GENERATION OF SYNTHETIC DATA – Cont'd.

- Bootstrap method, detailed examples:

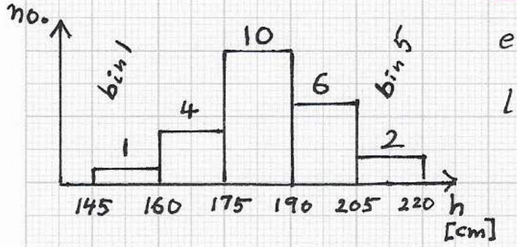
Ⓡ Ⓨ Ⓖ Ⓞ Ⓨ

Ⓨ Ⓞ Ⓨ Ⓖ Ⓨ

Ⓖ Ⓨ Ⓨ Ⓡ Ⓖ

Ⓨ Ⓨ Ⓖ Ⓨ Ⓞ

Population of red (R), orange (O), green (G) and yellow (Y) tennis ball – categorical data



Population of height (h) in 5 bins using 15cm-wide bins – binned data

corresponding data

bin 1: 158

bin 2: 162, 164, 170, 174

bin 3: 176, 177, 178, 179, 182, 183, 184, 185, 185, 188

bin 4: 192, 194, 196, 198, 200, 202

bin 5: 206, 210

MEAN $\bar{x} = 184.48$

Population of actual height values – numerical data

The bootstrap method can serve multiple purposes in the context of data classification. It is also general in the sense that it can be used for categorical, binned numerical or exact numerical data (see left images). The bootstrap method is primarily employed for the purpose of statistically characterizing the distribution of items of a large POPULATION that cannot be analyzed in its entirety. Based on just one ('sufficiently large' and 'population-representative') sample, consisting of N items drawn from the population, the bootstrap methods allow one to gain statistical insights into the population's statistical properties by generating many synthetic "resamples" of N items that are drawn from the one original sample, permitting drawing the same item from the original sample multiple times. The resamples can (i) serve as needed synthetic data for (material) classification and (ii) serve as a basis for determing CONFIDENCE INTERVALS for derived statistical properties of the unseen entire population.

Examples: Colors of tennis balls and heights of tennis players

■ GENERATION OF SYNTHETIC DATA – Cont'd.

- Bootstrap method, detailed examples:

MEAN	
$\bar{x}^0 = 183.33$	158, 164, 174, 176, 178, 182, 185, 185, 192, 196, 200, 210

h [cm]

Original sample of 12 randomly selected heights – selected from population of numerical data

Four bootstrap resamples:

MEAN:	
$\bar{x}^1 =$ 181.33	158, 164, 164, 176, 178, 178, 185, 185, 192, 196, 200, 200
$\bar{x}^2 =$ 182.83	158, 158, 174, 176, 182, 182, 185, 185, 192, 192, 200, 210
$\bar{x}^3 =$ 184.92	164, 164, 174, 176, 178, 185, 185, 185, 192, 196, 210, 210
$\bar{x}^4 =$ 182.17	158, 164, 176, 176, 176, 182, 185, 185, 192, 196, 196, 200

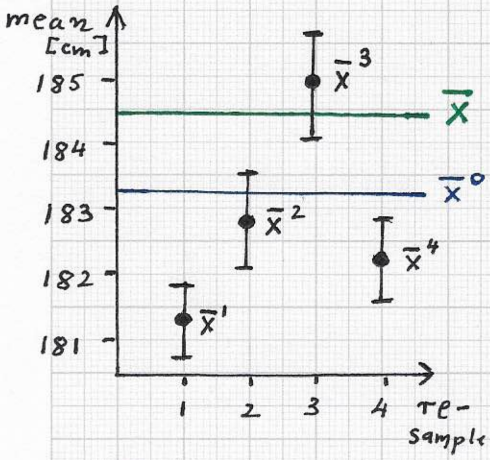
——— duplicated or triplicated data

For example, one could set N=12 and randomly select tennis player heights from the population of 23 heights, producing the sample shown in the figure (left). This sample is used to generate additional synthetic resamples with N=12 items. (In most applications, depending on the specific precision needs, several thousand resamples are required.) The statistical distribution properties of these resamples, when considered in a collective fashion, provide insight into the population's statistical properties. WHY? Given a "good original sample," the resamples can be viewed as "quasi-samples" taken from the actual population.

Four bootstrap resamples are shown in the example (Left); some of the randomly selected data in the resamples are duplicated or triplicated data of the original 12-data sample. The resamples have mean values $\bar{x}^j$, $j=1\dots4$, and standard deviation values $s^j = \left(\frac{1}{11}\sum_{i=1}^{12}(x_i^j - \bar{x}^j)^2\right)^{1/2}$, where sample j consists of data x_i^j, $i=1\dots12$.

■ GENERATION OF SYNTHETIC DATA – Cont'd.

