International Journal of Industrial Ergonomics 69 (2019) 23–28

Contents lists available at ScienceDirect

International Journal of Industrial Ergonomics

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

Sample size calculations for a functional human motion analysis: Application to vehicle ingress discomfort prediction

Hadi Ibrahim Masoud[a,*], Matthew P. Reed[b,c], Jionghua (Judy) Jin[c]

[a] Department of Industrial Engineering, University of Jeddah, Jeddah, Saudi Arabia
[b] University of Michigan Transportation Research Institute, Ann Arbor, USA
[c] Industrial and Operations Engineering, University of Michigan, Ann Arbor, USA

ARTICLE INFO

Keywords:
Sample size
Ingress
Human motion
Fraction disaccommodated
Functional data analysis

ABSTRACT

The ease of entering a vehicle, known as ingress, is one of the important ergonomic factors that car manufacturers consider during the process of vehicle design. Manufacturers frequently conduct human subject tests to assess ingress discomfort for different vehicle designs. Using subject tests, manufacturers are able to estimate the proportion of participants that report that they are discomfortable entering a vehicle, referred to in this paper as *fraction disaccommodated* (FD). Manufacturers then conduct statistical tests in order to determine if the FD of two vehicle designs are significantly different, and to determine the required sample size in testing the FD difference between two vehicle designs under pre-specified testing power. Since conducting human subject tests is often expensive and time consuming, another alternative is to estimate the FD using simulated human motion data. Determining the number of simulations that is required is an important statistical question that is dependent on the prediction performance of the simulation analysis. In this paper, a dual bootstrap approach is proposed to obtain the standard deviation of the estimated FD based on functional predictors. This standard deviation is then used to calculate the power in testing the difference between two estimated FDs.

1. Introduction

The ease of getting into a vehicle, known as ingress, is an important consideration for customer satisfaction in the automotive industry (Morgans and Thorness, 2013). This has motivated vehicle manufacturers to focus on assessing and improving ingress discomfort. The most straightforward way to assess ingress discomfort is to build prototypes or mockups and have human participants test these potential vehicle designs. Participants rate the ease of getting into the vehicle using a Likert scale. For example, using a 10-point scale, participants might rate a design 1 out of 10 if it is very difficult to get into the vehicle and 10 out of 10 if the ingress motion is exceptionally comfortable. These ingress ratings can also be transformed into binary responses using a cutpoint. Using cutpoint 5, for example, ratings below or equal to 5 are transformed to 0 (or "uncomfortable") and ratings above 5 are transformed to 1 (or "comfortable"). One metric of interest is the proportion of participants who rated the ingress discomfort of a design above a defined cutpoint, referred to as *fraction disaccommodated* (FD). As the population *FD* (true *FD*) for a certain vehicle design is unknown, the participants responses are usually considered as a sample for estimating the population ingress fraction disaccommodated, which is denoted as $\widetilde{FD}$ in this research.

As it is generally expensive and time-consuming to conduct tests with participants to assess ingress discomfort, manufacturers seek more efficient ways to assess ingress discomfort, including computer simulation (Wegner et al., 2007). Advances in digital human modeling technologies have provided the ability to simulate the ingress motion of people with a wide range of anthropometric features (Reed et al., 2006; Reed and Huang, 2008). However, even if accurate methods for simulating ingress motions are available, it is still necessary to predict the subjective responses from the motion data. Masoud et al. (2016) developed a systematic framework that used human motion trajectories to predict subjective ingress discomfort responses using a machine-learning approach based on support vector machines (SVM). By using this framework, the FD of a vehicle design can be predicted by conducting simulations for a wide range of drivers (e.g., tall and short, young and old) and predicting subjective responses from the simulated motion data. This simulation-based approach can expedite the vehicle design validation process and reduce the cost of testing participants in physical mockups. To differentiate between the estimated FD obtained using participant responses ($\widetilde{FD}$) and the predicted FD obtained using actual or simulated human motion data, we denote the latter as $\widehat{FD}$.

