \alpha_{ij} + \beta_{ij} x_{ijk} + \epsilon_{ijk}. \tag{M0}$$

(b) Test for equality of slopes for packages:

$$H_{02} : y_{ijk} = \alpha_{ij} + \beta_i x_{ijk} + \epsilon_{ijk} \quad (M2)$$

versus  $H_{a1} : (M0)$ .

The data can be classified into one of four classes in accordance with the slope of the degradation line. The data are classified as follows:

Class 0: The null hypotheses are rejected in both of the above tests.

Class 1: The null hypotheses are rejected in (a) and accepted in (b).

Class 2: The null hypotheses are accepted in (a) and rejected in (b).

Class 3: The null hypotheses are accepted in both of the above tests.

Alternatively, the data can be classified into one of four groups according to the intercept of the degradation line. One more step of tests is necessary for this grouping. If both of the above null hypotheses are rejected, then separate slopes and separate intercepts are estimated for both batches and packages. The data are classified into Group 0 in this case. If one of the above hypotheses is accepted and the other hypothesis is rejected, then go to Step 2-1. If both of the hypotheses are accepted, then go to Step 2-2.

**Step 2-1:**

(a) If  $H_{01}$  is accepted and  $H_{02}$  is rejected, then perform the following test of equality of intercepts for batches given equal batch slopes:

$$H_{03} : y_{ijk} = \alpha_j + \beta_j x_{ijk} + \epsilon_{ijk} \quad (M3)$$

versus  $H_{a3} : (M1)$ .

If  $H_{03}$  is rejected, then we conclude that the slopes are the same for batches, but the intercepts are different for both factors (Group 0). If  $H_{03}$  is accepted, we conclude that there is no batch-to-batch variation, but there is package-to-package variation (Group 1).

(b) If  $H_{01}$  is rejected and  $H_{02}$  is accepted in Step 1, then perform the following test of

equality of intercepts for packages given equal package slopes:

$$H_{04} : y_{ijk} = \alpha_i + \beta_i x_{ijk} + \epsilon_{ijk} \quad (M4)$$

versus  $H_{a4} : (M2)$ .

If  $H_{04}$  is rejected, then use separate intercepts for both batches and packages, but use the same slope for packages (Group 0). If  $H_{04}$  is accepted, we conclude that there is no package-to-package variation, but there exists batch-to-batch variation (Group 2).

**Step 2-2:** In the case that both  $H_{01}$  and  $H_{02}$  are accepted in Step 1, the following hypotheses are tested.

(a) Test for equality of intercepts for batches:

$$H_{05} : y_{ijk} = \alpha_j + \beta x_{ijk} + \epsilon_{ijk} \quad (M6)$$

versus

$$H_{a5} : y_{ijk} = \alpha_{ij} + \beta x_{ijk} + \epsilon_{ijk}. \quad (M5)$$

(b) Test for equality of intercepts for packages:

$$H_{06} : y_{ijk} = \alpha_i + \beta x_{ijk} + \epsilon_{ijk} \quad (M7)$$

versus  $H_{a5} : (M5)$ .

If both of the above hypotheses are rejected, then use separate intercept, but use the same slope for both batches and packages (Group 0). If  $H_{06}$  is accepted and  $H_{07}$  is rejected, then use the same slope for both factors, and same intercept for batches (Group 1). If  $H_{06}$  is rejected and  $H_{07}$  is accepted, then use the same slope for both factors, and same intercept for packages (Group 2). If both of the above hypotheses are accepted, then all the batches and packages have the same shelf life (Group 3), and the following model will be considered.

$$y_{ijk} = \alpha + \beta x_{ijk} + \epsilon_{ijk}. \quad (M8)$$

As a summary, the above test procedure produces the following classes and groups:

1. According to the slopes (determined in Step 1),

Class 0: different slopes for different batch, package combinations,

Class 1: a common slope among batches within each package,

Class 2: a common slope among packages within each batch,

Class 3: a common slope in all batches and packages.

