Repeatability and *Reproducibility* of Categorical Measurements (Classification)

Yariv N. Marmor
Emil Bashkansky
Tamar Gadrich

BRAUDE College of Engineering, Karmiel

Leuven - ENBIS 2024
16/09/2024, 13:50

The Purpose of the Presentation

To propose the model of estimating different components of classification precision, when it is provided by defined number of collaborators (classifiers) according to fixed and random model of their selection.

Definition of *“categorical”* measurement

“Classification of the analyzed property value of the objects under study (OUS) into one of K exclusive categories forming a comprehensive spectrum (scale) of the studied property will be considered as categorical measurement.” **(*)**

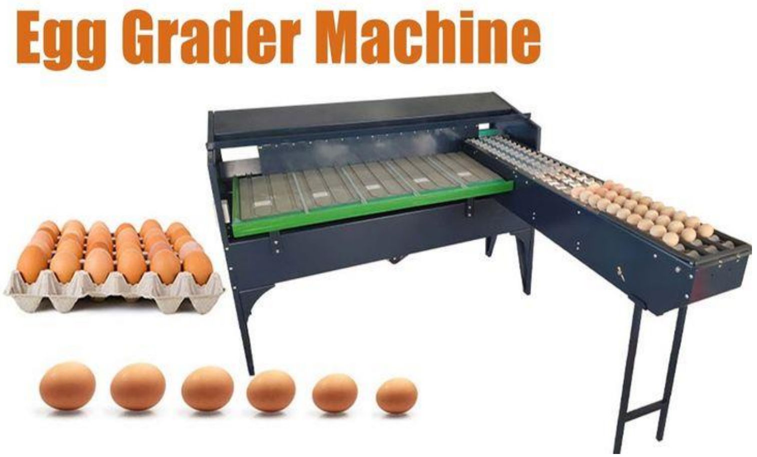
Note 1: The results of classification are presented by so-called categorical data. In cases where the spectrum of possible values consists only of two categories such data are binary, and the appropriate activity is also often called ***testing.***

Note 2: In this presentation categories are not ordered (*nominal* scale)

* T. Gadrich, E. Bashkansky, (2016) " A Bayesian approach to evaluating uncertainty of inaccurate categorical measurements", *Measurement* 91, 186–193.

Examples of *Classifiers* (*K* > 2)

Coins sorting machine



Egg classification machine

Plastic color sorting machine

Google language detector

Do
Re
Mi
Fa
Sol
La
Ti

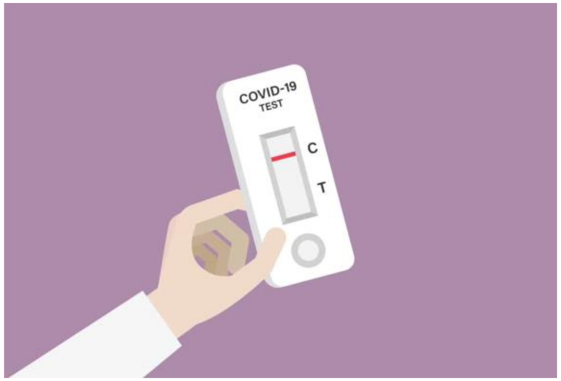
Musical tone recognition

Examples of *Binary Classifiers* (*K* = 2)

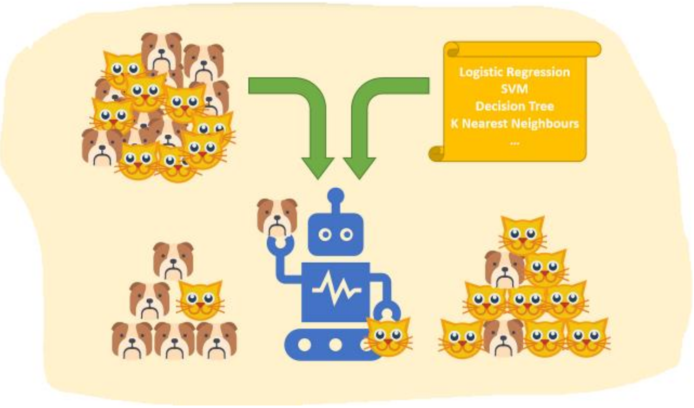
Pregnancy tester

Spam filtering

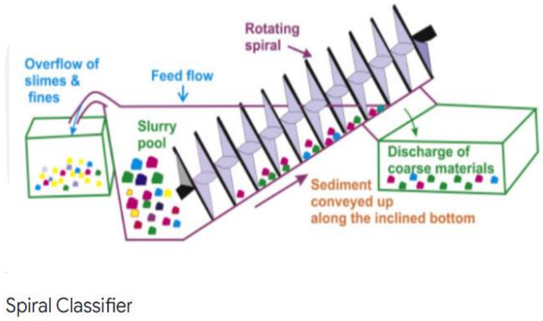
Counting machine: bank notes are classified as forged or accepted