* Corresponding author.
E-mail addresses: masoud@uj.edu.sa (H.I. Masoud), mreed@umich.edu (M.P. Reed), jhjin@umich.edu (J.J. Jin).

https://doi.org/10.1016/j.ergon.2018.09.010
Received 31 October 2017; Received in revised form 16 May 2018; Accepted 26 September 2018
Available online 04 October 2018
0169-8141/

In many cases, manufacturers are interested in knowing whether the ingress discomfort of one design is better than that of another. For this purpose, manufacturers may conduct a statistical hypothesis test to examine whether the FD of one design is significantly higher than that of another. Moreover, after a design change has been made, manufacturers seek to determine the minimum sample size that can provide a definitive assessment of the difference between two designs in terms of their FD values. In literature, many methods have been developed to test whether there is a significant difference between two proportions (Newcombe, 1998). Power calculations and sample size determination for testing the difference between proportions have also been studied (Faul et al., 2007; Cohen, 2013). In these methods, the responses used to estimate the proportions are assumed to be i.i.d (independent and identically distributed) and to follow a binomial distribution, i.e., each response has an equal probability of success (p) and the standard deviation of the sample proportions is equal to $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$. Although this assumption is appropriate for $\widetilde{FD}$, which is estimated using subjective responses, it is not immediately apparent that this relationship can be used to estimate the standard deviation of $\widehat{FD}$ due to the complex relationship between the motion model parameterization and the predicted subjective responses.

The objective of this paper is to develop a method for conducting power calculations in comparing two $\widehat{FD}$s in which the response proportion $\widehat{FD}$ are predicted from functional data obtained either from physical or virtual experiments. To conduct the power calculations, we must estimate the standard deviation of $\widehat{FD}$, referred to as σ_{FD} in this research. We developed a dual-bootstrapping approach that enables us to consider the two sources of variation in σ_{FD}. One is the modeling variation, which is due to the uncertainty of the estimated prediction model (σ_m) under different training datasets, and the other is the sampling variation due to the randomness of selecting test participants from the population (σ_s).

2. Methods

2.1. Data source

The data in this study was obtained from a vehicle ingress experiment that was conducted to study and reduce discomfort during ingress (Masoud et al., 2016). In brief, the experiment captured human motion data from 32 participants during vehicle ingress trials. Participants evaluated 17 vehicle designs that differed widely in the layout of the driver entry area. During each ingress test, reflective markers were used to record the location over time (trajectories) of 20 different joints. The trajectories of each joint were modeled by 27 B-spline coefficients. After participants completed an ingress trial (sample), they rated the ease of getting into the car on a 10-point scale, where 1 represents an unacceptable ingress experience and 10, an exceptionally comfortable ingress experience. The ingress discomfort rating was then transformed into a binary response using the cutpoint equal to 5, i.e., ratings below or equal to 5 were set as 0, and those above 5 were set as 1. In this research, the Cartesian trajectories of the 5 joints (left hip, right shoulder, right elbow, S1L5, and head) are used. The coordinates of these kinematic joints were identified by Masoud et al. (2016) as the most informative kinematic data for predicting ingress discomfort.

2.2. Method overview

A dual bootstrap or resampling approach was developed to estimate σ_{FD}, which includes two types of variation, σ_m and σ_s. A bootstrap approach is necessary because the complex relationship between the motion model and the predicted subjective responses precludes the use of the binomial distribution for estimating the standard deviation for the response proportion $\widehat{FD}$. As shown in Fig. 1, the first step is to generate a set of "bootstrap training datasets" by randomly resampling from the original dataset $(X_t,\ Y_t)$ obtained from physical participant-tests described in the previous section, where X_t is the human motion data and Y_t, the corresponding participant ingress discomfort response. Each of the generated bootstrap training datasets $(X_t^{*b},\ Y_t^{*b})$ $(b=1,\ldots, B)$ is used to train a prediction model using an SVM classifier (Masoud et al., 2016). With B bootstrap training datasets, we can obtain a set of prediction models, i.e., B different SVM classifiers, as shown in Fig. 1. The second step is to generate "bootstrap prediction datasets" for the two designs to be compared. As shown in Fig. 1, the bootstrap prediction datasets are generated by randomly resampling from X_p to generate J bootstrap prediction datasets $X_p^{*1},\ X_p^{*2},\ \ldots, X_p^{*j},\ \ldots,\ X_p^{*J}$. These bootstrap prediction datasets are then used along with one trained SVM model to predict J $\widehat{FD}$ for the design of interest (i.e., one $\widehat{FD}$ for each bootstrap prediction dataset). These predicted $\widehat{FD}$ are used to predict the sampling variance (σ_s) that arises due to the randomness in the prediction dataset. By repeating this process B times, through each of the SVM models, we can estimate the modeling variance (σ_m) induced by the uncertainty in the estimated prediction models. The details of each step are discussed in the following subsections.

