Comparing two bootstrapped regions in images: the D-test

Florentin Kucharczak, Inés Couso, Olivier Strauss, Denis Mariano-Goulart

Comparing two bootstrapped regions in images: the D-test.

Florentin Kucharczak[a,b,*], Inés Couso[c], Olivier Strauss[a], Denis Mariano-Goulart[b,d]

[a]LIRMM, Univ. Montpellier, 161 Rue Ada, 34095 Montpellier, France
[b]Department of Nuclear Medicine, Montpellier University Hospital, 371 Av. du Doyen Gaston Giraud, 34090 Montpellier, France
[c]Department of Statistics and OR, Univ. Oviedo, Luis Moya Blanco, 261, 33203 Gijón, Spain
[d]PhyMedExp, Univ. Montpellier, INSERM U1046, CNRS UMR 9214, 371 Avenue du Doyen G. Giraud, 34295 Montpellier, France

Abstract

Objectives: Many molecular imaging diagnoses involve comparing two regions of interest (ROIs) in the image or different images. Since the images are obtained by measuring a random phenomenon, such comparisons should be based on a statistical test to ensure reliability. Recent studies have shown that use of the bootstrap approach provides access to the statistical variability of reconstructed values in molecular images. However, although there is general agreement that this increase in information should make diagnosis based on molecular images more reliable, no approach has been proposed in the relevant literature to use bootstrap replicates to enhance the reliability of comparisons of two ROIs. In this paper, we propose to fill this gap by introducing the first statistical test that allows us to compare two sets of pixels/voxels for which bootstrap replicates are available.
Material and methods : After presenting the theoretical basis of this non-parametric statistical test, this article describes how to calculate it in practice. Finally, it proposes two experiments based on quantitative comparisons and expert judgment to assess its relevance.
Results : The results obtained are consistent with expert diagnosis on synthetic data. This validates the relevance of the D-test.
Conclusion : This paper presents the first statistical test to compare two ROIs in reconstructed images for which the statistical variability information is accessible.

Keywords:
Statistical test, ROI comparison, statistical variability, bootstrap replicates, positron emission tomography (PET).

[*]Corresponding author: Tel.: +336-17-55-58-94;
e-mail: florentin.kucharczak@chu-montpellier.fr;

1. Introduction

This type of comparison is usually visually performed by a physician expert. S/he can be assisted by semi-quantitative ROI description metrics such as the widely used standard uptake value (SUV) that corresponds to the ratio of the image-derived radiotracer concentration and the whole body injected radioactivity concentration (1). However, the reliability of this metric for comparison is poor and its limits are well documented (2; 3; 4; 5).

For organ specific tasks such as brain ROI comparison, some quantitative analysis software packages like Scenium (6) are provided by the positron emission tomography (PET) manufacturer. Mean SUV are measured in each cortical ROI and normalized to the mean whole brain SUV to produce cortical SUV ratios (SUVr). Lastly, these cortical SUVr are converted to standard deviation scores (SUVr SD), commonly called z-score, based on reference distributions from age-matched control populations (7). However, this comparison requires a database, which means that it is dependent on numerous parameters such as the reconstruction method used, the tomograph version, the ethnic origin of controls, etc. Thus, it is essential for the development of reliable statistical tools to help physicians compare two regions of interest and thus to ensure more reliable diagnosis (9; 10; 11).

For such a statistical tool, the spatial distribution in the ROI would have to be representative of the statistical distribution of each pixel, i.e. the phenomenon should be somewhat ergodic, or the prior distribution in each pixel should be known. Since the images used in nuclear medicine are obtained through a reconstruction process, the statistical properties of each reconstructed value is unknown and ergodicity cannot be hypothesized, although the statistics of the original measurements are known. Bootstrap-based methods have been proposed for quantifying the uncertainty associated with reconstructed activity values in reconstructed images (12; 13; 14). Bootstrap-based variability quantification is currently considered to be the ground truth approach. It consists of drawing repeated samples from the original data for generating samples of the reconstructed values. Those samples, also called replicates, are considered as being samples of the distribution underlying the reconstructed data. Bootstrap replicates have been shown to accurately estimate image noise for various reconstruction algorithms (13; 15; 8). The main limitation of bootstrap approaches could be the computation time. However, current computing capacities have overcome this limitation. Surprisingly, even though this approach is considered to be the gold standard method for statistical variability quantification, no direct application of the reconstructed replicates to improve the reliability of ROI comparison has been proposed in the relevant literature to our knowledge.

Here we aim to fill this gap by proposing a statistical method dedicated to the comparison of ROIs in images whose replicates are available by either bootstrap or repeated acquisitions/ reconstructions. We propose to perform this comparison using a statistical test, i.e. the *drop-out test*. The idea behind this test is as follows. Two regions R_X and R_Y will be considered different if, by randomly drawing values from both regions, swapping the values of the R_X region with those of the R_Y region causes discernible changes in the distribution of values of the host region. This test is based on a concept that is not often used in statistics, i.e. concentration intervals. Section 2 details the background of the study. Section 3 develops the theoretical basis and justifications of the drop-out test (D-Test). The use of the D-test assumes the stationarity of the regions to be compared, i.e. that any pixel/voxel is interchangeable in the same region. In practice, this assumption is seldom verified. Section 4 thus presents a modified version of the D-test, i.e. the so-called KD-test. Section 5 details the construction of the test and proposes a way to compute it in practice. Section 6 illustrates the performance of the D-Test and KD-Test with an experiment with simulated PET data.

2. Background

2.1. Notations

- $\mathbb{R}$ is the set of real values.
- $\mathbb{IR}$ is the set of bounded and closed real intervals.
- Let $I \in \mathbb{IR}$, then $\underline{I}$ (rsp. $\overline{I}$) denotes its lower bound (rsp. upper bound). Thus $I = [\underline{I}, \overline{I}]$.
- Let $I \in \mathbb{IR}$, then $|I| = \overline{I} - \underline{I}$ being the length of I.
- Let S be a random variable, with P_S denoting its induced probability measure.
- Given $\beta \in [0, 1]$, then a β–concentration interval $I_\beta \in \mathbb{IR}$ will be an interval satisfying:
  - $P_S(I_\beta) \geq \beta$
  - $\forall I \in \mathbb{IR}$ such that $P_S(I) \geq \beta$, $|I_\beta| \leq |I|$.

  Hence, a β–concentration interval covers the value of S with probability of at least β while having a minimal length among all intervals meeting that condition.

2.2. Materials and methods

Let R_X and R_Y be the two ROI to be compared. ROI R_X (resp. R_Y) consists of n_x (resp. n_y) pixels. If only one measurement is available, then one measured value $x_i \in \mathbb{R}$ (resp. $y_i \in \mathbb{R}$) is associated with the i^{th} pixel of ROI X (resp. ROI Y). We denote $\boldsymbol{x} = (x_1, \ldots, x_{n_x})$ and $\boldsymbol{y} = (y_1, \ldots, y_{n_y})$. From a resampling process, we obtain m replicates of the corresponding tuples in each region. Thus, ROI R_X (resp. ROI R_Y) is associated with m tuples ${}^t\boldsymbol{x} = ({}^tx_1, \ldots, {}^t x_{n_x})_{t=1\ldots m}$ (resp. $({}^ty_1, \ldots, {}^t y_{n_y})_{t=1\ldots m}$) of bootstrapped values.

3. Drop-out test (D-test)

3.1. Formal statement of the problem

Each pixel is associated with a numerical quantity that, due to the process adopted for its measurement, can be considered random. Thus, each pixel $j \in \{1, \ldots, n_x\}$ of ROI R_X is identified with a random variable X_j, $j = 1, \ldots, n_x$ which induces the distribution P_j. Similarly, each of the n_y pixels of ROI R_Y is identified with a random variable Y_j, $j = 1, \ldots, n_y$, thus inducing the distribution Q_j.

Subsequently, we can consider a random variable X resulting from choosing a pixel j at random (from among n_x pixels, using a discrete uniform distribution) and subsequently observing a value of the corresponding random variable X_j (analogously, we can consider a random variable Y, combining the choice of a pixel j among the n_y pixels of the second ROI and an observation of the corresponding variable Y_j). Finally we search for a suitable procedure to test the null hypothesis that the probability distribution of X (P) coincides with that of Y (Q), against the alternative that they do not coincide. Readers should note that the probability distribution of X can be understood as the convex linear combination of n_x distributions associated with the corresponding pixels, that is $P = \frac{P_1+\ldots+P_{n_x}}{n_x}$. Analogously, $Q = \frac{Q_1+\ldots+Q_{n_y}}{n_y}$. On the other hand, the distributions $P_1, \ldots, P_{n_x}$ do not necessarily coincide with each other (and similarly, we also do not assume that $Q_1 = \ldots = Q_{n_y}$). Our ultimate goal is to determine whether or not the probability distribution of the numerical observations in the first region matches that of the second region, without hypothesizing or trying to infer anything additional about the distribution within the various pixels that make up each region.

As explained in the previous section, we start from a sample of m independent tuples of n_x-dimensional (resp. n_y-dimensional) vectors.