Covid - 19 tester



Classification algorithm


Rotating spiral
Overflow of slimes & fines
Feed flow
Slurry pool
Discharge of coarse materials
Sediment conveyed up along the inclined bottom
Spiral Classifier

Spiral classifier

Ability of *Classifier* (General Case)

The conditional probabilities that an object will be classified as category ***k***, given that its actual/true category is ***i*** - $P_{k|i}$

$$1 \le i, k \le K; \qquad \sum_{k=1}^{K} P_{k|i} = 1$$

Ideal classifier: $every\ P_{k|i} = 0, except\ of\ P_{i|i}$

Classification (Confusion) Matrix and Repeatability for the General Case of K Categories

$$\boldsymbol{P} = \begin{pmatrix} p_{1|1} & p_{2|1} & \cdots & p_{k|1} & \cdots & p_{K|1} \\ p_{1|2} & p_{2|2} & \cdots & p_{k|2} & \cdots & p_{K|2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{1|i} & p_{2|i} & \cdots & p_{k|i} & \cdots & p_{K|i} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{1|K} & p_{2|K} & \cdots & p_{k|K} & \cdots & p_{K|K} \end{pmatrix} \qquad \boldsymbol{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix}$$

$$\boldsymbol{Repeatability} = \begin{bmatrix} Repeatability_1 \\ Repeatability_2 \\ \cdots \\ Repeatability_i \\ \cdots \\ Repeatability_K \end{bmatrix}$$ - (for the **same** classifier)

$$Repeatability_i = \frac{K}{K-1} \sum_{k=1}^{K} \left[p_{k|i} \cdot \left(1 - p_{k|i}\right) \right] = \frac{K}{K-1} \left(1 - \sum_{k=1}^{K} \left[p_{k|i} \right]\text{^}2\right) = VAR_{within}$$

$$0 \leq VAR_{within} \leq 1$$

The *Closeness of Agreement* between classifications obtained by <u>different</u> *Classifiers* participating in collaborative study (from here on - *Classifiers effect*)

$$Classifiers\ effect = \begin{bmatrix} Classifier\ effect_1 \\ Classifier\ effect_2 \\ \dots \\ Classifier\ effect_i \\ \dots \\ Classifier\ effect_K \end{bmatrix}$$

Where:

$$Classifiers\ effect_i = \frac{K}{K-1} \sum_{k=1}^{K} VAR\left(p_{k|i}\right)$$

$VAR\ \left(p_{k|i}\right)$ = classic variation of $p_{k|i}$ between collaborators/classifiers

According to **ISO 5725:2023 "Accuracy (trueness and precision of measurement methods and results)"** (differ from that in the Gauge R&R studies): Reproducibility = Repeatability + $Classifier\ effect$

Reproducibility describes the **total precision** of classification procedure (i.e., categorical measurements)

H_0 - *Homogeneity Hypothesis: Acceptance and Rejecting* (for the fixed group of classifiers)

H_0 : all classifiers' counts are drawn from the same population characterized by vector ($p_1, p_2, \ldots, p_k, \ldots p_K$).

If *SP* (segregation/separation power) statistics is less than some critical value,

H_0 is not rejected. Otherwise, see (*) for unbiased estimators of variation components, i.e.: *repeatability, classifier's effect* and *total precision*.

*****T. Gadrich, E. Bashkansky, I. Kuselman (2013) "Comparison of biased and unbiased estimators of variances of qualitative and semi-quantitative results of testing", *Accreditation and Quality Assurance*, **18** (2), 85-90

*T. Gadrich, E. Bashkansky, R. Zitikis, (2015) "Assessing variation: a unifying approach for all scales of measurement", *Quality and Quantity*, **49** (3) , 1145-1167