2. According to the intercepts (determined in Steps 1 and 2),

Group 0: different intercepts for different batch/package combinations,

Group 1: in Classes 1 and 3, a common intercept among batches within each package,

Group 2: in Classes 2 and 3, a common intercept among packages within each batch,

Group 3: given in Class 3, a common intercept in all batches and packages.

Figures 1 and 2 describe the classification scheme. Table 1 shows how the models are classified. The proposed test procedure can be generalized for more than two factors.

Of the 16 possible models in the given situation, 9 models were taken into consideration as the representative models. The other seven models are as follows:

$$y_{ijk} = \alpha_j + \beta_{ij}x_{ijk} + \epsilon_{ijk}, \quad (M0a)$$

$$y_{ijk} = \alpha_i + \beta_{ij}x_{ijk} + \epsilon_{ijk}, \quad (M0b)$$

$$y_{ijk} = \alpha + \beta_{ij}x_{ijk} + \epsilon_{ijk}, \quad (M0c)$$

$$y_{ijk} = \alpha_i + \beta_jx_{ijk} + \epsilon_{ijk}, \quad (M1a)$$

$$y_{ijk} = \alpha_j + \beta_ix_{ijk} + \epsilon_{ijk}, \quad (M2a)$$

$$y_{ijk} = \alpha + \beta_jx_{ijk} + \epsilon_{ijk}, \quad (M3a)$$

$$y_{ijk} = \alpha + \beta_ix_{ijk} + \epsilon_{ijk}. \quad (M4a)$$

The expiration dating period for the above seven models can be computed as that for the representative models. For example, the self life for  $M0a$ ,  $M0b$  and  $M0c$  can be calculated as that for  $M0$ , and the self life for  $MKa$  can be calculated as that for  $MK$ ,  $K = 1, \dots, 4$ . We will show how to estimate the shelf life for a given group of models in Section 2.3.

## 2.2 Special case

In practice, package usually does not physically impact the intercept. If we assume that the intercept of the degradation line is constant for all the packages from the same batch, then model (M0) becomes

$$y_{ijk} = \alpha_i + \beta_{ij}x_{ijk} + \epsilon_{ijk}. \quad (AM0)$$

In this case, we can perform the same test procedure as we proposed in Section 2.1.

In Step 1 of the test procedure, (M1) becomes

$$y_{ijk} = \alpha_i + \beta_j x_{ijk} + \epsilon_{ijk}, \quad (AM1)$$

and (M2) becomes (M4). In this step, we test (AM1) versus (AM0) and (M4) versus (AM0) simultaneously. In Step 2-1, (M3) becomes

$$y_{ijk} = \alpha + \beta_j x_{ijk} + \epsilon_{ijk}. \quad (AM3)$$

Thus, we test (AM3) versus (AM1) in (a). We do not need (b) in this step. In Step 2-2, the simultaneous test reduces to one test of (M8) versus (M7).

## 2.3 Estimation of shelf life

Shelf-life estimation investigated in Chen et al. (1997) can be re-written for the four groups discussed in this paper.

If the data fall into Group 0, the shelf lives are computed separately for different packages. For each package, the shelf life of each individual batch is computed and the minimum of all the shelf lives is used as the shelf life for that package. See Figure 3 (a).

If the data are in Group 1, the shelf lives for the packages are computed separately. The shelf life is computed after pooling the batches in each package. The shelf life of a package is calculated using a common slope and common intercept model for all the batches (Figure 3 (b)).

If the data are in Group 2, the data of the different packages are pooled within each batch. The shelf life is computed in each individual batch using a common slope and common intercept model for the different packages. The minimum of all the batch shelf lives is used as the common shelf life for all the packages (Figure 3 (c)).

If the data fall into Group 3, the data of all the packages and all the batches are pooled. The

shelf life is obtained using a common slope and common intercept model ( $M8$ ) for the whole data. See Figure 3 (d).

