Optimal Data Analysis
Vol. 9 (June 10, 2020), 172-177

2155-0182/10/$3.00

Generating Novometric Confidence Intervals in R: Bootstrap Analyses to Compare Model and Chance ESS

Nathaniel J. Rhodes, Pharm.D., M.Sc. and Paul R. Yarnold, Ph.D.

Chicago College of Pharmacy, and the Pharmacometrics Center of Excellence, Midwestern University

Optimal Data Analysis, LLC

We introduce a method for evaluating the upper and lower bounds of statistical confidence [e.g., an exact discrete confidence interval (CI)] in the expected effect strength for sensitivity of a decision model. The method presented herein uses bootstrap simulations to determine if a given effect is robust and not attributable to random chance (i.e., the lower bound of the model bootstrap CI is greater than the upper bound of the chance bootstrap CI). We evaluated relatively weak, moderate, and strong effects using the novometric bootstrap method. Our approach should serve to increase confidence in the veracity of decision models.

Scientists intrinsically seek "game-changing" changes in *state*: fixes, cures, and answers—*solutions* achieved via the scientific method. Statistics *is not needed* for discoveries which create *qualitative* transformations.

In pursuit of a game-changer, scientists extrinsically seek strong changes in *phase*: insights, improvements, and directions—*advances* achieved via the scientific method. Statistics *is useful* for quantifying chance-corrected (ESS) and complexity-corrected (D) accuracy that define the limits of predictable control.[1,2]

In reality, much empirical literature is challenged *theoretically* (construct validity), *methodologically* (inadequate and/or inappropriate sampling and/or measurement, ignoring crucial interactions, *not* investigating validity and reliability), and/or *statistically* (violated assumptions, suboptimal solutions). Statistics *is crucial* if methods are deficient, and as effects approach the decision criterion for statistical significance—a focus of novometric theory.[1]

This article focuses on the novometric methodology[1] of assessing degree of overlap in chance-adjusted classification accuracy (ESS) which exists between exact, discrete, confidence intervals (CIs) obtained for model *vs*. chance.[3,4] Herein we introduce a tool which we developed to generate such *novometric CIs* in R, thereby enabling researchers to directly assess *degree of ESS* overlap between model and chance.

Consider a model's point predictions. In simplest form, a model predicts an observation either *is* or *is not* a member of the positive class:

2155-0182/10/$3.00

for example a model which predicts presence *vs*. absence of SARS-2-CoV in the respiratory tract given some clinical data. This hypothetical model yields a confusion matrix reporting the number of sample observations in each of the cells created by crossing an observation's *actual* (rows) and *predicted* (columns) class status. If the strength of an observed effect is disputed, and/or if implications of the finding are crucial, then establishing the rigor and reproducibility of predictions is of great concern.

Independent verification of the model's predictions is obviously favored (i.e., the model has similar performance when used to classify a new independent random sample), but this is not always feasible. Fortunately, assessing validity of model predictions may also be accomplished by using simulation. For example, one may generate predictions using a re-sampling of the original observed data, and then compare this resampling to the predictions obtained from the model. If this process is repeated *n* times, one obtains *n* bootstrap predictions, which in turn can be used to calculate an exact discrete CI for the model's predictions.

Traditional statistical methodologies conclude that, if a given model prediction (true positive or true negative classification) exceeds a rate consistent with chance (i.e., the marginals observed exceed the marginals expected), then the predictions are unlikely to be due to chance alone. This is typically normed against a 1 in 20 false discovery rate or $p<0.05$. If model predictions are reproduced with similar performance by resampling (e.g., via jackknife or bootstrap analysis), an investigator gains confidence that error rates and 95% CIs are well estimated.

In contrast, *novometric theory* states the exact discrete CI for classification performance of the model applied to the actual data (*model*), *vs*. applied to the data with randomly scrambled class variable (*chance*), should not overlap.[4] The lower bound (LB) 2.5% CI for model metrics (e.g., PAC, ESS, sensitivity, specificity, etc.) should not fall below the upper bound (UB) of the corresponding 97.5% CI for chance.