Random *Dirichlet* - *Multinomial (DM)* Model - 1

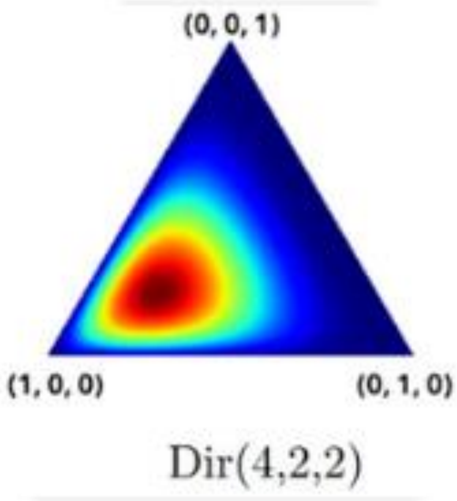
❑ ***L*** classifiers are randomly sampled from the population which classification abilities related to category ***i*** are distributed according to the Dirichlet distribution:

$$f\left(p_{1|i}, p_{2|i}, \cdots, p_{K|i}\right) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_{k|i}{}^{\alpha_{k|i}-1} \qquad \left(\sum_{k=1}^{K} p_{k|i} = 1 \right)$$

where: $\boldsymbol{\alpha_i} = \left(\alpha_{1|i}, \alpha_{2|i}, \cdots, \alpha_{K|i}\right)$ - *parameters of the Dirichlet distribution characterizing the i-th category classification;*

$E\ (p_{k|i}) = \alpha_{k|i} / \ \alpha_{0|i}$;

$\alpha_{0|i} = \sum_{k=1}^{K} \alpha_{k|i}$ - *concentration parameter* (*)



* Location is determined by repeatability, and dispersion ($\alpha_{0|i}$) is determined by degree of variation between classifiers.

Repeatability (*within*) and *Classifiers* (*between-classifiers*) *effect*

$$\boldsymbol{Repeatability_i} = \frac{K}{K-1}\sum_{k=1}^{K} E\left[p_{k|i}\cdot\left(1-p_{k|i}\right)\right] = \frac{K}{K-1}\frac{\alpha_{0|i}}{\alpha_{0|i}+1}\left[1-\sum_{k=1}^{K}\frac{\alpha_{k|i}^2}{\alpha_{0|i}^2}\right]$$

$$\boldsymbol{Classifier\ effect_i} = \frac{K}{K-1}\sum_{k=1}^{K} VAR\left(p_{k|i}\right) = \frac{K}{K-1}\frac{1}{\alpha_{0|i}+1}\left[1-\sum_{k=1}^{K}\frac{\alpha_{k|i}^2}{\alpha_{0|i}^2}\right] = \frac{\boldsymbol{Repeatability}}{\alpha_{0|i}}$$

$$\boldsymbol{Reproducibility_i} = \frac{K}{K-1}\left[1-\sum_{k=1}^{K}\frac{\alpha_{k|i}^2}{\alpha_{0|i}^2}\right]$$

$\alpha_{0\|i} \rightarrow 0$	$\alpha_{0\|i} \rightarrow \infty$
Repeatability = 0	Classifiers effect =0
$\boldsymbol{Classifiers\ effect_i} = \frac{K}{K-1}\left[1-\sum_{k=1}^{K}\frac{\alpha_{k\|i}^2}{\alpha_{0\|i}^2}\right]$	$\boldsymbol{Repeatability_i} = \frac{K}{K-1}\left[1-\sum_{k=1}^{K}\frac{\alpha_{k\|i}^2}{\alpha_{0\|i}^2}\right]$

Random *Dirichlet* - *Multinomial* (*DM*) Model - 2

- ❑ From an experiment consisting of *N* **repeated classifications** of each property that actually belongs to the category *i*

$$\left(n^{(l)}_{1|i}, n^{(l)}_{2|i}, \dots, n^{(l)}_{K|i} \middle| \boldsymbol{p} = \left(p_{1|i}, p_{2|i},, \dots, p_{K|i}\right)\right) \sim Multinomial(N, \boldsymbol{p} = \left(p_{1|i}, p_{2|i},, \dots, p_{K|i}\right))$$

where $\boldsymbol{p} = \left(p_{1|i}, p_{2|i},, \dots, p_{K|i}\right) \sim Dirichlet(\boldsymbol{\alpha_i})$

Estimator of classification matrix of the *l*-th classifier $(l = 1,2, \dots, L)$.