The ways of computing the shelf lives in Groups 0 and 1 are the same as those in the FDA procedure. If the equal slope and equal intercept hypotheses are tested using the combined data of all the packages, then the FDA estimation of the shelf life would be the same as ours for the data in Group 3. However, if the testing is conducted separately in each package, which is more common, then the FDA estimation is different from ours. If the data fall into Group 2, the proposed estimation procedure is substantially different from that recommended by the FDA. Because the FDA procedure uses one-way ANCOVA, usually the difference of the batches is tested separately in each package. Therefore, the shelf lives are estimated separately in each package, even though there is no difference of the degradation curves and intercepts among the packages. If all the packages and batches are combined (even though it is unusual) and one-way ANCOVA is used for testing, the FDA estimation of the shelf life is still different from ours. If we follow the FDA guideline, it is possible that the batches are pooled in some packages and the batches are separated in the other packages for the data in Groups 0 and 1.

### 3 Simulation Study

A Monte Carlo simulation study was conducted to evaluate the Type I error rate and power of the proposed test procedure. A design with eight time points with three batches and three packages was considered. For all the models discussed in Section 2, the values of time points  $x_{ij1}, \dots, x_{ij8}$  are 0, 3, 6, 9, 12, 18, 24 and 36 months, respectively.

First, we examined the performance of the individual  $F$ -tests for slopes and intercepts for both batches and packages. Here, we included some tests which are not included in the proposed test procedure. The probability of a Type I error was close to the nominal level. However, we do not report these results because the properties of the  $F$ -test are well-known. Second, we examined how well the proposed procedure identifies the models ( $M0 - M8$ ) and the groups (Group 0 - Group 3). The power is dependent on the variance of the errors and the number of time points in the model. However, the simulation was conducted as a preliminary study because these  $F$ -tests are a part of our model classification.

Two simulation experiments are reported in this paper. The distribution of the random error was chosen as  $\epsilon \sim N(0, 1)$  in Experiment 1, and as  $\epsilon \sim N(0, 2^2)$  in Experiment 2. The parameter

values were the same in both of the experiments. Table 2 shows the parameter values of the nine models chosen in this simulation. The values the parameters are close to those in typical stability data. For the common intercept model ( $M8$ ),  $\alpha = 101$  was chosen. For the models with different intercepts for either batches or packages ( $M3, M4, M6$  and  $M7$ ), the intercepts were 102, 101 and 100. For the models with different intercepts for both batches and packages ( $M0, M1, M2$  and  $M5$ ), the intercepts ranged from 100 to 102. For common slope models ( $M5 - M8$ ),  $\beta = -.16$  was chosen. For the models with different slopes for either batches or packages ( $M1 - M4$ ), the slopes were  $-.16, -.24$  and  $-.08$ . For the models with different slopes for both batches and packages ( $M0$ ), the slopes ranged from  $-.24$  to  $-.08$ . Four different nominal significance levels (.05, .1, .2 and .25) were tried for the  $F$ -tests.

### 3.1 Individual $F$ -tests

Table 3 shows the power of the tests listed in Section 2. The table does not include the tests of  $M2$  versus  $M0$ ,  $M4$  versus  $M2$ , and  $M7$  versus  $M5$ , because they are symmetric with  $M1$  versus  $M0$ ,  $M3$  versus  $M1$ , and  $M6$  versus  $M5$ , respectively, in our simulation. Instead,  $M5$  versus  $M1$  ( $y_{ijk} = \alpha_{ij} + \beta x_{ijk} + \epsilon_{ijk}$  versus  $y_{ijk} = \alpha_{ij} + \beta_i x_{ijk} + \epsilon_{ijk}$ ),  $M6$  versus  $M3$  ( $y_{ijk} = \alpha_i + \beta x_{ijk} + \epsilon_{ijk}$  versus  $y_{ijk} = \alpha_i + \beta_i x_{ijk} + \epsilon_{ijk}$ ) and  $M8$  versus  $M6$  ( $y_{ijk} = \alpha + \beta x_{ijk} + \epsilon_{ijk}$  versus  $y_{ijk} = \alpha_i + \beta x_{ijk} + \epsilon_{ijk}$ ) are added in the table. Ten thousand samples were generated for each test.