2.3. Generate bootstrap training datasets

Assume that $X_t = (x_1^t,\ x_2^t,\ \ldots,\ x_{n_0}^t)$ represents the original training dataset obtained from the human participant-tests, where x_i^t represents the vector of human motion data of one ingress sample, represented as B-spline coefficients; n_0, the number of samples; and $Y_t = (y_1^t,\ y_2^t,\ \ldots,\ y_{n0}^t)$, the participant's binary ingress discomfort responses, where y_i^t is the discomfort rating corresponding to the motion data sample x_i^t. A bootstrap training dataset $X_t^{*b} = (x_1^{*b},\ x_2^{*b},\ \ldots,\ x_{n0}^{*b})$, $Y_t^{*b} = (y_1^{*b},\ y_2^{*b},\ \ldots,\ y_{n0}^{*b})$ is generated by randomly resampling, with replacement, n_0 times from the original dataset X_t and Y_t, where * represents a bootstrap sample and b, the bootstrap replication index. This replication process is performed B times to generate a large number of bootstrap training datasets $X_t^{*1},\ X_t^{*2},\ \ldots,\ X_t^{*B}$ and $Y_t^{*1},\ Y_t^{*2},\ \ldots,\ Y_t^{*B}$. In this analysis, the number of bootstrap datasets, denoted as B, was set to 100.

2.4. Train SVM prediction models

In this step, the bootstrap training datasets are used to train SVM prediction models (Cortes and Vapnik, 1995). SVM is a supervised learning classifier that has gained popularity in recent years as it can handle nonlinear classification and is robust to outliers (Cherkassky and Ma, 2004; Pal and Foody, 2010).

In this work, each set of bootstrap training datasets, X_t^{*b} and Y_t^{*b}, was used to train a separate SVM classifier, thus generating B different SVM models $({}^{(1)}SVM,\ {}^{(2)}SVM,\ \ldots,\ {}^{(b)}SVM,\ \ldots, {}^{(B)}SVM)$. The SVM models were trained using a Gaussian RBF kernel. The parameters of the RBF kernel were optimized for the original datasets X_t and Y_t using grid search to minimize the bias between the FD estimated from the prediction model $(\widehat{FD})$ and that estimated from participant responses $(\widetilde{FD})$; i.e., $\sum_{d=1}^{D}\left(\widehat{FD}_d - \widetilde{FD}_d\right)^2$ is minimized, where d is the index of different vehicle designs. Details of training an SVM model for classifying functional data can be found in Masoud et al. (2016).

2.5. Generate bootstrap prediction datasets

Assume that $X_p = (x_1^p,\ x_2^p,\ \ldots,\ x_n^p)$ *and* $X_{p'} = (x_1^{p'},\ x_2^{p'},\ \ldots,\ x_n^{p'})$ represent the human motion data corresponding to two different designs indicated by subscripts p and p' respectively, where n represents the number of motion data samples obtained though visual experimental tests or computer simulations. The participants tested in Design p can be either different from those in Design p', referred to as independent