$$\hat{P}^{(l)} = \begin{pmatrix} \hat{p}^{(l)}_{1|1} & \hat{p}^{(l)}_{2|1} & \cdots & \hat{p}^{(l)}_{k|1} & \cdots & \hat{p}^{(l)}_{K|1} \\ \hat{p}^{(l)}_{1|2} & \hat{p}^{(l)}_{2|2} & \cdots & \hat{p}^{(l)}_{k|2} & \cdots & \hat{p}^{(l)}_{K|2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \hat{p}^{(l)}_{1|i} & \hat{p}^{(l)}_{2|i} & \cdots & \hat{p}^{(l)}_{k|i} & \cdots & \hat{p}^{(l)}_{K|i} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \hat{p}^{(l)}_{1|K} & \hat{p}^{(l)}_{2|K} & \cdots & \hat{p}^{(l)}_{k|K} & \cdots & \hat{p}^{(l)}_{K|K} \end{pmatrix} \quad \text{where } \hat{p}^{(l)}_{k|i} = \frac{n^{(l)}_{k|i}}{N}$$

Unbiased Estimators

Unbiased estimator for repeatability (within) variation:

$$\widehat{\boldsymbol{Repeatability}}_{\boldsymbol{u(i)}} = \frac{N}{N-1}\frac{1}{L}\sum_{l=1}^{L} \widehat{Repeatability}_i =$$

$$= \frac{N}{N-1}\frac{1}{L}\sum_{l=1}^{L} \frac{K}{(K-1)}\sum_{k=1}^{K} \hat{p}_{k|i}^{(l)} \cdot \left(1 - \hat{p}_{k|i}^{(l)}\right)$$

Unbiased estimator for between classifiers variation:

$$\widehat{\boldsymbol{Classifier\ effect}}_{\boldsymbol{u(i)}} = \frac{K}{K-1}\sum_{k=1}^{K} \frac{1}{L-1}\sum_{l=1}^{L} \left(\hat{p}_{k|i}^{(l)} - \frac{1}{L}\sum_{l=1}^{L} \hat{p}_{k|i}^{(l)}\right)^2$$

$$- \frac{\widehat{Repeatability}_{u(i)}}{N}$$

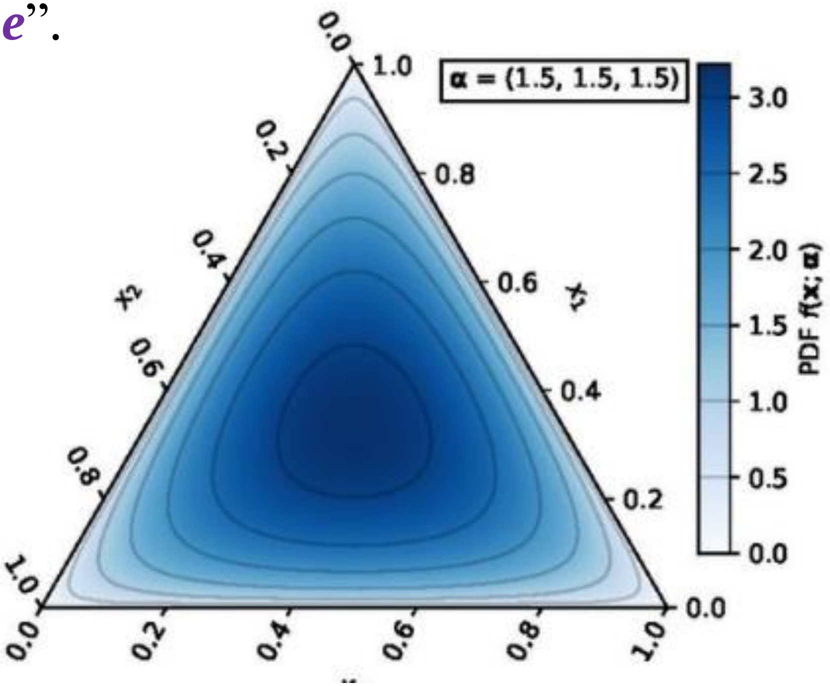
Unbiased estimator for total classification variation:

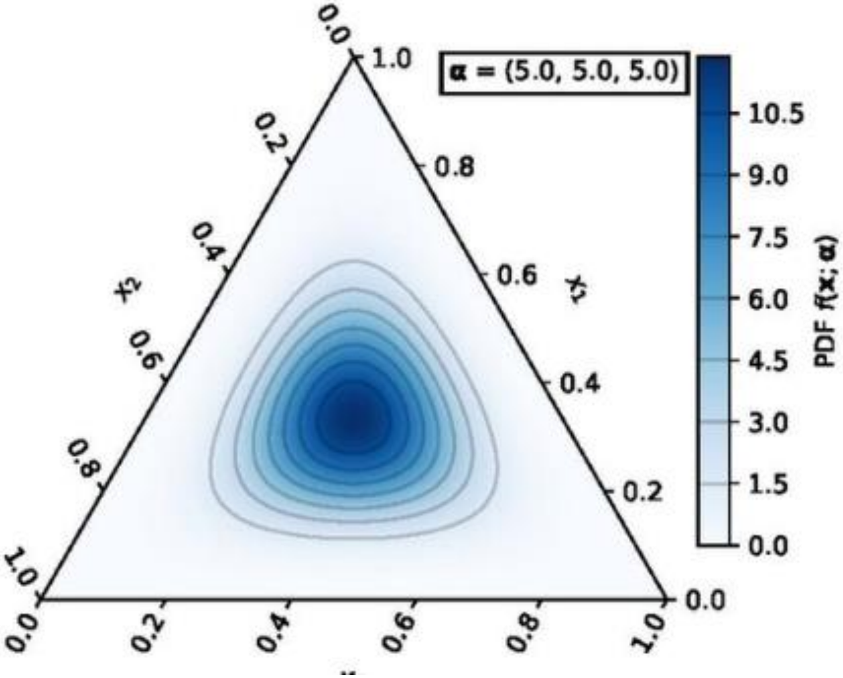
$$\widehat{VAR}_{TOTAL} = \frac{K}{K-1}\sum_{k=1}^{K} \frac{1}{L-1}\sum_{l=1}^{L} \left(\hat{p}_{k|i}^{(l)} - \frac{1}{L}\sum_{l=1}^{L} \hat{p}_{k|i}^{(l)}\right)^2 +$$

$$+ \frac{1}{L}\sum_{l=1}^{L} \frac{K}{K-1}\sum_{k=1}^{K} \hat{p}_{k|i}^{(l)} \cdot \left(1 - \hat{p}_{k|i}^{(l)}\right)$$

How to estimate vector $\boldsymbol{\alpha_i} = \left(\alpha_{1|i}, \alpha_{2|i}, \cdots, \alpha_{K|i}\right)$, characterizing *Population Distribution* of classification abilities?

It is considered that the key parameter is $\alpha_{0|i}$, characterizing the so called "***Dirichlet noise***".





We propose to estimate $\alpha_{0|i}$ by help of Repeatability to *Classifiers effect* ratio:

$$\hat{\alpha}_{0|i} = \frac{\text{Repeatability}_{u(i)}}{Classifier\ effect_{u(i)}} \qquad \text{and then} \qquad \hat{\alpha}_{k|i} = \hat{\alpha}_{0|i}\frac{1}{L}\sum_{L=1}^{L}\hat{p}_{k|i}^{(l)}$$

Simulation Study for $K = 3$

The population was $\boldsymbol{p} = (0.8200, 0.1667, 0.0133)$ (*)

Accordingly, we simulated for different $\boldsymbol{\alpha}$ and α_0:

- $\boldsymbol{\alpha} = (2.46, 0.5, 0.04)$, i.e., $\alpha_0 = 3$
- $\boldsymbol{\alpha} = (8.2, 1.667, 0.133)$, i.e., $\alpha_0 = 10$
- $\boldsymbol{\alpha} = (16.4, 3.333, 0.267)$, i.e., $\alpha_0 = 20$
- $\boldsymbol{\alpha} = (24.6, 5.0, 0.4)$, i.e., $\alpha_0 = 30$