The power was large for all the  $\alpha$  values in Experiment 1. However, it was low for several tests when  $\alpha = .05$  and  $\alpha = .1$  in Experiment 2. The power with  $\alpha = .25$  was satisfactory in Experiment 2.

### 3.2 Model identification and classification

Table 4 and Table 5 show how well the proposed procedure identified the models. For each of the models  $M0, M1, M3, M5, M6$  and  $M8$ , Ten thousand samples were generated. The right choices are indicated in bold face. In Experiment 1, the models were identified better with lower values of  $\alpha$  than with the bigger values, except for  $M0$ . The tests identified  $M0$  pretty well for all the  $\alpha$  values. In Experiment 2, accuracy of the model identification was lower than that of Experiment 1. This is due to the bigger variance of the random error. For  $M0, M1$  and  $M5$ , accuracy of the model identification was higher as the value of  $\alpha$  grew. For  $M6$  and  $M8$ , accuracy was lower as the value of  $\alpha$  grew. Note that the blocks of four numbers for different real models, for example  $M1$  and  $M3$  in the chosen model  $M0$  column, that are all exactly equal to each other in Table 4. The

initial random seed was chosen to be the same for each real model for comparing the performances of proposed test procedure among the different models.

Table 6 and Table 7 show how well the procedure classified the four groups of models (Group 0 - Group 3). The results are according to the results in Tables 4 and 5. (See also Table 1.) For  $M0$ , there was no classification error in Experiment 1. The classification error was lower for higher values of  $\alpha$  for  $M1$  and  $M5$  in both of the experiments and for  $M0$  in Experiment 2. For  $M0, M1$  and  $M5$ , the accuracy of classification was satisfactory for the four  $\alpha$  values in Experiment 1. The classification error was lower for lower values of  $\alpha$  for  $M3$  and  $M8$  in both of the experiments and for  $M6$  in Experiment 1. In both of the experiments, using higher  $\alpha$  values gave less chance of pooling the data in the case of heterogeneity within a factor.

Classification is more meaningful than model identification because the shelf-life estimation is according to the classification. Misidentification within a group does not affect the final shelf-life estimation. Tables 4 and 5 are provided only for illustration.

## 4 Example

We use the data of Shao and Chow (1994) to illustrate our procedure. The stability study was conducted on a 300 mg tablet of a drug product to establish appropriate shelf life. The data consist of five batches in two types of packages (bottle and blister). The tablets were tested for potency at 0, 3, 6, 9, 12 and 18 months.

We compare the FDA test procedure and the proposed procedure. The nominal significance level was chosen to be 0.25 for both procedures. First, we conducted the FDA test procedure for each package. In both packages, the tests indicated that the slopes were different by batch. According to the FDA guideline, the shelf lives will be estimated separately for the five batches and the minimum of the shelf lives will be taken for each package. Table 8 shows the ANCOVA table for the separate packages. Second, we examined the test procedure proposed in this paper. The model considered in this paper is

$$y_{ijk} = \alpha_{ij} + \beta_{ij}x_{ijk} + \epsilon_{ijk},$$

where  $i = 1, \dots, 5$ ,  $j = 1, 2$ ; and  $k = 1, \dots, 6$ . In Step 1 of the test procedure, Test (a) rejected  $H_{01}$  ( $P$ -value = .0027) and accepted  $H_{02}$  ( $P$ -value = .35). In Step 2-1, Test (b) accepted  $H_{04}$  ( $P$ -value is .88). Therefore, the final model for the data is  $M4$  and it fell into Group 2. As outlined in Section