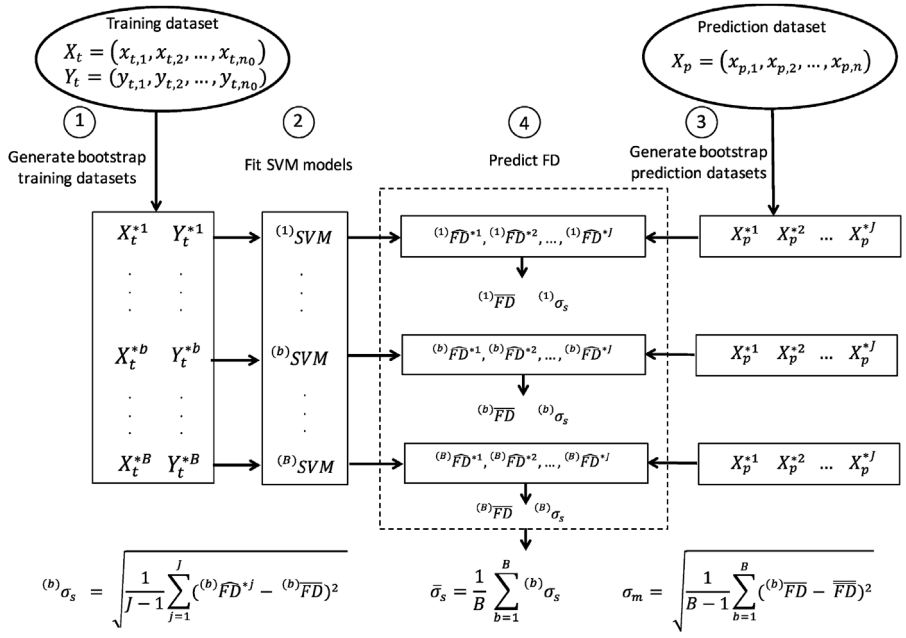

Fig. 1. Method overview.

datasets, or can be the same, referred to as paired datasets.

At this second bootstrap, bootstrap prediction datasets are generated for each of the two designs by following the same procedure as that used for generating the training datasets. The generated bootstrap prediction datasets for Design p and design p' are denoted as X_p^{*1}, X_p^{*2}, ...,X_p^{*j}, ..., X_p^{*J} and $X_{p'}^{*1}$, $X_{p'}^{*2}$, ...,$X_{p'}^{*j}$, ... $X_{p'}^{*J}$ respectively. Each dataset has sample size equal to n. The number of bootstrap datasets generated was set as $J = 100$.

2.6. Estimate FD and its standard deviation using bootstrap prediction datasets

In this step, the FD of the design of interest was estimated using the bootstrap prediction datasets. Specifically, a trained SVM model ($^{(b)}SVM$) was used to predict the ingress discomfort ratings Y_p^{*j} for the datasets X_p^{*j}. Subsequently, one sample of the estimated $\widehat{FD}$ was calculated as

$$\widehat{FD}^{*j} = \frac{1}{n}\sum_{i=1}^{n} y_i^{*j}. \tag{1}$$

where y_i^{*j} is the binary ingress discomfort predicted by the bth SVM model for the bootstrap iteration j. Similarly, this process was conducted on all J bootstrap prediction datasets, producing a set of estimates ($^{(b)}\widehat{FD}^{*1}$, $^{(b)}\widehat{FD}^{*2}$, ..., $^{(b)}\widehat{FD}^{*J}$.) The standard deviation of the sampling variation ($^{(b)}\sigma_s$) was then estimated by

$$^{(b)}\sigma_s = \sqrt{\frac{1}{J-1}\sum_{j=1}^{J}\left(^{(b)}\widehat{FD}^{*J} - {}^{(b)}\widehat{FD}^{*2}\right.} \tag{2}$$

where $^{(b)}\overline{FD} = \frac{1}{J}\sum_{j=1}^{J}{}^{(b)}\widehat{FD}^{*J}$. These steps were repeated on all B SVM models, thus producing the results of ($^{(1)}\overline{FD}$, $^{(b)}\overline{FD}$, ..., $^{(B)}\overline{FD}$) and ($^{(1)}\sigma_s$, $^{(b)}\sigma_s$, ...,$^{(B)}\sigma_s$. The standard deviation of the modeling variation was then calculated by

$$\sigma_m = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(^{(b)}\overline{FD} - \overline{\overline{FD}}\right)^2} \tag{3}$$

where $\overline{\overline{FD}} = \frac{1}{B}\sum_{b=1}^{B}{}^{(b)}\overline{FD}$. The average standard deviation of the sampling variance is then obtained by

$$\bar{\sigma}_s = \frac{1}{B}\sum_{b=1}^{B}{}^{(b)}\sigma_s \tag{4}$$

Finally, the total standard deviation of the FD for the design of interest (p) was calculated as

$$\sigma_{FD_p} = \sqrt{\sigma_m^2 + \bar{\sigma}_s^2} \tag{5}$$

All the steps explained in this section were used to estimate the standard deviation of the second design ($\sigma_{FD_{p'}}$). Moreover, for independent observations where the selected participants in Design p are different from those in design p', the standard deviation of the difference ($\widehat{FD}_p - \widehat{FD}_{p'}$) was calculated as

$$\sigma_{\Delta FD} = \sqrt{\sigma_{FD_p}^2 + \sigma_{FD_{p'}}^2} \tag{6}$$

In contrast, for paired observations where the same participant was selected under both designs, $\sigma_{\Delta FD}$ was calculated as