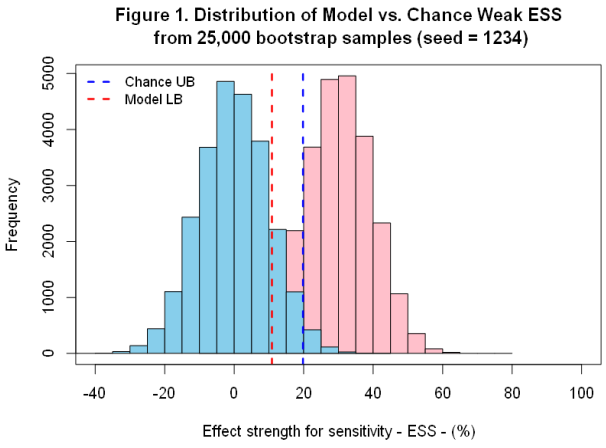
We consider three scenarios with ESS varying from weak to strong.[5,6] In each example the observations and predictions were resampled at 50% with replacement. Next, the observations were scrambled relative to the class variable, and scrambled (chance) data were resampled at 50% with replacement. Resampling was repeated for model and chance to obtain *n* bootstrap replicates (n=25,000 each). The overlap for the ESS of the model LB (i.e., the 2.5th percentile) and the chance UB (i.e., the 97.5th percentile) were then assessed. We used the function *NOVOboot()* which is available[7] in the ODA package for R: code for the analysis conducted herein is provided as an *Appendix* to this paper.

Example 1: Relatively Weak ESS (24%)

We first present an example of a model having a relatively weak effect, with ESS = 24%. Table 1 gives the confusion matrix obtained by using the model to classify the total sample, and Figure 1 summarizes the findings of this analysis.

Table 1: Training Confusion Matrix, Relatively Weak Model

	Predicted Class	
Actual Class	class = 0	class = 1
class = 0	62	38
class = 1	38	62



2155-0182/10/$3.00

Findings demonstrate that, for model and for chance, the exact, discrete CIs overlapped: dashed red (the 2.5th percentile of model) and dashed blue (the 97.5th percentile of chance) lines are incorrectly located relative to their respective distributions (light blue and rose, respectively). That is, if an effect is statistically significant, then the red dashed line should be on the same side as the rose histogram, whereas the blue dashed line should be on the same side as the blue histogram. Numerical results are summarized in Table 2.

Table 2: Model *vs*. Chance Bootstrap Distribution: Weak, Moderate, and Strong ESS

ESS:	Weak (24%)		Moderate (48%)		Strong (72%)	
Quantile	Model	Chance	Model	Chance	Model	Chance
0%	-8.28	-46.47	13.92	-40.06	47.34	-40.76
2.5%	10.87	-19.89	32.77	-19.95	62.10	-19.87
5%	14.00	-16.36	35.86	-16.35	64.41	-16.51
50%	30.11	0.00	50.57	0.00	76.09	0.00
95%	45.94	16.23	64.45	16.19	86.07	16.19
97.5%	48.85	19.81	67.27	19.80	88.00	19.63
100%	77.13	40.02	84.46	38.00	98.18	39.29

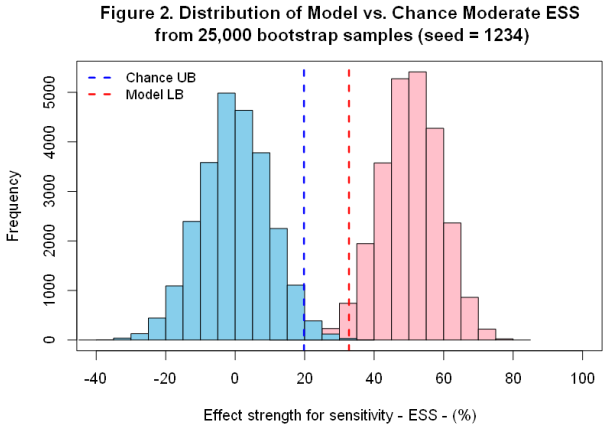
Example 2: Moderate ESS (48%)

Next, we considered a model achieving a moderate ESS (48%). Table 3 is the confusion matrix obtained using the model to classify the total sample.

Table 3: Training Confusion Matrix, Moderate Strength Model

	Predicted Class	
Actual Class	class = 0	class = 1
class = 0	74	26
class = 1	26	74

Figure 2 illustrates the net result of these simulations. As seen, for model (2.5% LB) and for chance (97.5% UB), exact, discrete CIs did not significantly overlap, and the red dashed line is on the same side as the rose histogram.



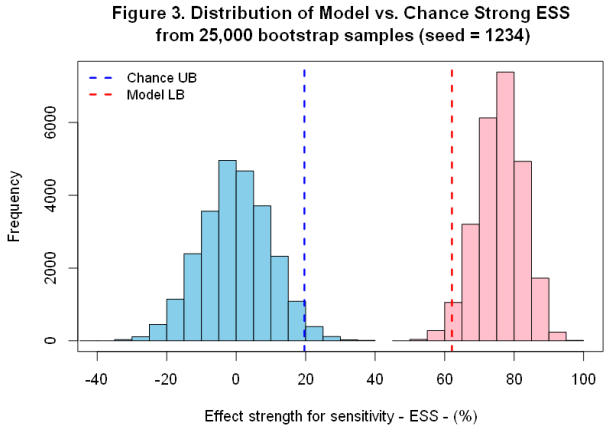
Example 3: Relatively Strong ESS (72%)

We next considered a model having a relatively strong effect as measured by ESS (72%). Table 4 is the confusion matrix obtained by using the model to classify the total sample.