2, the two packages can be pooled and the shelf life can be computed in each batch. According to the test results, we analyzed the data with packages combined. Table 9 shows summary results of the test procedure for the data. Table 10 shows the ANCOVA table. Note that the  $P$ -values in this table are not the same as those in the proposed test procedure. The ANCOVA table is obtained after combining the packages. The minimum of the five shelf lives will be the estimated shelf life. The estimated shelf life would be higher than both of the two estimates from the FDA procedure. The estimated shelf life could rise by pooling the containers with similar degradation rates, because the confidence interval becomes narrower for a bigger sample. Chow and Shao (1991) criticized that the minimum shelf life approach in the FDA guideline is too conservative for future production batches. The proposed procedure may reduce the conservatism of the shelf life estimation.

## 5 Conclusions

We proposed the procedure of testing in a two-way ANCOVA for stability data. A combination of simultaneous and sequential hypothesis testing was introduced and examined. The test procedure provides a quick examination of nonhomogeneity within a factor. As we observed in Section 4.1, the power was satisfactory for all the tests for  $\alpha = .25$ , but significantly lower for  $\alpha = .1$  and  $\alpha = .05$  for some tests in Experiment 2. The procedure classified the data quite well for  $M3$  and  $M8$  when  $\alpha = .05$  and  $\alpha = .1$ , and for  $M0$  when  $\alpha = .25$ . This procedure can be generalized for stability data with more than two factors. The power depends on the number of time points, packages, and batches. However, the simulation designs considered in this paper are typical drug stability designs. Therefore, our simulation results are meaningful. Future study effort will be devoted to find optimal designs by examining the designs with missing time points or other factors.

Note that the proposed test procedure is not aimed for

$$H_0 : \text{The model belongs to group } i$$

versus

$$H_a : \text{The model does not belong to group } i$$

or vice versa. The classification is simply a result from a combination of simultaneous and sequential hypothesis tests.

Recently, Fairweather et al. (1995) also proposed an extension of the FDA Guideline to two-way

setting. Their classification scheme begins with the full model and first test for the significance of the interaction. Then the significance of the main effects are examined.

The entire procedure is coded in one FORTRAN program so that the classification and model specification can be done at once. The FORTRAN program is obtainable from the authors.

## References

- Bancroft, T. A. (1964). "Analysis and inference for incompletely specified models involving the use of preliminary test(s) of significance," *Biometrics*, **20**, 427-442.
- Chen, J. J., Ahn, H. and Tsong, Y. (1997). "Shelf-life estimation for multifactor stability studies," *Drug Information Journal*, **31**, 573-587.
- Chen, J. J., Hwang, J.-S. and Tsong, Y. (1995). "Estimation of the shelf-life of drugs with mixed effects models," *Journal of Biopharmaceutical Statistics*, **5**, 131-140.
- Chow, S.-C. and Shao, J. (1989). "Test for batch-to-batch variation in stability analysis," *Statistics in Medicine*, **8**, 883-890.
- Chow, S.-C. and Shao, J. (1991). "Estimating drug shelf-life with random batches," *Biometrics*, **47**, 1071-1079.
- Fairweather, W. R., Lin, T. D. and Roswitha, K. (1995). "Regulatory, design, analysis aspects of complex stability studies," *Journal of Pharmaceutical Sciences*, **84**, 1322-1326.
- Food and Drug Administration, U.S. Department of Health and Human Services (1987). "Guideline for submitting documentation for the stability of human drugs and biologics," FDA, Rockville, Maryland.
- Lin, K. K., Lin, T. D. and Kelly, R. E. (1993), Stability of drugs. In: *statistics in the Pharmaceutical Industry*, 2nd ed. (Buncher, C. R., Tsay, J. Y., Eds.). Marcel Dekker, New York, pp. 419-444.
- Ruberg, S. J. and Stegeman, J. W. (1991). "Pooling data for stability studies: Testing the equality of batch degradation slopes", *Biometrics*, **47**, 1059-1069.
- Shao, J. and Chow, S.-C. (1994). "Statistical inference in stability analysis," *Biometrics*, **50**, 753-765.