3.2. Proposed family of tests

First, we choose a threshold $\beta > 0$ and, for each pixel, we randomly take one of the m observations and use the remaining $m-1$ observations to construct an estimate of a concentration interval at the β level, according to the estimation procedure described in Subsection 5.1. The corresponding concentration intervals are denoted $A_1, \ldots, A_{n_x}$ (for the n_x pixels of the first ROI), $B_1, \ldots, B_{n_y}$ (for the n_y pixels of the second ROI), and are fixed. For technical reasons, we assume that these n_x and n_y pixels are part of larger regions in each of which arbitrarily large sequences of n and n' pixels could have been respectively chosen. Let the remaining (randomly selected from each pixel) observations be respectively denoted $X_1, \ldots X_n, \ldots, Y_1, \ldots, Y_{n'}, \ldots$. For every $i = 1, \ldots, n_x$ and every $j = 1, \ldots, n, \ldots$ we define a random (binary) variable δ^i_j as follows:

$$\delta^i_j = \begin{cases} 1 & \text{if } X_j \in A_i \\ 0 & \text{otherwise.} \end{cases}$$

Analogously, for every $i = 1, \ldots, n_x$ and every $j = 1, \ldots, n', \ldots$, we define $(\delta^i_j)'$ as follows:

$$(\delta^i_j)' = \begin{cases} 1 & \text{if } Y_j \in A_i \\ 0 & \text{otherwise.} \end{cases}$$

Let us now consider, for each $j = 1, \ldots, n, \ldots$, the random variable N_j representing the number of concentration intervals A_i (each of them constructed from each pixel of region R_X) that contain the value of the random variable X_j. Mathematically:

$$N_j = \#\{i = 1, \ldots, n_x \,:\, A_i \ni X_j\} = \delta^1_j + \ldots \delta^{n_x}_j.$$

Analogously, for every $j = 1, \ldots, n', \ldots$, let N'_j denote the number of those concentration intervals (also from R_X) containing Y_j, that is:

$$M_j = \#\{i = 1, \ldots, n_x \,:\, A_i \ni Y_j\} = (\delta^1_j)' + \ldots (\delta^{n_x}_j)'.$$

Let us also consider the following random variables for each $i, k = 1, \ldots, n_x$, any n and n':

- $N^i = \#\{j \in \{1, \ldots, n\} : X_j \in A_i\}$,
- $M^i = \#\{j \in \{1, \ldots, n'\} : Y_j \in A_i\}$.
- $N^{i,k} = \#\{j \in \{1, \ldots, n\} : X_j \in A_i \cap A_k\}$,
- $M^{i,k} = \#\{j \in \{1, \ldots, n'\} : Y_j \in A_i \cap A_k\}$,

where we implicitly assume the convention $N^{i,i} = N^i$ and $M^{i,i} = M^i$ for all $i = 1, \ldots, n_x$. We also consider the following notation, for every pair of indices $i, k = 1, \ldots, n_x$:

$$\hat{p}^i = \frac{N^i}{n},\ \hat{q}^i = \frac{M^i}{n},\ \hat{p}^{i,k} = \frac{N^{i,k}}{n},\ \hat{q}^{i,k} = \frac{M^{i,k}}{n'}. \tag{1}$$

Based on these constructions, let us consider the following averages:

$$\overline{N} = \frac{\sum_{j=1}^{n} N_j}{n} = \frac{\sum_{i=1}^{n_x} N^i}{n} \text{ and } \overline{M} = \frac{\sum_{j=1}^{n'} M_j}{n'} = \frac{\sum_{i=1}^{n_x} M^i}{n'}.$$

Remark 1. *Let us recall that the random variables δ_j^i and $(\delta_j^i)'$ respectively follow a Bernoulli distribution with success parameter $p_j^i = P_j(A_i)$ and $q_j^i = Q_j(A_i)$, $i = 1, \ldots, n_x$. Their respective expectations and variances are:*

- $E(\delta_j^i) = p_j^i$, $V(\delta_j^i) = p_j^i \cdot (1 - p_j^i)$, *and*
- $E((\delta_j^i)') = p_j^i$, $V((\delta_j^i)') = q_j^i \cdot (1 - q_j^i)$.

Furthermore, the covariance between two of those variables can be calculated as follows:

$$Cov(\delta_j^i, \delta_j^k) = E(\delta_j^i \cdot \delta_j^k) - E(\delta_j^i) \cdot E(\delta_j^k) = \\ P_j(A_i \cap A_k) - P_j(A_i) \cdot P_j(A_k) = p_j^{i,k} - p_j^i \cdot p_j^k.$$

And analogously,

$$Cov((\delta_j^i)', (\delta_j^k)') = q_j^{i,k} - q_j^i \cdot q_j^k.$$

Thus, $N_j = \delta_j^1 + \ldots + \delta_j^{n_x}$ is the sum of n_x (not necessarily independent) Bernoulli random variables and its expectation and variance are respectively:

- $E(N_j) = \sum_{i=1}^{n_x} p_j^i$ *and*
- $V(N_j) = \sum_{i=1}^{n_x} V(\delta_j^i) + \sum_{k \neq i} Cov(\delta_j^i, \delta_j^k)$.

Thus, the variance of N_j can be easily expressed in terms of the probability distribution P_j, and more concretely in terms of the probabilities $P_j(A_i)$ and $P_j(A_i \cap A_k)$, $i = 1, \ldots, n_x$, $k = 1, \ldots, n_x$. Something analogous can be said about the distribution of each of the random variables $N_j' = (\delta_j^1)' + \ldots (\delta_j^{n_x})'$, $j = 1, \ldots, n_y$, this time with respect to the probability distribution Q_j.

For the sake of concision and readability, hereafter we will use the symbols μ_j and σ_j^2 to denote the expectation and variance of N_j (respectively μ_j' and $(\sigma_j')^2$ to denote the mean and variance of N_j'). We will furthermore denote $\overline{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mu_j$ and $\overline{\mu}' = \frac{1}{n'} \sum_{i=1}^{n'} \mu_j'$.

Remark 2. *Hereafter, we will consider sequences of random variables* $(\overline{N}_n, T_n, D_n, K_n$ *etc) and parameters* $(\sigma_n, c_n, d_n$ *etc), all indexed by* n*. Furthermore, to involve a single index in the corresponding sequences, we will assume that there is a fixed natural value* r *such that either* $n' = n + r$*, or* $n = n' + r$*. From a theoretical viewpoint, we need to refer to numerable sequences of variables and parameters, since we intend to determine the asymptotic distributions of certain statistical sequences in order to be able to determine the asymptotic size of the D-test and the KD-test to be respectively defined in Sections 3 and 4. However, in practice, when using these tests in concrete situations, we will assume that the numbers of pixels,* n_x *and* n_y*, observed respectively in regions* R_x *and* R_y*, are large enough to approximate the distribution of the corresponding statistics by such an asymptotic distribution. In such cases, we will implicitly identify,* n_x *and* n_y *with sufficiently large values of* n *and* n' *respectively. In cases where it does not give rise to confusion, we will avoid the use of the subscript* n *in order to simplify the notation. Thus, as long as there is no confusion, we will use the notation* $\overline{N}, T, D, K$*, etc. instead of* $\overline{N}_n, T_n, D_n, K_n$*, etc.*

We aim to test the null hypothesis $H_0 : P = Q$ against the alternative hypothesis $H_1 : P \neq Q$. Hence, we propose a nested family of tests, $(DT_\alpha)_{\alpha \in (0,1)}$, each based on the comparison of between the absolute value of a certain statistic and the standard normal $1 - \frac{\alpha}{2}$-quantile.
Specifically, the statistic is calculated as follows:

Definition 1. *The* D-statistic *is defined as follows:*

$$D = \frac{\overline{N} - \overline{M}}{\sqrt{T}}, \textit{ with} \tag{2}$$

$$T = \frac{1}{n} \sum_{i=1}^{n_x} \sum_{k=1}^{n_x} (\hat{p}^{i,k} - \hat{p}^i \cdot \hat{p}^k) + \frac{1}{n'} \sum_{i=1}^{n_x} \sum_{k=1}^{n_x} (\hat{q}^{i,k} - \hat{q}^i \cdot \hat{q}^k)$$

Based on the calculation of the absolute value of the D-statistic, we propose the following decision procedure to test $H_0 : P = Q$ against $H_1 : P \neq Q$:

Definition 2. *Let us consider the test of the null hypothesis* $H_0 : P = Q$ *against the alternative hypothesis* $H_1 : P \neq Q$*.* D-test *is the procedure that consists of rejecting* H_0 *when* $|D| > z_{1-\frac{\alpha}{2}}$*, and not rejecting it otherwise.*

In the next subsection, we will prove that the asymptotic size of each test in the above family is not α, but $f(\alpha) < \alpha$, where $f : (0,1) \to (0,1)$ is a strictly increasing function. Furthermore, we will show that the gap between α and $f(\alpha)$ depends on the degree of difference between the distributions over the different pixels in the same ROI, i.e. the extent to which each region can be considered stationary. The more similar those distributions then the less the difference $\alpha - f(\alpha)$.
In particular, if $P_1 = \ldots = P_{n_x} = P$ and $Q_1 = \ldots = Q_{n_y} = Q$ (regardless of the degree of resemblance between P (region R_X) and Q (region R_Y)) the size $f(\alpha)$ will coincide with α, for every α. In other words, in this particular situation, the mapping f is the identity function. In the general case, the difference $\alpha - f(\alpha)$ will depend on a pair of unknown population parameters that quantify those differences inside regions R_X and R_Y.