- Bootstrap method, detailed examples:



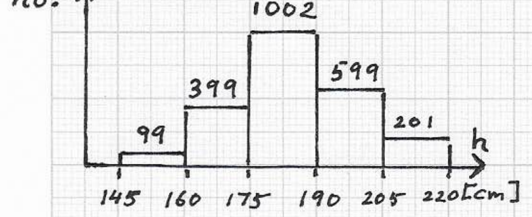
$\bar{x}$ = population mean

$\bar{x}^0$ = orig. sample mean

$\bar{x}^j$ = mean of resample j

Means of population, original sample and 4 resamples; confidence intervals of resamples.

⇒ "A 'BAD' ORIGINAL SAMPLE CAN MISREPRESENT A POPULATION!"



Binned distribution of means $\bar{x}^j$ for 2300 resamples; distribution is a distribution of derived statistical property (mean) of the resamples – depending on original sample's "representative quality".

One can compute a variety of statistical properties for the resamples, e.g., their mean values. It is then possible to compute CONFIDENCE INTERVALS for the respective properties of interest. For example, the "95% CONFIDENCE INTERVAL" c_{95}^j for the mean $\bar{x}^j$ of resample j is (approximately)

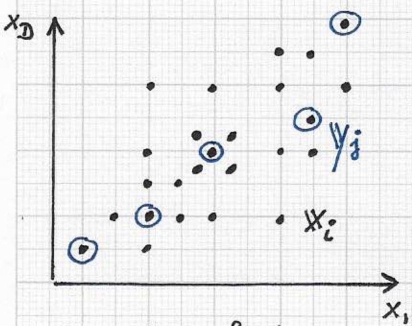
$$c_{95}^j = \bar{x}^j +/- 1.96\, s^j/\sqrt{12}$$

for the above example with 12 data per resample. (See statistics textbooks.) The example shown (left) indicates that the original sample with mean $\bar{x}^0$ has substantial influence on the means $\bar{x}^j$, $j=1\ldots4$, of the resamples and their associated confidence intervals.

Note: In the univariate case, a confidence interval essentially states that there is a certain probability (e.g., 95%) that some statistical property lies inside the interval. In the case of multi-dimensional feature spaces, the "interval" must, in principle, be generalized to a region bounded by a hyperellipsoid or more general boundary.

■ GENERATION OF SYNTHETIC DATA – Cont'd.

- **Bootstrap method:**



• given feature point

○ sampled feature point

$\mathbf{x}_i = (x^i_1, \dots, x^i_D)^T$

$\mathbf{y}_j = (x^j_1, \dots, x^j_D)^T$

$\{\mathbf{x}_i\}_{i=1}^N$ original data

$\{\mathbf{y}_j\}_{j=1}^K$ sample data

Distance:

$$d_i = dist(\mathbf{x}_i, \{\mathbf{y}_j\}) = \min \{\|\mathbf{x}_i - \mathbf{y}_j\|\}_{j=1}^K$$

Sum of squared distances:

$$D_{N,K} = \sum_{i=1}^N d_i^2$$

$D_{N,K}$ = measure of location quality of $\{\mathbf{y}_j\}$

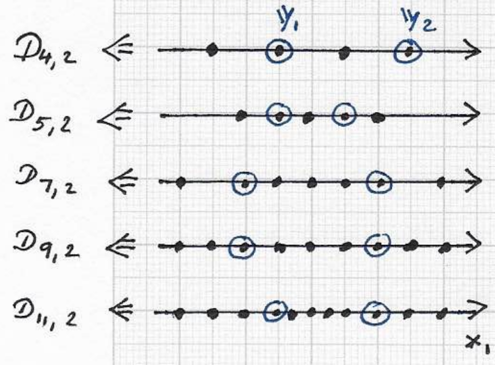
In the following, it is assumed that feature points/vectors are known only for one or a very small number of objects (materials); while this given discrete feature value data set could be 'large', there is a need for generating a 'large' number of additional synthetic feature value data sets for a training-based classification technique. When choosing bootstrapping as method of choice for synthetic data generation, a relatively 'small' sample set consisting of original feature data must be determined. This sample set should – ideally – be optimal in an at least twofold sense: **(i)** the locations of the selected feature points in D-dimensional feature space should optimally 'represent' the locations of all original feature points, and **(ii)** the local density of the ultimately selected points should be 'very close' to the point density of the original feature point set.

(i) Determining an optimal sample set $\{\mathbf{y}_j\}_{j=1}^K$, as far as location is concerned, requires one to minimize a distance between $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{y}_j\}_{j=1}^K$:

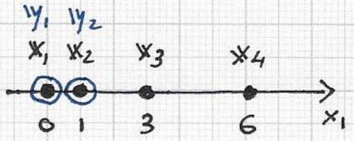
$D_{N,K} \longrightarrow \min$. (See left figure.)