Hongshik Ahn  
Department of Applied Mathematics and Statistics  
State University of New York at Stony Brook  
Stony Brook, NY 11794 - 3600  
U.S.A.

James J. Chen  
Division of Biometry and Risk Assessment  
National Center for Toxicological Research  
Food and Drug Administration  
Jefferson, AR 72079  
U.S.A.

Tsae-Yun D. Lin  
Division of Biometrics IV  
Center for Drug Evaluation and Research  
Food and Drug Administration  
Rockville, MD 20857  
U.S.A.

Table 1: Classification of the models by the proposed test procedure.

Class	Model	Group	Model
Class 0	M0	Group 0	M0, M1, M2 and M5
Class 1	M1 and M3	Group 1	M3 and M6
Class 2	M2 and M4	Group 2	M4 and M7
Class 3	M5, M6, M7 and M8	Group 3	M8

Table 2: Parameter values for the simulation.

	<u>Model</u>	<u>Parameter Values</u>
Intercept	<i>M8</i>	$\alpha = 101$
	<i>M3, M4, M6, M7</i>	$\alpha_1 = 102, \alpha_2 = 101, \alpha_3 = 100$
	<i>M0, M1, M2, M5</i>	$\alpha_{11} = 100.25, \alpha_{12} = 101.75, \alpha_{13} = 100.5,$ $\alpha_{21} = 102, \alpha_{22} = 100.75, \alpha_{23} = 101.5,$ $\alpha_{31} = 101, \alpha_{32} = 101.25, \alpha_{33} = 100$
Slope	<i>M5 – M8</i>	$\beta = -.16$
	<i>M1 – M4</i>	$\beta_1 = -.16, \beta_2 = -.24, \beta_3 = -.08$
	<i>M0</i>	$\beta_{11} = -.22, \beta_{12} = -.1, \beta_{13} = -.2,$ $\beta_{21} = -.08, \beta_{22} = -.18, \beta_{23} = -.12,$ $\beta_{31} = -.16, \beta_{32} = -.14, \beta_{33} = -.24$

Table 3: The power (%) for the individual  $F$ -tests. Ten thousand samples were generated in each case.

Test	Exp.	d.f.	$\alpha$			
			.05	.10	.20	.25
$M1$ vs. $M0$	1	6, 54	.8906	.9401	.9753	.9849
	2		.2859	.4095	.5729	.6345
$M3$ vs. $M1$	1	6, 60	.9546	.9794	.9931	.9944
	2		.3485	.4825	.6515	.7065
$M5$ vs. $M1$	1	2, 60	1	1	1	1
	2		.8805	.9143	.9458	.9542
$M6$ vs. $M3$	1	2, 66	.9999	1	1	1
	2		.7771	.8654	.9340	.9492
$M6$ vs. $M5$	1	6, 62	.9554	.9811	.9933	.9947
	2		.3529	.4846	.6516	.7080
$M8$ vs. $M6$	1	2, 68	1	1	1	1
	2		.8708	.9294	.9681	.9761

Table 4: Number of model choices in 10,000 samples in Experiment 1. The right choice is identified as a bold faced number.