Such parameters are respectively:

$$\theta_1 = \sum_{j=1}^{n_x} \sum_{j'=j+1}^{n_x} \big[E(N_j) - E(N_{j'})\big]^2$$

and

$$\theta_2 = \sum_{j=1}^{n_y} \sum_{j'=j+1}^{n_y} \big[E(M_j) - E(M_{j'})\big]^2.$$

The greater their values, the greater the difference between α and $f(\alpha)$. In particular, when both parameters are 0 (which happens when the distributions over the n_x pixels of R_x are similar, and also when the distributions over the n_y pixels of R_y are similar), the mapping f coincides with the identity.

3.3. Asymptotic distribution of the D-test statistic

In this section, we will prove that, for each $\alpha \in (0,1)$, the asymptotic distribution of the statistic of Equation (2) is that of the standard normal. This requires some preliminary results.

Lemma 1. *Suppose the following conditions are satisfied:*

(a) The sums of covariances $\sum_{j=1}^{n} \sum_{k \neq i} (p_j^{ik} - p_j^i \cdot p_j^k)$ and $\sum_{j=1}^{n'} \sum_{k \neq i} (q_j^{ik} - q_j^i \cdot q_j^k)$ are non-negative.

(b) $\epsilon > 0$ and at least one index $i = 1, \ldots, n_x$ such that $p_j^i = P_j(A_i) > \epsilon$, for every $j = 1, \ldots, n, \ldots$

Consider an arbitrary but fixed $r \in \mathbb{N} \cup \{0\}$. Suppose that either $n' = n + r$ or $n = n' + r$. Then the sequence of random variables whose n^{th} term is:

$$\frac{(\overline{N} - \overline{M}) - (\overline{\mu} - \overline{\mu}')}{\sigma_n} \tag{3}$$

where $\sigma_n = \sqrt{\dfrac{1}{n^2} \displaystyle\sum_{j=1}^{n} \sigma_j^2 + \dfrac{1}{(n')^2} \sum_{j=1}^{n'} (\sigma_j')^2}$ *converges in law to a standard normal.*

Proof: *According to our experiment (one measurement taken at random in each pixel, independently of the remaining pixels), the random variables in the sequence $N_1, \ldots, N_n, \ldots, M_1, \ldots, M_{n'}, \ldots$ are independent and thus are the quotients $\frac{N_1}{n}, \ldots, \frac{N_n}{n}, \ldots, \frac{M_1}{n'}, \ldots, \frac{M_n}{n'}, \ldots$. According to the central limit theorem, if such a sequence of independent variables satisfies the so-called Lyapunov condition, i.e. if:*

$$\lim_{\substack{n \to \infty \\ (n'=n+r)}} \frac{E\left[\frac{1}{n^3} \sum_{j=1}^{n} |N_j - \mu_j|^3 + \frac{1}{(n')^3} \sum_{j=1}^{n'} |M_j - \mu_j'|^3\right]}{\left(\frac{1}{n^2} \sum_{j=1}^{n} \sigma_j^2 + \frac{1}{(n')^2} \sum_{j=1}^{n'} (\sigma_j')^2\right)^{\frac{3}{2}}} = 0,$$

then the quotient

$$\frac{(\overline{N} - \overline{M}) - (\overline{\mu} - \overline{\mu}')}{\sigma_n}$$

converges in law to the standard normal distribution.

Regarding the Lyapunov condition, the numerator of the n term of the sequence is:

$$
\begin{aligned}
&E\left[\frac{1}{n^3}\sum_{j=1}^{n}|N_j-\mu_j|^3+\frac{1}{(n')^3}\sum_{j=1}^{n'}|M_j-\mu'_j|^3\right]\\
&=E\left[\frac{1}{n^3}\sum_{j=1}^{n}|\sum_{i=1}^{n_x}\left(\delta_j^i-p_j^i\right)|^3+\frac{1}{(n')^3}\sum_{j=1}^{n'}|\sum_{i=1}^{n_x}\left((\delta_j^i)'-q_j^i)\right)|^3\right]\\
&\leq\frac{1}{n^3}\sum_{j=1}^{n}\sum_{i=1}^{n_x}|E(\delta_j^i)-p_j^i|^3+\frac{1}{(n')^3}\sum_{j=1}^{n'}\sum_{i=1}^{n_x}|E((\delta_j^i)')-q_j^i|^3\\
&=\frac{1}{n^3}\sum_{j=1}^{n}\sum_{i=1}^{n_x}(1-p_j^i)^3p_j^i+(1-p_j^i)(p_j^i)^3+\frac{1}{(n')^3}\sum_{j=1}^{n'}\sum_{i=1}^{n_x}(1-q_j^i)^3q_j^i+(1-q_j^i)(q_j^i)^3\\
&\leq\frac{1}{n^3}\sum_{j=1}^{n}\sum_{i=1}^{n_x}(1-p_j^i)p_j^i+\frac{1}{(n')^3}\sum_{j=1}^{n}\sum_{i=1}^{n_x}(1-q_j^i)q_j^i
\end{aligned}
$$

The last inequality is obtained by taking out, as a common factor, the product $p_j^i(1-p_j^i)$ in the first n summands and the product $q_j^i(1-q_j^i)$ in the last n' summands, based on the fact that $(1-p)^2+p^2\leq(1-p)+p=1$ for all $p\in[0,1]$.
Let us now study the denominator of the Lyapunov quotient. To do so, we must check the variance of some statistics involved in our computations. The variance of each N_j is:

$$
\begin{aligned}
\sigma_j^2&=\sum_{i=1}^{n_x}V(\delta_j^i)+\sum_{k\neq i}Cov(\delta_j^i,\delta_j^k)=\sum_{i=1}^{n_x}\sum_{k=1}^{n_x}(p_j^{ik}-p_j^i\cdot p_j^k)\\
&=\sum_{i=1}^{n_x}\left(p_j^i(1-p_j^i)+\sum_{k\neq i}(p_j^{ik}-p_j^i\cdot p_j^k)\right)
\end{aligned}
$$

as previously calculated. Something similar applies to the variance of each M_j. Thus, the denominator of the Lyapunov quotient is:

$$
\left(\frac{1}{n^2}\sum_{j=1}^{n}\left[\sum_{i=1}^{n_x}\left(p_j^i(1-p_j^i)+\sum_{k\neq i}(p_j^{ik}-p_j^i\cdot p_j^k)\right)\right]+\right.
$$
$$
\left.\frac{1}{(n')^2}\sum_{j=1}^{n'}\left[\sum_{i=1}^{n_x}\left(q_j^i(1-q_j^i)+\sum_{k\neq i}(q_j^{ik}-q_j^i\cdot q_j^k)\right)\right]\right)^{\frac{3}{2}}.
$$

Now, according to Assumption (a), the following weighted average of sums of covariances *is non-negative:*

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{k\neq i}(p_j^{ik}-p_j^i\cdot p_j^k)+\frac{1}{(n')^2}\sum_{j=1}^{n'}\sum_{k\neq i}(q_j^{ik}-q_j^i\cdot q_j^k)$$

Therefore, according to this assumption, the denominator of the Lyapunov quotient is greater than or equal to

$$\left(\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n_x}p_j^i(1-p_j^i)+\frac{1}{(n')^2}\sum_{j=1}^{n'}\sum_{i=1}^{n_x}q_j^i(1-q_j^i)\right)^{\frac{3}{2}}.$$

In short, under the above assumption regarding the sum of the covariances, the Lyapunov quotient proves to be less than

$$\frac{\frac{1}{n^3}\sum_{j=1}^{n}\sum_{i=1}^{n_x}(1-p_j^i)p_j^i+\frac{1}{(n')^3}\sum_{j=1}^{n}\sum_{i=1}^{n_x}(1-q_j^i)q_j^i}{\left(\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n_x}p_j^i(1-p_j^i)+\frac{1}{(n')^2}\sum_{j=1}^{n'}\sum_{i=1}^{n_x}q_j^i(1-q_j^i)\right)^{-\frac{1}{2}}}.$$

Now, if $n'=n+r$ with $r\geq 0$, the above numerator is less than

$$\frac{1}{n^3}\sum_{j=1}^{n}\sum_{i=1}^{n_x}(1-p_j^i)p_j^i+\frac{1}{n^3}\sum_{j=1}^{n}\sum_{i=1}^{n_x}(1-q_j^i)q_j^i$$

and the denominator is greater than

$$\left(\frac{1}{(n')^2}\sum_{j=1}^{n}\sum_{i=1}^{n_x}p_j^i(1-p_j^i)+\frac{1}{(n')^2}\sum_{j=1}^{n'}\sum_{i=1}^{n_x}q_j^i(1-q_j^i)\right)^{-\frac{1}{2}},$$

and thus the quotient is upper bounded by

$$\left(\frac{n+r}{n}\right)^3\left(\sum_{j=1}^{n}\sum_{i=1}^{n_x}p_j^i(1-p_j^i)+\sum_{j=1}^{n+r}\sum_{i=1}^{n_x}q_j^i(1-q_j^i)\right)^{-\frac{1}{2}}.$$