$$\sigma_{\Delta FD} = \sqrt{\sigma_{FD_p}^2 + \sigma_{FD_{p'}}^2 - 2\,cov\left(\widehat{FD}_p, \widehat{FD}_{p'}\right)} \tag{7}$$

where

$$cov\left(\widehat{FD}_p, \widehat{FD}_{p'}\right) = \frac{1}{(B)(J-1)}\sum_{b=1}^{B}\sum_{j=1}^{J}\left(^{(b)}\widehat{FD}_p^{*j} - {}^{(b)}\overline{FD}_p\right)\left(^{(b)}\widehat{FD}_{p'}^{*j} - {}^{(b)}\overline{FD}_{p'}\right) \tag{8}$$

It should be clarified that the covariance between $\widehat{FD}_p$ and $\widehat{FD}_{p'}$ will only be evident if the bootstrap is done simultaneously for the two designs, such that a participant who is chosen in the bootstrap prediction sample X_p^{*b} is also chosen in $X_{p'}^{*b}$.

2.7. Estimate σ_{FD_p} under different sample sizes

The value of σ_{FD_p} is a function of the sample size (n) of the original prediction datasets X_p. To analyze the effect of the sample size on σ_{FD_p},

σ_{FD_p} needs to be estimated under various sample sizes (n^*) for a given design p. The resultant estimates of σ_{FD_p} can then be used to evaluate the testing power for comparing two designs under different sample sizes n^*.

To calculate σ_{RP_p} for a sample size n^*, we used a technique called oversampling (Japkowicz, 2000). Assume that $X_p = (x_1^p, x_2^p, ..., x_n^p)$ represents the human motion data of the design of interest, where n represents the number of participants. Assume that we are interested in calculating σ_{FD_p} for a dataset with a different sample size n^*. For this purpose, we generated bootstrap prediction datasets $X_p^{*1}, X_p^{*2}, ..., X_p^{*J}$ from the dataset X_p, where we oversampled with replacement from X_p such that each bootstrap prediction dataset X_p^{*j} ($j = 1,...,J$) has a new sample size n^*. Once the bootstrap prediction datasets are generated, the estimates of σ_{FD_p} and $\sigma_{\Delta FD}$ can be obtained for this new sample size n^*

2.8. Hypothesis testing and sample size determination

Once ($\sigma_{\Delta FD}$) is obtained, it is possible to test whether there is a significant difference between the FD of the two designs. The hypothesis was formulated as follows:

H_0: $\Delta FD = 0$
H_1: $\Delta FD \neq 0$

where $\Delta FD = FD_1 - FD_2$

To conduct this hypothesis test, a standardized test statistic was defined as

$$z = \frac{\Delta\widehat{FD}}{\sigma_{\Delta FD}} \tag{9}$$

where $\Delta\widehat{FD} = \overline{\widehat{FD}}_1 - \overline{\widehat{FD}}_2$. If a normal distribution is assumed for z, the null hypothesis H_0 is rejected when z is larger than $z_{\alpha/2}$, where α is the pre-specified significance level.

The power to detect a difference of $\delta = \Delta FD = FD_1 - FD_2$ was calculated as

Power = $2P(|z| > z_{\alpha/2} \mid \delta)$, which can be rewritten as follows:

$$\text{Power} = 1 - \Phi\left[z_{\alpha/2} - \frac{\delta}{\sigma_{\Delta FD}}\right] + \Phi\left[-z_{\alpha/2} - \frac{\delta}{\sigma_{\Delta FD}}\right] \tag{10}$$

where Φ is the cumulative distribution function of the standard normal distribution.

3. Results

3.1. Comparing FD predictions: $\widehat{FD}$ vs. $\widetilde{FD}$

In this section, we compared the FD predicted using human motion data ($\widehat{FD}$) with that predicted using direct participant responses ($\widetilde{FD}$). As previously explained in the introduction, these two ($\widehat{FD}$ and $\widetilde{FD}$) are the estimates of the true (unknown) FD. The closeness between $\widehat{FD}$ and $\widetilde{FD}$ will give us an indication of how well the prediction model is performing and give us more confidence to use $\widehat{FD}$ for evaluating future vehicle designs when human participant responses are not available.