Real model	$\alpha$	Chosen model								
		<i>M0</i>	<i>M1</i>	<i>M2</i>	<i>M3</i>	<i>M4</i>	<i>M5</i>	<i>M6</i>	<i>M7</i>	<i>M8</i>
<i>M0</i>	.25	<b>9795</b>	114	54	0	0	37	0	0	0
	.20	<b>9678</b>	181	75	0	0	66	0	0	0
	.10	<b>9221</b>	381	180	0	0	218	0	0	0
	.05	<b>8620</b>	580	286	0	0	514	0	0	0
<i>M1</i>	.25	2472	<b>7484</b>	0	42	0	2	0	0	0
	.20	1954	<b>7976</b>	0	67	0	3	0	0	0
	.10	1022	<b>8787</b>	0	183	0	8	0	0	0
	.05	487	<b>9061</b>	0	435	0	16	1	0	0
<i>M3</i>	.25	2472	1914	0	<b>5612</b>	0	0	2	0	0
	.20	1954	1602	0	<b>6441</b>	0	1	2	0	0
	.10	1022	925	0	<b>8045</b>	0	1	7	0	0
	.05	487	492	0	<b>9004</b>	0	1	16	0	0
<i>M5</i>	.25	1597	875	872	3	3	<b>6603</b>	28	10	9
	.20	1194	761	755	4	5	<b>7205</b>	45	16	15
	.10	542	431	477	4	3	<b>8295</b>	125	72	51
	.05	219	252	265	4	3	<b>8688</b>	289	143	137
<i>M6</i>	.25	1597	251	875	627	0	1704	<b>4946</b>	0	0
	.20	1194	160	760	605	0	1489	<b>5792</b>	0	0
	.10	542	58	480	377	0	875	<b>7668</b>	0	0
	.05	219	19	268	237	0	469	<b>8786</b>	0	2
<i>M8</i>	.25	1597	251	255	627	620	1101	579	603	<b>4367</b>
	.20	1194	160	175	605	585	929	557	560	<b>5235</b>
	.10	542	58	56	377	424	485	396	390	<b>7272</b>
	.05	219	19	22	237	246	237	242	232	<b>8546</b>

Table 5: Number of model choices in 10,000 samples in Experiment 2. The right choice is identified as a bold faced number.

Real model	$\alpha$	Chosen model								
		$M0$	$M1$	$M2$	$M3$	$M4$	$M5$	$M6$	$M7$	$M8$
$M0$	.25	<b>5451</b>	1122	890	3	4	2519	8	3	0
	.20	<b>4811</b>	1182	912	8	6	3066	8	6	1
	.10	<b>3241</b>	1141	845	22	9	4635	49	37	21
	.05	<b>2123</b>	935	718	31	18	5843	154	92	86
$M1$	.25	2368	<b>4422</b>	93	1817	11	844	285	32	128
	.20	1870	<b>4114</b>	72	2232	12	1002	429	44	225
	.10	949	<b>3006</b>	61	2955	12	1172	1003	87	755
	.05	445	<b>1906</b>	25	3250	17	1115	1450	119	1673
$M3$	.25	2368	1566	99	<b>4673</b>	5	330	869	1	89
	.20	1870	1249	77	<b>5097</b>	7	325	1214	3	158
	.10	949	655	64	<b>5306</b>	9	216	2159	4	638
	.05	445	309	29	<b>484</b>	13	118	2800	7	1432
$M5$	.25	1597	661	668	217	207	<b>4137</b>	722	555	1236
	.20	1194	523	557	242	203	<b>4079</b>	860	650	1692
	.10	542	249	271	186	209	<b>3426</b>	1059	730	3328
	.05	219	111	124	145	144	<b>2497</b>	1042	725	4993
$M6$	.25	1597	251	815	627	60	1676	<b>4458</b>	28	488
	.20	1194	160	698	605	62	1459	<b>5092</b>	30	700
	.10	542	58	390	377	90	838	<b>5994</b>	37	1674
	.05	219	19	199	237	69	442	<b>5954</b>	27	2834
$M8$		The same as in TABLE IV								

Table 6: Number of group choices in 10,000 samples in Experiment 1. The right choice is identified as a bold faced number.