An analogous argument would serve to prove that the quotient is upper bounded by the quantity

$$\left(\frac{n'+r}{n'}\right)^3\left(\sum_{j=1}^{n}\sum_{i=1}^{n_x}p_j^i(1-p_j^i)+\sum_{j=1}^{n+r}\sum_{i=1}^{n_x}q_j^i(1-q_j^i)\right)^{-\frac{1}{2}}. \tag{4}$$

in the case that $n=n'+r$, with $r\geq 0$.
We easily observe that Expression 4 tends to 0, when

$$\lim_{n\to\infty}\sum_{j=1}^{n}\sum_{i=1}^{n_x}p_j^i(1-p_j^i)+\sum_{j=1}^{n+r}\sum_{i=1}^{n_x}q_j^i(1-q_j^i)=\infty. \tag{5}$$

According to Assumption (b), the above equality holds (i.e., the above sequence of sums tends to infinity).
In summary, the so-called Lyapunov condition is satisfied, and therefore the sequence of random variables in Equation 3 converges in law to a standard normal. □

Remark 3. *We assume that a convex linear combination of the two following sums of covariances:*

$$\sum_{j=1}^{n}\sum_{k\neq i}(p_j^{ik} - p_j^i \cdot p_j^k) \text{ and } \sum_{j=1}^{n'}\sum_{k\neq i}(q_j^{ik} - q_j^i \cdot q_j^k)$$

is non-negative. This is a natural assumption in our problem. Informally speaking, this assumption is based on the fact that we can assume that the probability distributions P_i $i = 1, \ldots, n_x$ (corresponding to n_x pixels of the same region R_X), or at least a high proportion of them, are sufficiently similar to each other so that the (set-)difference between the corresponding concentration intervals $A_i \Delta A_j = (A_i \cap \overline{A}_j) \cup (\overline{A}_i \cap A_j)$ is small enough for the following inequalities to be satisfied:

$$P_j(A_i \cap A_k) \geq P_j(A_i) \cdot P_j(A_k) \text{ and } Q_j(A_i \cap A_k) \geq Q_j(A_i) \cdot Q_j(A_k)$$

for most of the triplets (i, j, k).

Remark 4. *We also assume that Equation (5) is satisfied. For the above premise to apply, we assume that some $\epsilon > 0$ such that $p_j^i > \epsilon$, for every pair (i, j). In fact, in such cases the n^{th} term of the sequence is lower bounded by $\epsilon(1-\epsilon)n \cdot n_x$, which clearly tends to infinity when $n \to \infty$. This last assumption about the lower bound of the probabilities p_j^i is consistent with our idea that:*

- *the distributions at different pixels in region R_X are not too different from each other, hence p_j^i is not too different from p_i^i, and*
- *the fact that A_i are concentration intervals at the β level (i.e. $p_i^i > \beta$) for some sufficiently large β.*

Now, to be able to apply the test we need no unknown parameter to appear in the test statistic expression. Thus, we have to get rid of the values of the population variances in Equation (3), and replace them by the corresponding sample variances. But first we will prove some of the auxiliary results.

Lemma 2. *Consider the concentration intervals $A_1, \ldots, A_{n_x}$ and the sequences of variables $X_1, \ldots, X_n, \ldots$ and $Y_1, \ldots, Y_{n'}, \ldots$. For each $n, n' \in \mathbb{N}$ consider the collection of random variables $\hat{p}^i$, $\hat{q}^i$, $\hat{p}^{i,k}$ and $\hat{q}^{i,k}$, $i, k = 1, \ldots, n_x$, as defined in Equation (1). Then:*

$$E(\hat{p}^i) = \overline{p}^i,\ E(\hat{p}^{i,k}) = \overline{p}^{i,k}, E(\hat{q}^i) = \overline{q}^i,\ E(\hat{q}^{i,k}) = \overline{q}^{i,k},$$

where $\overline{p}^i$, $\overline{p}^{i,k}$ $\overline{q}^i$ and $\overline{q}^{i,k}$ respectively denote the parameters:

- $\overline{p}^i = \dfrac{p_1^i + \ldots + p_n^i}{n}$, $\overline{p}^{i,k} = \dfrac{p_1^{i,k} + \ldots + p_n^{i,k}}{n}$
- $\overline{q}^i = \dfrac{q_1^i + \ldots + q_{n'}^i}{n}$, $\overline{q}^{i,k} = \dfrac{q_1^{i,k} + \ldots + q_{n'}^{i,k}}{n'}$

Proof: The proof is immediate, based on the linearity of the expectation. □

Lemma 3. *Consider a sequence of independent Bernoulli random variables $X_n \equiv B(1, p_n)$ (not necessarily equally distributed), and the sequences of variables $\overline{X}_n$ and constants $\overline{p}_n$ calculated from the above as:*

$$\overline{X}_n = \frac{X_1 + \ldots + X_n}{n}, \quad \overline{p}_n = \frac{p_1 + \ldots + p_n}{n}, n \in \mathbb{N}.$$

Then, given a pair of arbitrarily small $\epsilon > 0$ and $\delta > 0$, $n_0 \in \mathbb{N}$ such that

$$P\left(|\overline{X}_n - \overline{p}_n| > \epsilon\right) < \delta, \; \forall\, n \geq n_0.$$

Proof: *Given $n \in \mathbb{N}$, taking into account the fact that the variables $X_1, \ldots, X_n$ are independent and that the variance of each of them is upper bounded by $\frac{1}{4}$, we conclude that the variance of $\overline{X}_n$ is upper bounded by $\frac{1}{4n}$. Thus, according to Bienaymé-Chebyshev's inequality, we can prove the thesis of the lemma.* □

Corollary 1. *Consider the concentration intervals $A_1, \ldots, A_{n_x}$ and the sequences of variables $X_1, \ldots, X_n, \ldots$ and $Y_1, \ldots, Y_{n'}, \ldots$. For each n and each n', consider the collection of random variables $\hat{p}^i$, $\hat{q}^i$, $\hat{p}^{i,k}$ and $\hat{q}^{i,k}$, $i, k = 1, \ldots, n_x$, as defined in Equation (1). Then, given a pair of arbitrarily small $\epsilon > 0$ and $\delta > 0$, and given an arbitrary pair $i, k = 1, \ldots, n_x$, we have $n_0 \in \mathbb{N}$ such that*

- $P\left(|\hat{p}^{i,k} - \overline{p}^{i,k}| > \epsilon\right) < \delta$
- $P\left(|\hat{p}^{i} - \overline{p}^{i}| > \epsilon\right) < \delta$
- $P\left(|\hat{q}^{i,k} - \overline{q}^{i,k}| > \epsilon\right) < \delta$
- $P\left(|\hat{q}^{i} - \overline{q}^{i}| > \epsilon\right) < \delta$,

for all $n \geq n_0$[1].

Lemma 4. *Consider two values $0 < p, q < 1$ and two positive random variables X and Y defined in the same probability space. If, for a pair of arbitrarily small positive quantities $\epsilon > 0$ and $\delta > 0$, $P(|X - p| > \epsilon) < \delta$ and $P(|Y - q| > \epsilon) < \delta$, then $P(|XY - pq|) > 2\epsilon(1 + \frac{\epsilon}{2}) < 2\delta$.*

Corollary 2. *Consider the concentration intervals $A_1, \ldots, A_{n_x}$ and the sequences of variables $X_1, \ldots, X_n, \ldots$ and $Y_1, \ldots, Y_{n'}, \ldots$. For each $n \in \mathbb{N}$ and for $n' = n + r$, consider the collection of random variables $\hat{p}^i$, $\hat{q}^i$, $\hat{p}^{i,k}$ and $\hat{q}^{i,k}$, $i, k = 1, \ldots, n_x$, as defined in Equation (1). Now consider the random variable*

$$T_n = \frac{1}{n} \sum_{i=1}^{n_x} \sum_{k=1}^{n_x} (\hat{p}^{i,k} - \hat{p}^i \cdot \hat{p}^k) + \frac{1}{n} \sum_{i=1}^{n_x} \sum_{k=1}^{n_x} (\hat{q}^{i,k} - \hat{q}^i \cdot \hat{q}^k)$$

and the constant

$$d_n = \frac{1}{n_x} \sum_{i=1}^{n_x} \sum_{k=1}^{n_x} (\overline{p}^{i,k} - \overline{p}^i \cdot \overline{p}^k) + \frac{1}{n_y} \sum_{i=1}^{n_x} \sum_{k=1}^{n_x} (\overline{q}^{i,k} - \overline{q}^i \cdot \overline{q}^k)$$

Then the sequence $\left(\frac{T_n}{d_n}\right)_{n \in \mathbb{N}}$ converges in probability to 1.

[1] Note that the index n appears implicitly in the expressions defining the statistics $\hat{p}^{i,k}$,, $\hat{p}^i$, $\hat{q}^{i,k}$, $\hat{q}^i$ as well as in the definitions of the constants $\overline{p}^{i,k}$, $\overline{p}^i$, $\overline{q}^{i,k}$ and $\overline{q}^i$.