Table 4: Training Confusion Matrix, Relatively Strong Model

	Predicted Class	
Actual Class	class = 0	class = 1
class = 0	86	14
class = 1	14	86

Figure 3 illustrates the net result of these simulations. Our findings demonstrate that for model and for chance, the exact, discrete CIs did not overlap at all.



2155-0182/10/$3.00

In summary, our simulation-based analysis demonstrated that a relatively weak effect (ESS=24%) was found non-significant relative to the chance distribution. On the other hand, analyses with moderate (48%) and strong (72%) ESS had model and chance distributions that did not (respectively) significantly or even partially overlap. The merits of evaluating the exact discrete CIs for model and chance have been previously discussed, and here are visually presented.[1,3,4] The present simulation-based methodology is a handy means of visually evaluating model and chance CIs to evaluate statistical significance, as is specified in Axiom One of novometric analyses.[4]

Our method has the additional advantage that the full distribution of bootstrap replicates can be evaluated once generated. Thus, investigators conducting hypothesis screening studies may select models exhibiting weak (ESS<25) to moderate (ESS<50) effects for follow-up research when the 90% exact discrete CI of model and chance do not overlap (i.e., $p < 0.1$). In this manner investigators can establish their own specific level of significance for follow-up research depending on their specific needs. Likewise, when multiple comparisons are made, the *a priori* level of alpha can be corrected using a Bonferroni-Sidak-type adjustment.[5,6]

The scientific enterprise is faced with a mounting crisis of trust magnified by recent events including the COVID-19 pandemic. Rigor and reproducibility of scientific activities are needed when the implications of evidence-based decisions are grave, stakes are high, and effects are marginal. Threats to rigor include spurious findings (e.g., "data dredging" or "*p*-hacking") and in general the lack of sound methodology. Threats to reproducibility range from lack of methodologic transparency[8] to methodologic bias.[9,10] An urgent need exists for robust analytic methods to guard against such threats and increase confidence in the veracity of research findings.

We conclude that the predictions produced by a given model can be evaluated for statistical rigor by examining the extent of the overlap between the exact discrete CIs for model and chance. To support analytic rigor, we recommend that investigators critically evaluate the strength of empirical effects using tools such as our simulation-based approach to guard against potentially non-reproducible weak and marginal effects, especially when stakes are high and model implications are profound.

References

[1]Yarnold PR, Soltysik RC (2016). Maximizing predictive accuracy. Chicago, IL: ODA Books.

[2]Yarnold PR (2015). Distance from a theoretically ideal statistical classification model defined as the number of additional equivalent effects needed to obtain perfect classification for the sample. *Optimal Data Analysis, 4*, 81-86.

[3]Yarnold P R (2018). Comparing exact discrete 95% CIs for model *vs*. chance ESS to evaluate statistical significance. *Optimal Data Analysis, 7*, 82-84.

[4]Yarnold PR (2020). Reformulating the first axiom of novometric theory: Assessing minimum sample size in experimental design. *Optimal Data Analysis, 9*, 7-8.

[5]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: Guidebook with software for Windows*. APA Books.

[6]Rhodes NJ (2020). Statistical power analysis in ODA, CTA and novometrics (Invited). *Optimal Data Analysis, 9*, 21-25.

[7]Rhodes N J, Yarnold P R (2020). *ODA: a package and R-interface for the MegaODA software suite.* Version: 1.0.1.3.

2155-0182/10/$3.00

[8]Piller C (2020). Who's to blame? These three scientists are at the heart of the Surgisphere COVID-19 scandal. *Science*. doi: 10.1126/science.abd2252.

[9]Sciama Y (2020). Is France's president fueling the hype over an unproven coronavirus treatment? *Science*. doi: 10.1126/science.abc1786.

[10]Voss A. (2020). *Official Statement from International Society of Antimicrobial Chemotherapy (ISAC)*. Statement on IJAA paper. Published April 3, 2020. Available: https://www.isac.world/news-and-publications/official-isac-statement

Author Notes

No conflict of interest was reported.