Real model	$\alpha$	Classification			
		Group 0	Group 1	Group 2	Group 3
<i>M0</i>	.25	<b>10000</b>	0	0	0
	.20	<b>10000</b>	0	0	0
	.10	<b>10000</b>	0	0	0
	.05	<b>10000</b>	0	0	0
<i>M1</i>	.25	<b>9958</b>	42	0	0
	.20	<b>9933</b>	67	0	0
	.10	<b>9817</b>	183	0	0
	.05	<b>9564</b>	436	0	0
<i>M3</i>	.25	4386	<b>5614</b>	0	0
	.20	3557	<b>6443</b>	0	0
	.10	1948	<b>8052</b>	0	0
	.05	980	<b>9020</b>	0	0
<i>M5</i>	.25	<b>9947</b>	31	13	9
	.20	<b>9915</b>	49	21	15
	.10	<b>9745</b>	129	75	51
	.05	<b>9424</b>	293	146	137
<i>M6</i>	.25	4427	<b>5513</b>	0	0
	.20	3603	<b>6397</b>	0	0
	.10	1955	<b>8045</b>	0	0
	.05	975	<b>9023</b>	0	0
<i>M8</i>	.25	3204	1206	1223	<b>4367</b>
	.20	2458	1162	1145	<b>5235</b>
	.10	1141	773	814	<b>7272</b>
	.05	497	479	478	<b>8546</b>

Table 7: Number of group choices in 10,000 samples in Experiment 2. The right choice is identified as a bold faced number.

Real model	$\alpha$	Classification			
		Group 0	Group 1	Group 2	Group 3
<i>M0</i>	.25	<b>9982</b>	11	7	0
	.20	<b>9971</b>	16	12	1
	.10	<b>9862</b>	71	46	21
	.05	<b>9619</b>	185	110	86
<i>M1</i>	.25	<b>7727</b>	2102	43	128
	.20	<b>7058</b>	2661	56	225
	.10	<b>5188</b>	3958	99	755
	.05	<b>3491</b>	4700	136	1673
<i>M3</i>	.25	4363	<b>5542</b>	6	89
	.20	3521	<b>6311</b>	10	158
	.10	1884	<b>7465</b>	13	638
	.05	901	<b>7647</b>	20	1432
<i>M5</i>	.25	<b>7063</b>	939	762	1236
	.20	<b>6353</b>	1102	853	1692
	.10	<b>4488</b>	1245	939	3328
	.05	<b>2951</b>	1187	869	4993
<i>M6</i>	.25	4339	<b>5085</b>	88	488
	.20	3511	<b>5697</b>	92	700
	.10	1828	<b>6371</b>	127	1674
	.05	879	<b>6191</b>	96	2834
<i>M8</i>		The same as in TABLE 6			

Table 8: Analysis of covariance tables for the data from Shao and Chow (1994) using the FDA procedure for the separate packages.

Package: Bottle					
Source	SS	DF	MS	F	P-value
Slope & Int.	22.33	8	2.79	2.91	0.025
Slope	16.75	4	4.19	4.36	0.011
Intercept	5.59	4	1.40	1.46	0.253
Residual	19.19	20	.96		

  

Package: Blister					
Source	SS	DF	MS	F	P-value
Slope & Int.	24.17	8	3.02	2.32	0.061
Slope	16.55	4	4.14	3.18	0.036
Intercept	7.62	4	1.90	1.46	0.251
Residual	26.03	20	1.30		

Table 9: Results of the proposed test procedure for the data in Shao and Chow (1994).

Step	Model	Residual SS	df	F
1	0	45.22	40	
	1	78.52	48	(M1) vs. (M0): $F_{8,40} = 3.68$
	2	51.81	45	(M2) vs. (M0): $F_{5,40} = 1.16$
2	4	53.83	50	(M4) vs. (M2): $F_{5,45} = 0.35$

Table 10: Analysis of covariance tables for the data from Shao and Chow (1994) using the two packages combined.

---

Packages Combined					
Source	SS	DF	MS	F	P-value
Slope & Int.	39.87	8	4.98	4.63	0.0003
Slope	26.77	4	6.69	6.22	0.0004
Intercept	13.09	4	3.27	3.04	0.0255
Residual	53.83	50	1.08		

---