Lemma 5. *For each $n \in \mathbb{N}$ and $n' = n + r$ consider the quantities:*

$$\begin{aligned}\sigma_n^2 &= \frac{1}{n^2}\sigma_j^2 + \frac{1}{(n')^2}(\sigma_j')^2 \\ &= \frac{1}{n}\sum_{i=1}^{n_x}\sum_{k=1}^{n_x}(p^{i,k} - p^i \cdot p^k) + \frac{1}{n'}\sum_{i=1}^{n_x}\sum_{k=1}^{n_x}(q^{i,k} - q^i \cdot q^k),\end{aligned}$$

$$d_n = \frac{1}{n}\sum_{i=1}^{n_x}\sum_{k=1}^{n_x}(\overline{p}^{i,k} - \overline{p}^i \cdot \overline{p}^k) + \frac{1}{n'}\sum_{i=1}^{n_x}\sum_{k=1}^{n_x}(\overline{q}^{i,k} - \overline{q}^i \cdot \overline{q}^k),$$

and

$$c_n = \frac{1}{n^3}\sum_{j \neq j'}\left(\sum_{i=1}^{n_x}(p_j^i - p_{j'}^i)\right)^2 + \frac{1}{(n')^3}\left(\sum_{j \neq j'}(q_j^i - q_{j'}^i)\right)^2.$$

Then:

$$\sigma_n^2 = d_n - c_n.$$

As a corollary of Lemma 5 and Corollary 2, we have:

Corollary 3. *The sequence of random variables $\left(\frac{T_n - c_n}{\sigma_n^2}\right)_{n\in\mathbb{N}}$ converges in probability to the constant 1.*

Now, in the light of the two previous results, and taking the basic properties of convergence in probability into account, we see that:

Corollary 4. *Under the null hypothesis, the sequence of random variables*

$$\frac{\overline{N_n} - \overline{M_n}}{\sqrt{T_n - c_n}} \tag{6}$$

converges in law to a standard normal.

Now, as a consequence of Lemma 1 and Corollary 2 we can easily prove the following result:

Lemma 6. *For every $n \in \mathbb{N}$ consider the constant $a_n = \frac{d_n}{\sigma_n^2} = \frac{d_n}{d_n - c_n}$. Then the sequence*

$$\sqrt{a_n} \cdot \frac{\sum_{j=1}^{n}\frac{N_j}{n} - \sum_{j=1}^{n'}\frac{M_j}{n'}}{\sqrt{T_n}}$$

converges in law to a standard normal.

Remark 5. *In the previous section we proposed a nested family of hypothesis tests indexed by $\alpha \in (0, 1)$. For each α we proposed to reject the null hypothesis if and only if The absolute value $\frac{\sum_{j=1}^{n}\frac{N_j}{n} - \sum_{j=1}^{n'}\frac{M_j}{n'}}{\sqrt{T_n}}$ exceeds the quantile $1 - \frac{\alpha}{2}$ of the standard normal. Now, the condition*

$$\frac{\overline{N_n} - \overline{M_n}}{\sqrt{T_n}} > z_{1-\frac{\alpha}{2}}$$

is equivalent to

$$\sqrt{a_n} \cdot \frac{\overline{N_n} - \overline{M_n}}{\sqrt{T_n}} > \sqrt{a_n} \cdot z_{1-\frac{\alpha}{2}}.$$

The size of this test is therefore

$$f(\alpha) = 2\left[1 - \phi(a_n \cdot z_{1-\frac{\alpha}{2}})\right].$$

We can easily see that and $a_n = \frac{d_n}{d_n - c_n} \geq 1$ *for all* $n \in \mathbb{N}$*. Thus, for a given* a_n*, the function* f *is strictly increasing and* $f(\alpha) < \alpha$ *for all* $\alpha \in (0,1)$*. We also note that when the distributions at the different pixels (within each region) are coincident, then* $c_n = 0$ *and therefore the equality* $a_n = 1$ *holds. In that case,* f *coincides with the identity function.*

To conclude this section, we can say that a nested family of hypothesis tests has been constructed to test $H_0 : P = Q$ against $H_1 : P \neq Q$. The asymptotic test size corresponding to each index $\alpha \in (0,1)$ is $f(\alpha) = 2\left[1 - \phi(a_n \cdot z_{1-\frac{\alpha}{2}})\right] < 2\left[1 - \phi(z_{1-\frac{\alpha}{2}})\right] = \alpha$. From the sample data, we cannot give a precise value of the difference $\alpha - f(\alpha)$. We simply know that the difference is 0 when the equalities $P_1 = \ldots = P_{n_x}$ and $Q_1 = \ldots = Q_{n_y}$ are satisfied.

In case we want to use a test but do not need to know its size (probability of rejection under the null hypothesis) we can use the above test (the $D-$ test). Otherwise, if we need to select a specific size, we have to modify the test statistic. This modification will involve a higher computational cost.

In the next section we construct a new family of modified tests (KD-tests), $KT_\alpha)_{\alpha \in (0,1)}$, with approximate size α for each of them. Each test is based on the use of another statistic that results from the application of a variation to the statistic used in this section. That variation involves determining an unbiased estimator of the c_n value that allows us to construct an unbiased estimate of the $\sigma_n^2 = d_n - c_n$ value. The c_n estimator involves more complex calculations and is based on the selection of two random values (instead of 1) from each of the pixels of the two regions.

4. Known-size D-test (KD-test)

In the previous section we proposed to use the statistic:

$$D = \frac{\overline{N} - \overline{M}}{\sqrt{T}}$$

to test the hypothesis $H_0 : P = Q$ against $H_1 : P \neq Q$. The so-called D-test consists of comparing the absolute value of this statistic with the quantile z and rejecting the hypothesis when the former is greater than the latter. We have also seen that, under the null hypothesis, the asymptotic distribution of the sequence

$$\sqrt{a_n} \cdot \frac{\overline{N_n} - \overline{M_n}}{\sqrt{T_n}}$$

is standard normal, and that $a_n > 1$, so that for a sufficiently large $n, n' \in \mathbb{N}$, the size of the α-$D-$test is approximately $f(\alpha) = 2\left[1 - \phi(a_n \cdot z_{1-\frac{\alpha}{2}})\right]$, which is upper bounded by a known threshold α, but is not precisely determined (as a_n is unknown for us).

In the next subsections, we will construct a family of tests of known size.

4.1. Unknown variance of the previous statistic

As we mentioned above, the asymptotic variance of

$$\sqrt{a_n} \cdot \frac{\overline{N_n} - \overline{M_n}}{\sqrt{T_n}}$$

is 1, and therefore, for a sufficiently high n value, the variance of our $D-$statistic

$$\frac{\overline{N_{n_X}} - \overline{M_{n_X}}}{\sqrt{T_{n_X}}}$$

is approximately equal to $\frac{1}{a_{n_X}}$, which is never greater than 1.

This implies that the variance of $|D|$ (absolute value of the $D-$statistic) is also less than the variance of the absolute value of a standard Gaussian.

In the next subsections, we will construct a new statistic as a modification of the D-statistic, in which the denominator represents an unbiased estimate of the variance of $\overline{N} - \overline{M}$, even in the case where the distributions over the different pixels of the same region do not coincide. As mentioned above, this new statistic involves a greater computational investment.

4.2. The KD-statistic and the KD-test

We randomly choose two observations, out of the m available observations in each pixel, and construct concentration intervals with the remaining $m-2$ observations. We thus have two independent tuples of dimension n_x in the R_X region, $(x_{1r}, \dots, x_{n_x r}),\ r = 1,2$ and two tuples of dimension n_y in the R_Y region, $(y_{1r}, \dots, y_{n_y r}),\ r = 1,2$ that can be considered as observations of two independent copies of $(X_1, \dots, X_{n_x})$ and $(Y_1, \dots, Y_{n_y})$, respectively. We define N_{jr}, $j = 1, \dots, n_x$ M_{jr}, $j = 1, \dots, n_y$, $1, 2$ as follows:

- $N_{jr} = \#\{i \in \{1, \dots, n_x\} : A_i \ni X_{jr}\},\ r = 1,2$
- $M_{jr} = \#\{i \in \{1, \dots, n_x\} : A_i \ni Y_{jr}\},\ r = 1,2.$

Definition 3. *The* KD-statistic *is defined as follows:*

$$K = \frac{\overline{N} - \overline{M}}{\sqrt{T - R}} \tag{7}$$

where

$$T = \frac{1}{n}\sum_{i=1}^{n_x}\sum_{k=1}^{n_x}(\hat{p}^{i,k} - \hat{p}^i \cdot \hat{p}^k) + \frac{1}{n'}\sum_{i=1}^{n_x}\sum_{k=1}^{n_x}(\hat{q}^{i,k} - \hat{q}^i \cdot \hat{q}^k)$$

and

$$R = \frac{1}{n^3}\sum_{j \neq j'}\left(\frac{N_{j1}}{n_x} - \frac{N_{j'1}}{n_x}\right) \cdot \left(\frac{N_{j2}}{n_x} - \frac{N_{j'2}}{n_x}\right) + \frac{1}{(n')^3}\sum_{j \neq j'}\left(\frac{M_{j1}}{n_x} - \frac{M_{j'1}}{n_x}\right) \cdot \left(\frac{M_{j2}}{n_x} - \frac{M_{j'2}}{n_x}\right).$$

Based on the calculation of the absolute value of the KD-statistic, we proceed as follows to test hypothesis $H_0 : P = Q$ against $H_1 : P \neq Q$.