Table 1 shows the comparison results between the estimates $\widehat{FD}$ and $\widetilde{FD}$ in which the top-rated 7 out of 17 vehicle designs are selected based on their $\widetilde{FD}$. For each design, the $\widehat{FD}$ was calculated such that the motion data of that design was treated as the prediction dataset, and the remaining 6 designs were used to train the SVM model. Table 1 shows that the predicted $\widehat{FD}$ agrees well with the estimated $\widetilde{FD}$ (correlation, r = 0.93). The average absolute bias among these 7 designs was equal to 0.04.

It is worthwhile to note that in the above comparison, the SVM models were trained using 6 vehicle designs that had relatively high $\widetilde{FD}$ values ($\widetilde{FD} > 0.4$). Choosing designs with relatively similar $\widetilde{FD}$ values (i.e., all high or all low FD values) helps reduce the bias between $\widehat{FD}$ and $\widetilde{FD}$. This is reasonable for design comparison as, in practice, manufacturers often try to choose the best design among a group of good designs. The method presented in this work, however, can still be applied if the model was trained using the data of all vehicle designs. Table 2 shows a comparison between $\widehat{FD}$ and $\widetilde{FD}$ for the same 7 vehicle designs; however, the SVM model was trained using all vehicle designs (excluding the predicted vehicle design). We can observe that $\widehat{FD}$ and $\widetilde{FD}$ are still highly correlated (r = 0.88); however, the average absolute bias (0.076) of the results in Table 2 is larger than that (0.04) of the results in Table 1.

Table 1
Comparison of $\widehat{FD}$ and $\widetilde{FD}$ estimates (SVM model trained using the top 7 vehicle designs).

Vehicle Design	$\widehat{FD}$	$\widetilde{FD}$	Bias
1	0.6667	0.6667	0
2	0.7143	0.7143	0
3	0.5417	0.458	0.0837
4	0.8095	0.8517	−0.0422
5	0.85	0.85	0
6	0.8	0.7	0.1
7	0.7368	0.6842	0.0526

Table 2
Comparison of $\widehat{FD}$ and $\widetilde{FD}$ estimates (SVM trained using all vehicle designs).

Vehicle Design	$\widehat{FD}$	$\widetilde{FD}$	bias
1	0.476	0.6667	−0.1907
2	0.6667	0.7143	−0.0476
3	0.4167	0.4583	−0.0416
4	0.8095	0.8571	−0.0476
5	0.80	0.85	−0.05
6	0.60	0.70	−0.1
7	0.7368	0.6842	0.0526

3.2. Statistical testing and sample size calculations for comparing two vehicle designs

In this section, we apply the proposed method to estimate σ_{FD_p} for 2 of the 7 selected vehicle designs (designs 2 and 4). The motion data of these two designs were used for predicting $\widehat{FD}$, while the data for the remaining 5 designs were used as training data for training the SVM models. The predicted $\widehat{FD}$ of the 2 selected designs were 0.6667 and 0.8095, respectively. In this analysis, only the 18 participants who evaluated both designs were included in the prediction dataset. Two analyses were conducted: one, to determine whether these two designs show significantly different FDs and the other, to determine the sample size required for detecting a difference of $\delta = 0.2$ between the two designs with a testing power no less than 70%.

Using the original training data, 100 bootstrap training datasets were generated each with a sample size $n_0 = 104$. Each bootstrap training dataset was then used to train an SVM prediction model. Bootstrap prediction datasets were then generated for each design. As we use paired observations in this comparison, the bootstrap resampling was performed for both designs simultaneously such that the covariance structure was appropriately captured. These bootstrap prediction datasets were then used to estimate the modeling and sampling variance for both designs. The standard deviation $\sigma_{\Delta FD}$ was then calculated based on Eq. (7) for the paired observations, which yields $\sigma_{\Delta FD} = 0.0966$.

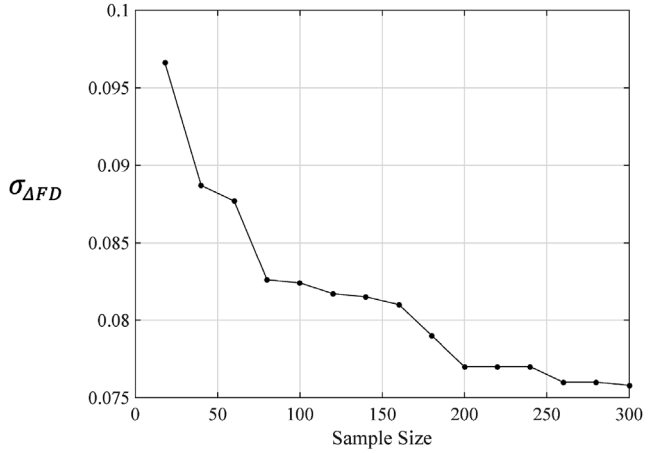

Fig. 2. Effect of increasing sample size on $\sigma_{\Delta FD}$.