Appendix

Code for bootstrap resampling procedure using *NOVOboot()* function within the *ODA* package (version 1.0.1.3)[7] for R.

```
# Evaluating an ESS of 24% (a weak effect)
ess <- 100*(((0.62+0.62)/2)-0.5)/0.5

# Evaluating an ESS of 48% (a moderate effect)
ess <- 100*(((0.74+0.74)/2)-0.5)/0.5

# Evaluating an ESS of 72% (a strong effect)
ess <- 100*(((0.86+0.86)/2)-0.5)/0.5

#Enter data for primary confidence interval analysis
library(ODA)
library(epitools)
#Table Weak ESS 24%
data <-
matrix(c(62,38,38,62),ncol=2,nrow=2,dimnames=list(
c(0,1),c(0,1)))
data.raw <- expand.table(data)
data.tab <-

list(table(cbind(data.raw[1],data.raw[2]),dnn=c("v1","
x")))
oda.list.1 <- list()
oda.list.1[[1]] <- do.call("list",data.tab)

#Table Moderate ESS 48%
data <-
matrix(c(74,26,26,74),ncol=2,nrow=2,dimnames=list(
c(0,1),c(0,1)))
data.raw <- expand.table(data)
data.tab <-
list(table(cbind(data.raw[1],data.raw[2]),dnn=c("v1","
x")))
oda.list.2 <- list()
oda.list.2[[1]] <- do.call("list",data.tab)

#Table Strong ESS 72%
data <-
matrix(c(86,14,14,86),ncol=2,nrow=2,dimnames=list(
c(0,1),c(0,1)))
data.raw <- expand.table(data)
data.tab <-
list(table(cbind(data.raw[1],data.raw[2]),dnn=c("v1","
x")))
oda.list.3 <- list()
oda.list.3[[1]] <- do.call("list",data.tab)

# Run the bootstrap analyses and plot the CI
distributions

## Weak ESS = 24%
NOVOboot(data=oda.list.1,run=1,predictor=1,outco
me=1,seed=1234, nboot=25000)

hist(novo.boot.1$ess.model,xlim=c(-
40,100),main="Figure 1. Distribution of Model vs.
Chance Weak ESS\n from 25,000 bootstrap
samples (seed = 1234)",col="pink",xlab="Effect
strength for sensitivity - ESS - (%)")
hist(novo.boot.1$ess.chance,col="skyblue",xlim=c(-
40,100),add=T)
abline(v=c(quantile(novo.boot.1$ess.model,probs=0.
025),quantile(novo.boot.1$ess.chance,probs=0.975)
),col=c("red", "blue"), lty=c(2,2), lwd=c(2,2))
box()
legend("topleft", c("Chance UB", "Model LB"),
col=c("blue", "red"), lwd=2, lty=2, cex=0.9, bty = "n")

## Moderate ESS = 48%
NOVOboot(data=oda.list.2,run=1,predictor=1,outco
me=1,seed=1234, nboot=25000)

hist(novo.boot.1$ess.model,xlim=c(-
40,100),main="Figure 2. Distribution of Model vs.
Chance Weak ESS\n from 25,000 bootstrap
samples (seed = 1234)",col="pink",xlab="Effect
strength for sensitivity - ESS - (%)")
hist(novo.boot.1$ess.chance,col="skyblue",xlim=c(-
40,100),add=T)
```

2155-0182/10/$3.00

```
abline(v=c(quantile(novo.boot.1$ess.model,probs=0.
025),quantile(novo.boot.1$ess.chance,probs=0.975)
),col=c("red", "blue"), lty=c(2,2), lwd=c(2,2))
box()
legend("topleft", c("Chance UB", "Model LB"),
col=c("blue", "red"), lwd=2, lty=2, cex=0.9, bty = "n")

## Strong ESS = 72%
NOVOboot(data=oda.list.3,run=1,predictor=1,outco
me=1,seed=1234, nboot=25000)

hist(novo.boot.1$ess.model,xlim=c(-
40,100),main="Figure 3. Distribution of Model vs.
Chance Weak ESS\n from 25,000 bootstrap
samples (seed = 1234)",col="pink",xlab="Effect
strength for sensitivity - ESS - (%)")
hist(novo.boot.1$ess.chance,col="skyblue",xlim=c(-
40,100),add=T)
abline(v=c(quantile(novo.boot.1$ess.model,probs=0.
025),quantile(novo.boot.1$ess.chance,probs=0.975)
),col=c("red", "blue"), lty=c(2,2), lwd=c(2,2))
box()
legend("topleft", c("Chance UB", "Model LB"),
col=c("blue", "red"), lwd=2, lty=2, cex=0.9, bty = "n")
```