Definition 4. *Let us consider the test of the null hypothesis $H_0 : P = Q$ against the alternative hypothesis $H_1 : P \neq Q$.* KD-test *refers to the procedure that consists of rejecting H_0 when $|K| > z_{1-\frac{\alpha}{2}}$, and not rejecting it otherwise.*

In the next subsection we justify why the asymptotic size of this test is α.

4.3. Asymptotic distribution of the KD-test statistic

Theorem 1. *Under the null hypothesis $H_0 : P = Q$, the sequence of KD-statistics $(K_n)_{n\in\mathbb{N}}$ converges in law to an $N(0,1)$ distribution.*

Corollary 5. *The test constructed in Subsection 4.2 has an asymptotic α size.*

5. Technical details on test construction

5.1. Parzen-Rozenblatt based concentration interval estimation

The D-test is based on estimations of most specific concentration intervals. Let $\boldsymbol{s} = \{s_i, \dots, s_n\}$ be n samples of the random variable S. We can suppose without any loss of generality that the samples are sorted in ascending order: $s_1 \leq s_2 \leq \cdots \leq s_n$. The Parzen-Rosenblatt estimate of $f_S(u)$, which is the density of S at location $u \in \mathbb{R}$, is given by:

$$\hat{f}_S(u) = \frac{1}{\Delta} \sum_{i=1}^{n} \kappa \left(\frac{u - s_i}{\Delta} \right), \tag{8}$$

where $\Delta \in \mathbb{R}^+$ is a bandwidth and κ is a kernel, i.e. a unimodal positive function summing to one: $\int_{\mathbb{R}} \kappa(u)du = 1$. As recommended in (18), we propose to use the Epanechnikov kernel defined by:

$$\kappa(u) = \left\{ \begin{array}{cc} \frac{3}{4.\sqrt{5}}.\left(1 - \frac{u^2}{5}\right), & if\, |u| \leq \sqrt{5} \\ 0 & else, \end{array} \right.$$

and estimate the bandwidth Δ by $\Delta = 0.79.R.n^{-\frac{1}{5}}$ where R is the interquartile range of $\boldsymbol{s}$.

We suggest to reduce the search for the most specific β-concentration interval to finding two indices $i, j \in \{1, \dots, n\}$ such that $[s_i, s_j]$ is the best choice among others for being the most specific β-concentration interval, i.e. being the smallest interval such that $\int_{s_i}^{s_j} \hat{f}_S(u)\, du \geq \beta$.

We thus compute, for each sample value, the $\hat{f}_S(s_i)$ value by using Eq. (8). We normalize this distribution to obtain a discrete distribution over all the samples: $\forall i \in \{1, \dots, n\}$, $\rho_i = \lambda.\hat{f}_S(s_i)$, with $\lambda = \frac{1}{\sum_{i=1}^{n} \hat{f}_S(s_i)}$.

Now, let (.) be a permutation that sorts the discrete distribution in descending order: $\rho_{(1)} \geq \rho_{(2)} \geq \dots \rho_{(n)}$. Finding the most specific β-concentration intervals consists of finding the highest value k such that $\sum_{i=\underline{k}}^{\overline{k}} \rho_i \leq \beta$, where $\underline{k}$ is the highest value such that $\max_{i=1\dots\underline{k}}\ \rho_i < \rho_{(k)}$ and $\overline{k}$ is the smallest value such that $\min_{i=\overline{k}\dots n}\ \rho_i < \rho_{(k)}$.

The most specific β-concentration interval is defined as $I_\beta = [s_{\underline{k}}, s_{\overline{k}}]$.

5.2. D-test computations

The D-test aims at comparing two ROIs R_X and R_Y made of n_x and n_y pixels respectively. To the i^{th} pixel of ROI X (resp. ROI Y) is associated a set of m (bootstrapped) measured values $({}^1x_i, \ldots, {}^m x_i)$ (resp. $({}^1y_i, \ldots, {}^m y_i)$).
The D-statistic is computed in five steps.

Step 1
The first step, called the drop-out step, consists of creating two tuples $\mathring{\boldsymbol{x}} = (\mathring{x}_1, \ldots, \mathring{x}_{n_x})$ and $\mathring{\boldsymbol{y}} = (\mathring{y}_1, \ldots, \mathring{y}_{n_y})$. $\mathring{\boldsymbol{x}}$ (resp. $\mathring{\boldsymbol{y}}$) is obtained by randomly selecting, for each $i \in \{1, \ldots, n_x\}$, (resp. $i \in \{1, \ldots n_y\}$), a single index in the set $\{1, \ldots, m\}$, and then taking the corresponding value from the bootstrapped tuple of measured values $({}^1x_i, \ldots, {}^m x_i)$ (resp. $({}^1y_i, \ldots, {}^m y_i)$). The selected samples in R_X are removed from the original set which only have $m-1$ values left. We can assume, without loss of generality, that these sets are renumbered in order to associate, for each i^{th} pixel of ROI X, a set of $m-1$ (bootstrapped) measured values $\boldsymbol{x}_i = \{{}^1x_i, \ldots, {}^{m-1} x_i\}$.

Step 2
The second step involves associating, to each i^{th} set $\boldsymbol{x}_i = ({}^1x_i, \ldots, {}^{m-1} x_i)$, a concentration interval $[\underline{x}_i, \overline{x}_i]$ by using the Parzen-Rozenblatt based concentration interval estimation presented in Section 5.1.

Step 3
The third step consists of computing $\{N^i\}_{i=1\ldots n_x}$, $\{M^i\}_{i=1\ldots n_x}$, $\{N^{i,j}\}_{i,j=1\ldots n_x}$ and $\{M^{i,j}\}_{i,j=1\ldots n_x}$.
For each $i \in \{1 \ldots n_x\}$, N^i is the number of values of $\mathring{\boldsymbol{x}}$ that belong to $[\underline{x}_i, \overline{x}_i]$ and M^i is the number of values of $\mathring{\boldsymbol{y}}$ that belong to $[\underline{x}_i, \overline{x}_i]$.
For each $i, j \in \{1 \ldots n_x\}$, $N^{i,j}$ is the number of values of $\mathring{\boldsymbol{x}}$ that belong to $[\underline{x}_i, \overline{x}_i] \cap [\underline{x}_j, \overline{x}_j]$ and $M^{i,j}$ is the number of values of $\mathring{\boldsymbol{y}}$ that belong to $[\underline{x}_i, \overline{x}_i] \cap [\underline{x}_j, \overline{x}_j]$.

Step 4
The fourth step consists of using the above computed values to estimate the D-Test parameters.
We set:

i) $\overline{N} = \frac{1}{n_x} \sum_{i=1}^{n_x} N^i$,

ii) $\overline{M} = \frac{1}{n_x} \sum_{i=1}^{n_x} M^i$,

iii) $\nu_x = \frac{1}{n_x^2} \sum_{i=1}^{n_x} \sum_{j=1}^{n_x} N^{i,j} - \frac{1}{n_x} \cdot N^i \cdot N^j$,

iv) $\nu_y = \frac{1}{n_y^2} \sum_{i=1}^{n_x} \sum_{j=1}^{n_x} M^{i,j} - \frac{1}{n_y} \cdot M^i \cdot M^j$.

Step 5
This final step involves computing the D-test (see Expression (2)):

$$D = \frac{\overline{N} - \overline{M}}{\sqrt{\nu_x + \nu_y}}.$$

5.3. Computations of the KD-test

Computing the KD-statistic follows almost the same steps as computing the D-statistic, with three notable differences:

1 – during step 1, two values need to be randomly extracted from each set, thus four sets are created $\mathring{\boldsymbol{x}} = \{\mathring{x}_1, \dots, \mathring{x}_{n_x}\}$, $\mathring{\boldsymbol{x}}' = \{\mathring{x}'_1, \dots, \mathring{x}'_{n_x}\}$, $\mathring{\boldsymbol{y}} = \{\mathring{y}_1, \dots, \mathring{y}_{n_y}\}$ and $\mathring{\boldsymbol{y}}' = \{\mathring{y}'_1, \dots, \mathring{y}'_{n_y}\}$. The set $\boldsymbol{x}_i$ thus contains only $m-2$ values: $\boldsymbol{x}_i = \{{}^1x_i, \dots, {}^{m-2}x_i\}$,

2 – during step 3, doubling of the drop-out allows us to compute four new sets of values: $\{E_i\}_{i=1\dots n_x}$, $\{E'_i\}_{i=1\dots n_x}$, $\{F_i\}_{i=1\dots n_y}$, $\{F'_i\}_{i=1\dots n_y}$, .