To test whether the FD of the above two designs differs significantly, the critical value, as defined in Eq. (9), was calculated as z = 1.48. If $\alpha = 0.05$ is used, we cannot reject H_0. As $z < z_{\alpha/2} = 1.96$, i.e., we cannot conclude that there is a significant difference between the FD of the above two designs based on the currently available motion data from these 18 participants.

To further answer how many participants would be required to detect a difference of $\delta = 0.2$ between those two designs, we create bootstrap prediction datasets by applying the oversampling strategy for a wide range of sample sizes, based on which $\sigma_{\Delta FD}$ was estimated for each sample size and the testing power was calculated using Eq. (10). Fig. 2 and Fig. 3 show the effect of increasing the sample size of the prediction datasets from $n^* = 18$ to $n^* = 300$ on $\sigma_{\Delta FD}$ and the testing power. It should be noted that these curves are theoretically smooth; the irregularity results from the underlying data from which the samples are drawn. From Fig. 2, we can observe that $\sigma_{\Delta FD}$ decreases as we increase the sample size of the prediction datasets. Fig. 2 also shows that there is a slow decrease in $\sigma_{\Delta FD}$ after the sample size increases to $n = 200$. This shows that the modeling variance (σ_m) is not affected by the increase in the sample size of the prediction dataset. Fig. 3 shows that 180 samples of each design are needed to achieve 70% power in detecting a 0.2 difference in FD.

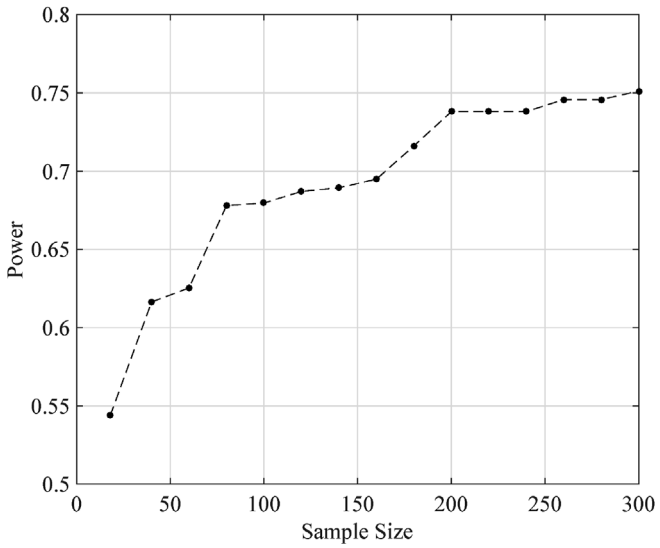

Fig. 3. Effect of sample size on power to detect a 0.2 difference in FD.

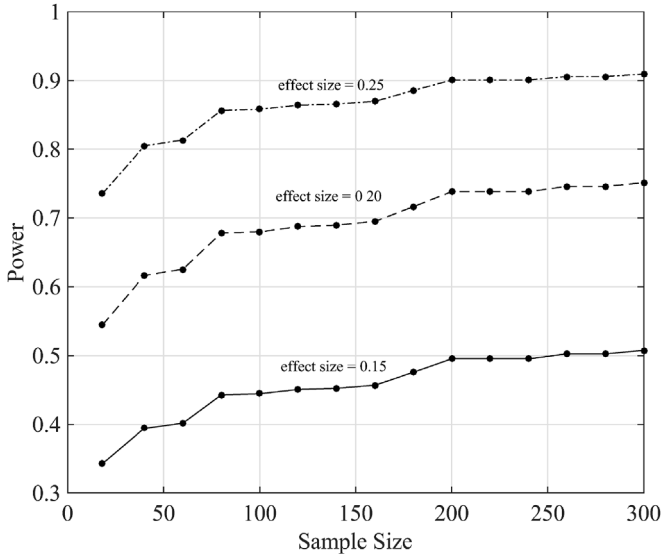

Fig. 4. Effect of increasing sample size on power for a range of effect sizes.