GENERATION OF SYNTHETIC DATA - Cont'd.

- Bootstrap method:

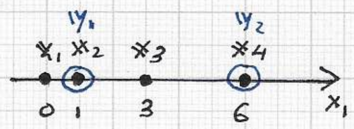


1D feature space; 2 of the given numerical feature values selected from original feature value set with 4, 5, 7, 9, 11 feature data



$\Rightarrow d_1 = 0,\ d_2 = 0,\ d_3 = 2,\ d_4 = 5;$

$\sum_i d_i^2 = 29$



$\Rightarrow d_1 = 1,\ d_2 = 0,\ d_3 = 2,\ d_4 = 0;$

$\underline{\underline{\sum d_i^2 = 5 = \min}}$

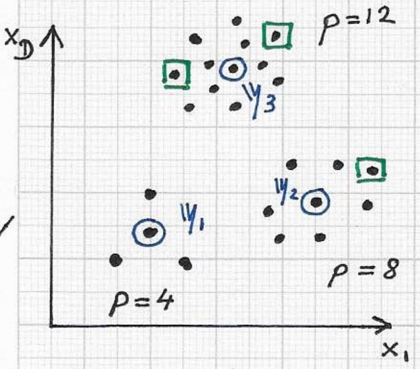
Unique optimal selection of two data, y_1, y_2, from four original data x_1, x_2, x_3, x_4.

The minimization problem $D_{N,K} \to \min$ is a combinatorial optimization problem, where one wants to determine a sample of K 'locationally optimal' feature points y_j — WHERE THE RESULTING OPTIMAL VALUE OF $D_{N,K}^{min}$ SHOULD ALSO BE SMALLER THAN A THRESHOLD D_{max}. For example, given N original feature data and a specific value of K, there exist $\binom{N}{K}$ possibilities to select K data; when allowing the parameter K to take on values 0, 1, 2, ..., N, then the total number of possibilities will be $\sum_{K=0}^{N} \binom{N}{K} = 2^N$. Since it is practically impossible to compute $D_{N,K}$ values for all possible selections of K feature data, one must employ an EFFICIENT COMBINATORIAL OPTIMIZATION METHOD to determine a 'near-optimal' solution, viewed as actually optimal solution with distance value $D_{N,K}^{min}$. SHOULD $D_{N,K}^{min} > D_{max}$, ONE MUST INCREASE THE VALUE OF K and repeat the process until a K-value is found for which $D_{N,K}^{min} \leq D_{max}$. Having determined an optimal/near-optimal sample set of K original (feature data) points, one still must ensure that the (relative) local feature point densities are (approximately) preserved.

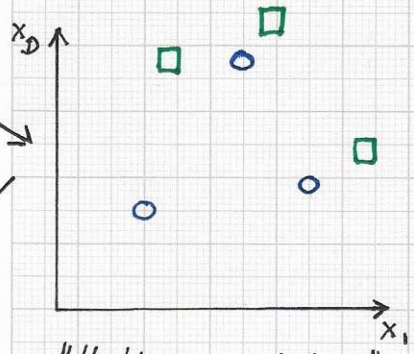
GENERATION OF SYNTHETIC DATA – Cont'd.

- Bootstrap method:

(ii) Assuming that one knows – or can estimate – the local probability density value $p(\mathbb{x})$ in a 'proper neighborhood' of a selected original feature point $\mathbb{y}_j$, one must ensure that, relatively, the original densities in the neighborhoods of selected points $\mathbb{y}_{j_1}$ and $\mathbb{y}_{j_2}$ are preserved. See example shown (left): In addition to the initially selected (location-wise optimal) 3 feature points $\mathbb{y}_1, \mathbb{y}_2, \mathbb{y}_3$ 3 more points are selected (□) in the local neighborhoods of $\mathbb{y}_2, \mathbb{y}_3$, to preserve the probability density ratio 1:2:3 of these **neighborhoods**. (If it was not required to preserve such density ratios, "p-ratios," precisely, one could quantize the **p**-values via a small number of bins, e.g., 4 bins to represent density levels 1, 2, 3, 4. Such a quantization of density values would also ensure that only a 'relatively small number' of points □ are added.)



Original feature point data set with 24 points (•). K=3 points selected for a sample (○ selected optimally w.r.t. location); 3 additional points selected to preserve density ratio (□): 4:8:12 (1:2:3).



"Highly representative" sample of original population of feature points: selected sample points chosen to preserve locations and associated densities of original feature points.

Generate high-quality bootstrap samples.

THE SAMPLE RESULTING FROM STEPS (i) AND (ii) IS USED TO GENERATE NEEDED SYNTHETIC DATA OBTAINED VIA BOOTSRAP RESAMPLING.

BH