For each $i \in \{1 \dots n_x\}$, E_i (rsp.E'_i) is the number of intervals $[\underline{x}_i, \overline{x}_i]$ $(i = 1 \dots n_x)$ that contain the i^{th} value of $\mathring{\boldsymbol{x}}$ (rsp. $\mathring{\boldsymbol{x}}'$), and F_i (rsp. F'_i) is the number of intervals $[\underline{x}_i, \overline{x}_i]$ $(i = 1 \dots n_x)$ that contain the i^{th} value of $\mathring{\boldsymbol{y}}$ (rsp. $\mathring{\boldsymbol{y}}'$).

3 – during step 4, we need to add the computation of two parameters:

v) $\varsigma_x = \frac{1}{n_x^4} \sum_{i=1}^{n_x} \sum_{j=1}^{n_x} (E_i - E_j).(E'_i - E'_j)$,

vi) $\varsigma_y = \frac{1}{n_x^3.n_y} \sum_{i=1}^{n_y} \sum_{j=1}^{n_y} (F_i - F_j).(F'_i - F'_j)$.

4 – The final step, i.e. computing the KD-statistic, is very similar to computing the D-statistic:

$$K = \frac{\overline{N} - \overline{M}}{\sqrt{\nu_x + \nu_y - \varsigma_x - \varsigma_y}}.$$

5.4. Some practical advice

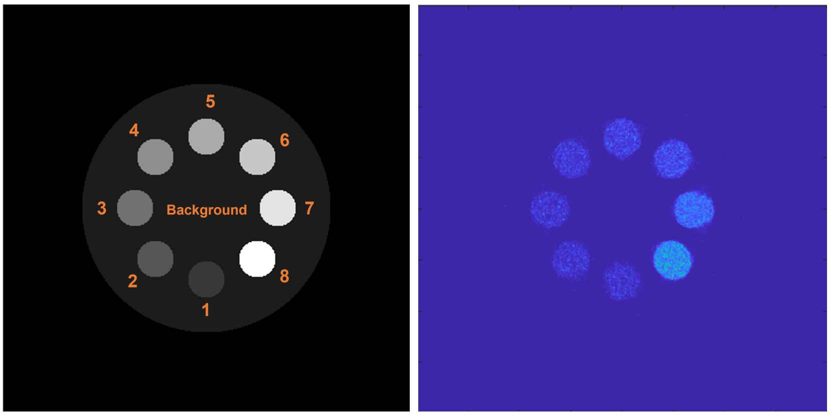
Although this test is non-parametric and does not require prior knowledge of the distribution of values in the regions being compared, the values of a few parameters must be defined, namely the significance level α, the probability of rejecting the null hypothesis falsely, and β, the concentration rate of representative intervals.

- For α, the usual values are 1% and 5%. Based on the experiments we carried out, these values are quite suitable. The values of the associated thresholds are respectively $\tau_{0.01} = z_{1-\frac{0.01}{2}} = 2.57$ for $\alpha = 0.01$ and $\tau_{0.05} = z_{1-\frac{0.05}{2}} = 1.96$ for $\alpha = 0.05$.

- Concerning β, its value defines the extent to which the interval $[\underline{x}_i, \overline{x}_i]$ represents the distribution of values associated with the i^{th} pixel/voxel. The larger beta is, the greater the degree of representativeness, but also the wider the interval is and the more sensitive it is to outliers. Experimentally, we noted that choosing $\beta = 0.9$ is a good trade-off between the degree of representativeness and the sensitivity to outliers.

6. Experimentation

In this paper, we propose an application of the proposed statistical test over the problem of ROI comparison in molecular imaging, specifically in PET.



(a) (b)

Figure 1: (a) Simulated phantom. (b) Example of a transaxial slice reconstruction with 30 ML-EM iterations of a 30s GATE simulation of phantom (a)

6.1. Simulation setup

In order to obtain ground truth data, all experiments are performed using the widely used GATE (16) statistical simulation framework. The chosen imaging modality was positron emission tomography (PET).

6.1.1. GATE simulation

For the simulation, an ECAT system configuration for PET scan was designed as it is appropriate for modelling recent PET scanners such as the Siemens Healthineers mCT20 Flow (Knoxville, TN, U.S.A.) that we use at out institution. The generated sinograms had 400 bins of projections and 312 angular views each. For these experiments, attenuation, random and scatter were not taken into account. Only transaxial coincidences were recorded.

6.1.2. Phantom description

We used a phantom made of 8 distinctive cylinders enclosed in a larger cylinder. Each of the 8 cylinders of the same diameter (8 cm) were simulated with different fluorodeoxyglucose (^{18}F-FDG) activity (respectively with 300 kBq, 310 kBq, 325 kBq, 375 kBq, 475 kBq, 600 kBq, 1000 kBq, 1500 kBq activity). The bigger cylinder, referred to as ”background”, of 50 cm diameter is filled with 200 kBq of ^{18}F-FDG. A schematic representation of a section of the phantom with ground truth cylinder indexing highlighted can be found in Figure 1(a).

6.1.3. Bootstrapping and image reconstruction

For the bootstrap approach, for each of the 109 transaxial slices, the total 30 s acquisition is divided into 30 sub-sinograms of 1 s each. From these 30 sub-sinograms, we created a series of 15k bootstrapped sub-sinograms. We grouped those bootstrapped sub-sinograms 30 by 30 to obtain 500 bootstrapped sinograms (13). Then, we obtained 500 bootstrap

$\mathbf{ROI}_j$ $\mathbf{ROI}_i$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
■	**KD-test**								**D-test**							
1	■	1	1	1	1	1	1	1	■	1	1	1	1	1	1	1
2	0	■	1	1	1	1	1	1	0	■	1	1	1	1	1	1
3	0.22	0	■	1	1	1	1	1	0.22	0	■	1	1	1	1	1
4	1	0.96	0.13	■	1	1	1	1	1	0.91	0.2	■	1	1	1	1
5	1	1	1	1	■	1	1	1	1	1	1		■	1	1	1
6	1	1	1	1	1	■	1	1	1	1	1	1	1	■	1	1
7	1	1	1	1	1	1	■	1	1	1	1	1	1	1	■	1
8	1	1	1	1	1	1	1	■	1	1	1	1	1	1	1	■

Table 1: Sensitivity (in orange cells) versus specificity (in blue cells) table for the 5% threshold.

replicates that, once reconstructed, can be considered as samples of the distribution underlying the reconstructed data.

For the reconstruction process, for all transaxial slices, each of the 500 bootstrap replicates are reconstructed with a 2D ML-EM statistical iterative algorithm (17) with 30 iterations. An example of phantom reconstruction is presented in Figure 1(b).

6.2. Results

In Tables (1) and (2), we present the results of the experiments that were carried out over the reconstructions of the phantom presented in Figure 1(a) for two different significance levels (1% and 5%).

In order to have a range of comparisons and to be able to characterize the performance of the proposed test in terms of sensitivity and specificity, the reconstruction of $n = 10$ adjacent trans-axial slices was performed in 2D mode. Each ROI_i of each slice was compared to the same ROI_i in the other slices that were considered to be equal resulting in $n(n-1)/2$ test results values. In the same way, each ROI_i of each slice are compared with KD-test and D-test to ROI_j in all the reconstructed slices, which also corresponds to $n(n-1)/2$ tests.

Notice that in the experiment section of this paper, the medical interpretation of sensitivity and specificity are used. Thus, sensitivity can be seen as the probability of rejecting H_0 if a difference is observed between two ROI distributions' and the specificity is seen as the probability of H_0 if two ROIs have the same distribution.

Under ambiguous conditions (ROIs of close activity or very noisy), the results of both the D-test and KD-test may vary. In order to take the variability related to drop-out choice into account, each comparison was performed 10 times and the test result was averaged.

Sensitivity and specificity levels are 100% when comparing all combinations of ROI of index greater than or equal to 5. Comparison of ROI 4 (375 kBq) against ROI 2 (310 kBq) gives good results for a 5% significance level (specificity of 96% and 91%, respectively, for the KD-Test and D-Test with 100% sensitivity). The results for this specific comparison are worse for the 1% significance level (specificity drops to 49% and 35%, respectively, for KD-Test and D-Test still with 100% sensitivity).

For ROI lower than or equal to 3 (ROI simulated with 300 kBq, 310 kBq, 325 kBq of radiotracer), the two tests are not able to discriminate these ROI between them.

$\mathbf{ROI}_j$ / $\mathbf{ROI}_i$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
	KD-test								**D-test**							
1		1	1	1	1	1	1	1		1	1	1	1	1	1	1
2	0		1	1	1	1	1	1	0		1	1	1	1	1	1
3	0	0		1	1	1	1	1	0	0		1	1	1	1	1
4	0.98	0.49	0		1	1	1	1	1	0.35	0		1	1	1	1
5	1	1	1	1		1	1	1	1	1	1	1		1	1	1
6	1	1	1	1	1		1	1	1	1	1	1	1		1	1
7	1	1	1	1	1	1		1	1	1	1	1	1	1		1
8	1	1	1	1	1	1	1		1	1	1	1	1	1	1	

Table 2: Sensitivity (in orange cells) versus specificity (in blue cells) table for the 1% threshold.