While

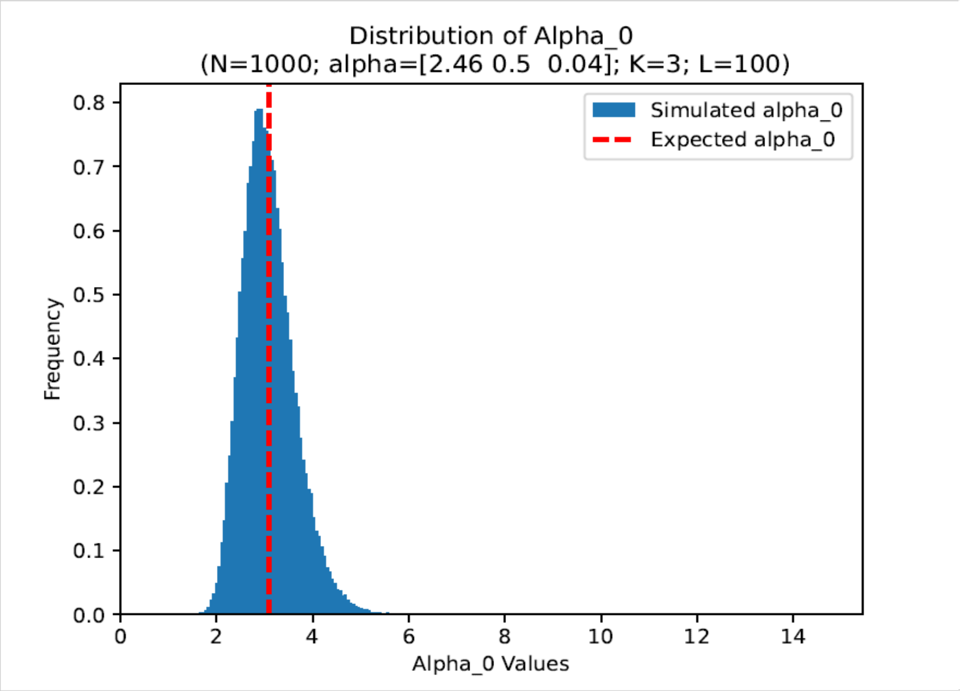
- L = 5, 10, 20, 30 or 100
- N = 1000

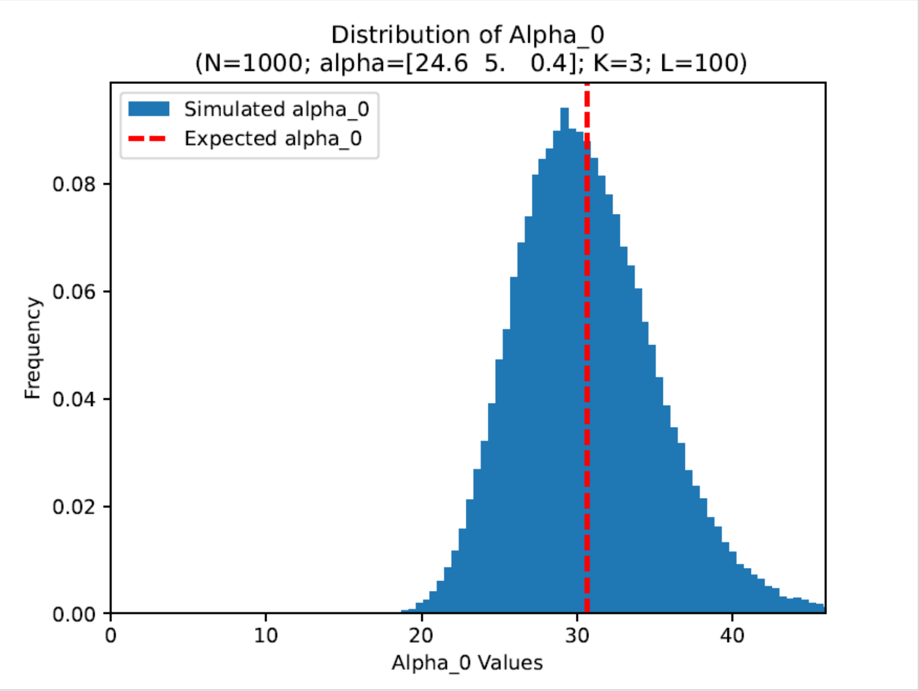
* E. Bashkansky, S. Dror, R. Ravid, P. Grabov (2007) “Effectiveness of a Product Quality Classifier”, *Quality Engineering*, vol. 19, issue 3, pp.235-244