3.3. Estimating the sample size for different effect sizes (δ)

Figure 4 illustrates the improvement in power as we increase the sample size of the prediction datasets from $n^* = 18$ to $n^* = 300$ for a range of effect sizes (δ): $\delta = 0.15$, $\delta = 0.20$, and $\delta = 0.25$. We can observe that for effect size $\delta = 0.25$ the number of samples required to achieve a 70% power is 18 samples. We also observe that we cannot achieve a 70% power for $\delta = 0.15$ due to the high modeling variance. Data from more participants would be needed for the training dataset in order to reduce the modeling variance. A similar oversampling technique can be used to estimate the behavior of the modeling variance as we add samples to the training dataset.

4. Discussion and conclusion

In this paper, a new dual bootstrap method was proposed to conduct power calculations for a virtual or physical experimental determination of a binomial outcome based on a functional model. This method was elaborated in the context of estimating the percentage of a driver population who would rate vehicle ingress as “uncomfortable.”

One important contribution is the separate consideration of modeling variance (i.e., the uncertainty of the prediction models) and sampling variance (the uncertainty due to differences in responses among the sampled individuals.) The case study demonstrated that the model variance imposes a bound on the tradeoff between effect size and power. In other words, in some circumstances, a better model may be necessary to achieve a desired power to detect a particular effect size; sampling more individuals for the particular comparison of interest will not help.

The sample size estimated using the method presented in this paper represents the upper bound for the samples needed. The method assumes that samples are randomly chosen for the population of participants. If the sampling of participants is strategically done such that easy-to-predict participants are sampled less than total number of samples required may be less.

The results for sample size determination shown in the case study are limited by the particular ingress dataset that were used and by the trained SVM prediction models. Different datasets or a different approach to modeling the relationship between the measurable quantities (in this case, motions) and the binary responses to be predicted (in this case, discomfort ratings) might produce different model behaviors. However, the general approach developed here would still be applicable. One essential requirement of using the proposed method is that

the data underlying the prediction model must be available to enable this resampling approach. A different method for estimating model variance would be needed if the original data were not available.

Acknowledgment

This research was jointly supported by Ford Motor Company and NSF-CMII-MES Award 1233108. The authors would like to thank Nancy Wang, Ksenia Kozak, Jian Wan, and Gianna Gomez-Levi in Ford Motor Company for making these data available. The authors would also like to thank Prof. Kerby Shedden, Director of the Center for Statistical Consultation and Research (CSCAR), for his time and invaluable statistical consultation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ergon.2018.09.010.

References

Cherkassky, V., Ma, Y., 2004. Practical selection of SVM parameters and noise estimation for SVM regression. Neural Network. 17 (1), 113–126.
Cohen, J., 2013. Statistical Power Analysis for the Behavioral Sciences. Academic press.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.
Faul, F., et al., 2007. G* Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39 (2), 175–191.
Japkowicz, N., 2000. The class imbalance problem: significance and strategies. In: Proc. Of the Int'l Conf. on Artificial Intelligence.
Masoud, H.I., et al., 2016. Predicting subjective responses from human motion: application to vehicle ingress assessment. J. Manuf. Sci. Eng. 138 (6), 061001 (2016).
Morgans, S., Thorness, B., 2013. POWER: How JD Power III Became the Auto Industry's Adviser, Confessor, and Eyewitness to History. Greenleaf Book Group.
Newcombe, R.G., 1998. Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat. Med. 17 (8), 873–890.
Pal, M., Foody, G.M., 2010. Feature selection for classification of hyperspectral data by SVM. IEEE Trans. Geosci. Rem. Sens. 48 (5), 2297–2307.
Reed, M.P., Huang, S., 2008. Modeling Vehicle Ingress and Egress Using the Human Motion Simulation Framework. N. 2008-01-1896, SAE Technical Paper.
Reed, M.P., et al., 2006. The HUMOSIM Ergonomics Framework: a New Approach to Digital Human Simulation for Ergonomic Analysis. SAE Technical Paper.
Wegner, D., et al., 2007. Digital human modeling requirements and standardization. In: N. 2007-01-2498, SAE International Conference of Digital Human Modeling.