6.3. Experts visual comparison

The results highlighted in Tables (1) and (2) are interesting but do not clarify the efficiency D-Test and the KD-Test are in comparison to visual inspection by physician experts.

Using the same experimental setup as described in the previous section, we proposed nuclear medicine physicians of our institution (Montpellier University Hospital - 7 physicians answered) to independently compare different regions of interest two-by-two.

As ROI 7 and 8 are really easy to compare to other ROIs, we limited the random selection of the ROI indice to indices from 1 to 6 (see Figure 1(a)). A sequence of 20 ROI indices were randomly drawn. For each ROI, the corresponding reconstructed slice (among the 10 adjacent) was also randomly drawn to ensure that every ROI have different statistics even if they share the same count number. Reconstructed ROI of the sequence were segmented and presented to the experts as in Figure 2. They thus were asked to perform 19 comparisons (16 different and 3 identical pairs of ROI).

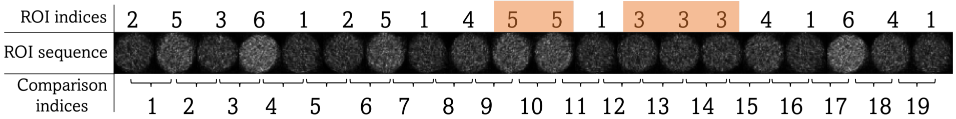


Figure 2: Sequence of randomly chosen ROI drawn from different adjacent slices given to physician experts for visual comparison. The identical adjacent ROI are highlighted in orange.

The results are presented in Table (3). For the D-test and KD-test, the results are given for both 1% and 5%. For each comparison, the tests were performed 10 times to account for drop-out variability. The results are expressed in percentage of right answers (1 for identical ROI, 0 for different) for expert evaluation and in proportion to the acceptance (1) or rejection (0) of the null hypothesis for D-test and KD-tests.

Table (3) shows that where the D-test and KD-test failed to reject the null hypothesis in the most critical cases (ROI 1 *vs* ROI 2, comparison 5 in the table), the visual inspection also failed to discriminate them (86% of wrong answers for physicians). In cases where the ROIs have the same distribution (comparison 10, 13 and 14), no tests rejected the null hypothesis whatever the chosen significance level. When no experts failed to find a difference between the two tested ROI, the statistical tests always reject H_0, except for comparison 8 (ROI 1 *vs* ROI 4).

The fact that the tests never rejected H_0 when the regions had identical activity is consistent with its expectations as it is designed to minimize the incorrect rejection

CI	**1 - 4**	**5**	**6-7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17-18**	**19**
GT	**0**	**0**	**0**	**0**	**0**	1	**0**	**0**	1	1	**0**	**0**	**0**	**0**
Experts	0	0.86	0	0	0.14	1	0	0.29	1	1	0.43	0.14	0	0.14
D-test*	0	1	0	0.6	0.2	1	0	0.9	1	1	1	0.3	0	0.1
D-test**	0	1	0	0.4	0.1	1	0	0.8	1	1	0.6	0	0	0
KD-test*	0	1	0	0.5	0.3	1	0	1	1	1	1	0.2	0	0.3
KD-test**	0	1	0	0.2	0	1	0	1	1	1	0.7	0	0	0

Table 3: Visual expert comparison *versus* D-test and KD-test evaluations to compare different segmented ROI. CI: Comparison indices, GT: Ground truth, *: 1%, **: 5%

of H_0. Considering a significance level α set at 5% for both tests, it tends to be less conservative than 1% but still not enough to wrongly reject H_0.

7. Conclusion

Here we presented a new test to compare two bootstrapped regions in images: the D-test. This research is innovative because, to our knowledge, it is the first statistical test proposed in the literature to compare two regions of interest for which the statistical variability information is accessible. The problem of comparing regions of interest is substantial in some medical disciplines such as nuclear medicine. In case of a poor signal-to-noise ratio or when the activity difference between the two regions of interest is too small, it may be very hard for the physician experts to discriminate those regions in a reliable way.

The statistical test that we have proposed provides a reliable solution to this problem if bootstrap replicates are available (always if list-mode data are registered). It is a non-parametric test for which no hypothesis on the statistical nature of the data is necessary. The core idea of this test is based on a drop-out procedure that relies on concentration intervals built from bootstrap replicates that provide several iterations of the random variable associated with each reconstructed pixel. The KD-test is a variant of the D-Test for which the size and therefore the rejection thresholds are guaranteed. The D-Test applies when the distribution of each region can be assumed to be stationary. When no stationarity can be ensured (or estimated) then using the KD-test is more adapted. We have shown that the behavior of the two tests are very close in the experiments we carried out. This could be easily explained by the fact that, in these experiments, the distribution in each region was stationary.

The procedures described respectively in the D-test and the KD-test do not treat the two compared regions symmetrically. In both cases, we chose as reference the concentration intervals calculated from one of them (region R_X by default) and checked whether the drop-out values of both regions were equally compatible with these concentration intervals. It seems logical to ask whether it makes sense to construct a symmetrical version of both tests. Clearly, the same process can be repeated, taking region R_Y as a reference, and the corresponding statistic D_y can be calculated. It is also clear that the D-test based on this new statistic will not result in the same decision (rejection/non-rejection) for all patients. A natural way to make the initial test symmetrical is to reject the null hypothesis when the maximum value $\max\{|D_x|, |D_y|\}$ exceeds a pre-specified threshold $z_{1-\frac{\alpha}{2}}$. The significance level α' of this new (symmetric) test is clearly lower than α. In fact, the probability of rejection decreases, as a stricter condition for the rejection is imposed. As future work, we propose to compare the power of these symmetrical tests with the respective powers of the non-symmetrical D-tests and KD-tests. An increase in power would mean a higher sensitivity for a similar specificity level, $1 - \alpha$, and the computational cost of these tests would only double the cost of the original tests.

As future prospects, it would be interesting to assess the ability of this statistical test to correctly compare regions of interest in a particular clinical application for which the ground truth about the definitive diagnosis is known. This will require a prospective clinical study. The applications are potentially numerous: from the diagnosis of neurodegenerative dementia to the monitoring of responses to treatment of hematological pathologies or solid cancers.

References

[1] G. Lucignani *et al.*, The use of standardized uptake values for assessing FDG uptake with PET in oncology: a clinical perspective. Nucl. Med. Commun., 25, 651–656, 2004

[2] R. Boellaard *et al.*, Effects of noise, image resolution, and roi definition on the accuracy of standard uptake values: a simulation study. J Nucl Med. 45, 1519-1527, 2004

[3] J. Feuardent *et al.*, Reliability of uptake estimates in FDG-PET as a function of acquisition and processing protocols using the CPET. IEEE Trans. Nucl. Sc., 52, 1447–1452, 2005

[4] E. Laffon *et al.*, The Effect of Renal Failure on 18F-FDG Uptake: A Theoretic Assessment. J. Nucl. Med. Technol., 36, 200–202, 2008

[5] A. de Langen *et al.*, Repeatability of 18F-FDG Uptake Measurements in Tumors: A Meta-analysis. J. Nucl. Med. 53, 701–708, 2012

[6] A. Jena *et al.*, Reliability of semiquantitative F-FDG PET parameters derived from simultaneous brain PET/MRI: a feasibility study. Eur J Radiol. 83, 1269–1274, 2014

[7] F. Kucharczak *et al.*, Brain 18F-FDG PET analysis via interval-valued reconstruction. Proof of concept for Alzheimer's disease diagnosis. Annals of Nuclear Medicine, 34, 565–574, 2020

[8] P.J. Markiewicz *et al.*, Assessment of bootstrap resampling performance for PET data. Phys Med Biol. 60(1), 279-299, 2015

[9] A. Sitek, Data analysis in emission tomography using emission-count posteriors. Phys. Med. Biol., 57, 6779, 2012

[10] F. Kucharczak *et al.*, Interval-based reconstruction for uncertainty quantification in PET. Phys. Med. Biol., 26, 035014, 2018

[11] M. Filipovic *et al.*, PET Reconstruction of the posterior image probability, including multimodal images. IEEE Trans. Med. Im., 38, 1643–1654, 2019

[12] M. Dahlborn, Estimation of image noise in PET using a bootstrap method. Nucl. Sci. Symp. Conf. Record 4, 2075-2079, 2001

[13] I. Buvat, A non-parametric bootstrap approach for analyzing the statistical properties of SPECT and PET images. Phys. Med. Biol. 47, 1761-1775, 2002

[14] C. Lartizien *et al.*, Comparison of bootstrap resampling methods for 3-D PET imaging. IEEE Trans. Med. Im., 29, 1442–1454, 2010

[15] M. Ibaraki *et al.*, Bootstrap methods for estimating PET image noise: experimental validation and application to evaluation of image reconstruction algorithms. Annals of Nucl. Med., 28, 172–182, 2014

[16] S. Jan *et al.*, GATE: a simulation toolkit for PET and SPECT. Phys. Med. Biol. 49, 4543 – 4561, 2004

[17] L. Shepp *et al.*, Maximum likelihood reconstruction for emission tomography. IEEE Trans. Med. Im., 113–122, 1982

[18] B.W. Silverman, Density estimation for statistics and data analysis. Chapman and Hall, 1986