Influence of α_0 magnitude

$\alpha_0 = 3$

$\alpha_0 = 30$

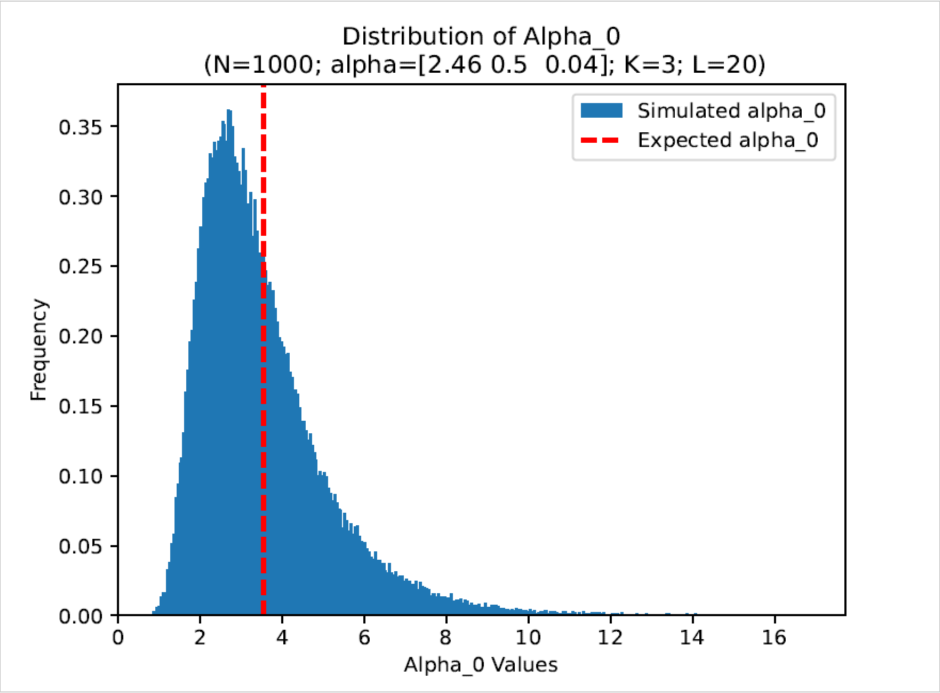


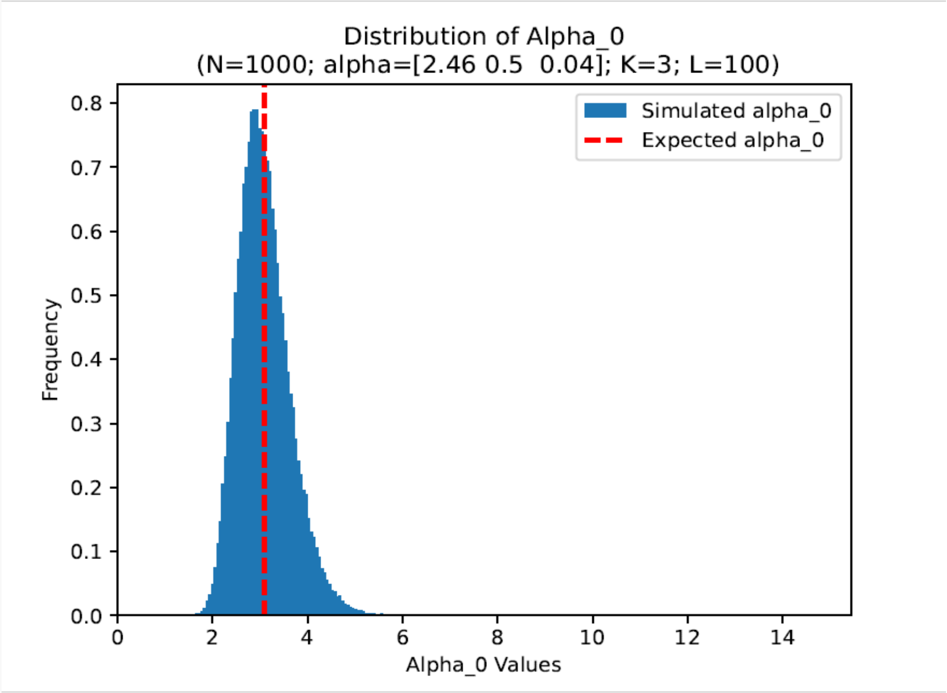


Influence of the number of *classifiers* - *L*

$L = 20$

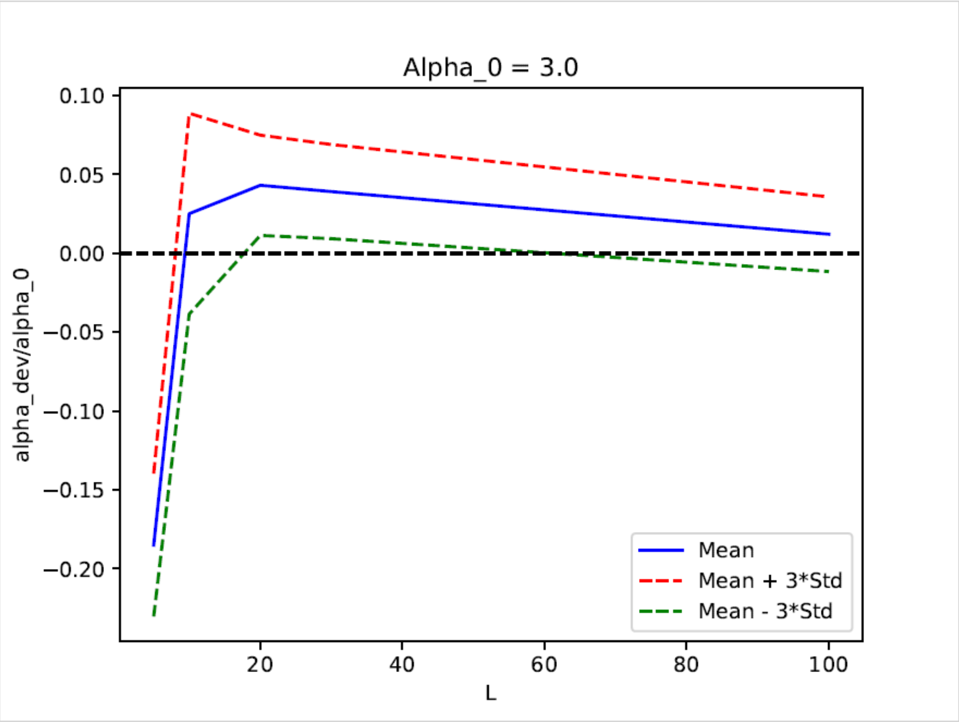
$L = 100$

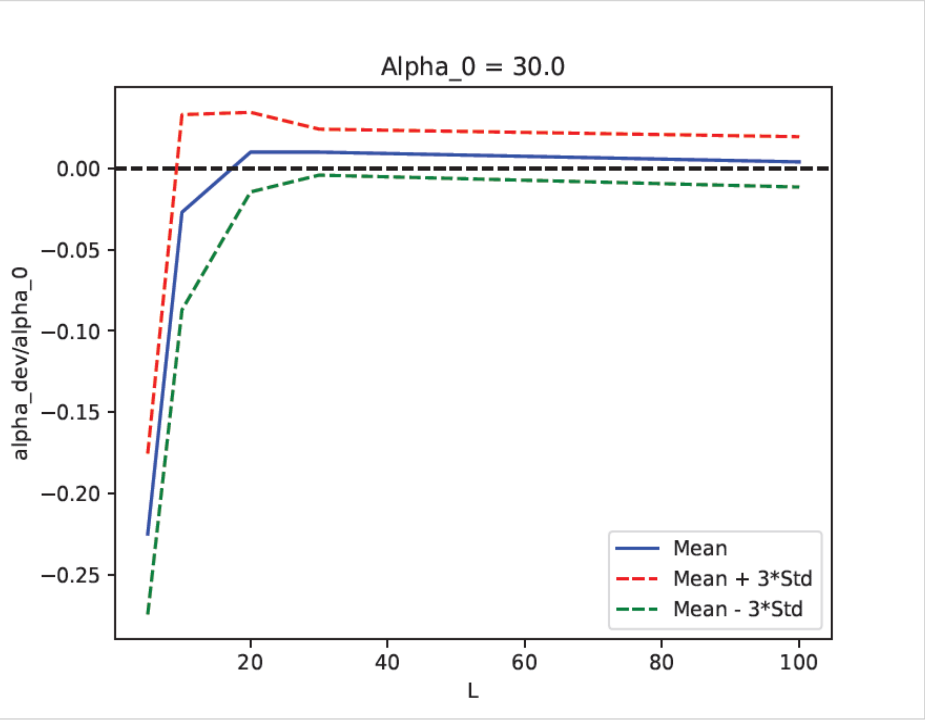




$\alpha_0 = 3$

How the number of *classifiers* L influences the *relative deviation* of α_0 estimator





L = 30 brings quite good results!

Summary

1. Classification precision is particularly crucial in scenarios where the cost of false output is high, e.g. medical diagnosis, search engine results, product quality control etc.
2. A statistical model for analyzing classification's precision and its *intra* and *inter* components from collaborative studies is presented.
3. Unbiased estimators of repeatability, classifiers' effect and total precision are provided for both the fixed and the random model.
4. Using simulation, it was found that the greatest influence on the accuracy and effectiveness of the model estimations in a random model is exerted mainly by the number of classifiers.

Thank You for Your Attention!

E-mail: *myariv@braude.ac.il*