Selected Topics in Electrical Engineering: Flow Cytometry Data Analysis

Bilge Karaçalı, PhD
Department of Electrical and Electronics Engineering
Izmir Institute of Technology

Outline

- Probability binning
  - Cox method for comparing cell histograms using K-S statistic
  - Application to equalized histograms for bin cell counts
  - Multivariate extensions

Motivation

- Comparing control and test dataset histograms can determine if the two datasets differ from one another in terms of their respective fluorescence intensity distributions
- However, the statistical tests employed to that end do not identify a “region of difference”
  - Where the “positives” in the test dataset reside on the fluorescence intensity scale
- This prevents the determination of any gates for the expected “positives”

Recalling K-S Test

- The Kolmogorov-Smirnov test computes a P value to assess the similarity of two cell populations
  - in terms of observed fluorescence intensities
- But the test does not provide where on the fluorescence intensity scale the perceived difference lies
- Identifying these “difference regions” requires modifying the comparison test to accommodate limited intensity ranges

Cox Method

- Observation:
  - Suppose a dataset of fluorescence intensities of size n observed on a cell population is given
  - Let n_k represent the number of cells with fluorescence intensity in bin B_k defined by

$$B_k = [I_{k-1}\ I_k]$$

for $k = 1,2, \ldots, K$, where $I_0 = 0$

  - If n_k are sufficiently large, one can then assume that they are realizations of Poisson random variables N_k
    - with potentially different expected number of successes
  - In that case, n_k can be treated as estimates of the mean of N_k, that can be defined in terms of the unknown underlying probability distribution of fluorescence intensities $p(i)$ as

$$n_k \cong \overline{N_k} = n \int_{i \in B_k} p(i)di$$

$$\therefore \int_{i \in B_k} p(i)di \cong \frac{n_k}{n}$$

Cox Method

- Observation (continued):
  - Now, let a control dataset and a test dataset be given
  - Suppose that there are n_k^{cont} and n_k^{test} cells respectively of the control and the test datasets in bin B_k
  - Under the null hypothesis that asserts that both datasets are drawn from the same distribution, we expect

$$\frac{n_k^{cont}}{n^{cont}} \cong \frac{n_k^{test}}{n^{test}} \cong \int_{i \in B_k} p(i) di$$

  - By the same token, the entity

$$Z_k = \frac{\left(\frac{n_k^{cont}}{n^{cont}} - \frac{n_k^{test}}{n^{test}} \right)}{\sqrt{\frac{n_k^{cont}}{(n^{cont})^2} + \frac{n_k^{test}}{(n^{test})^2}}}$$

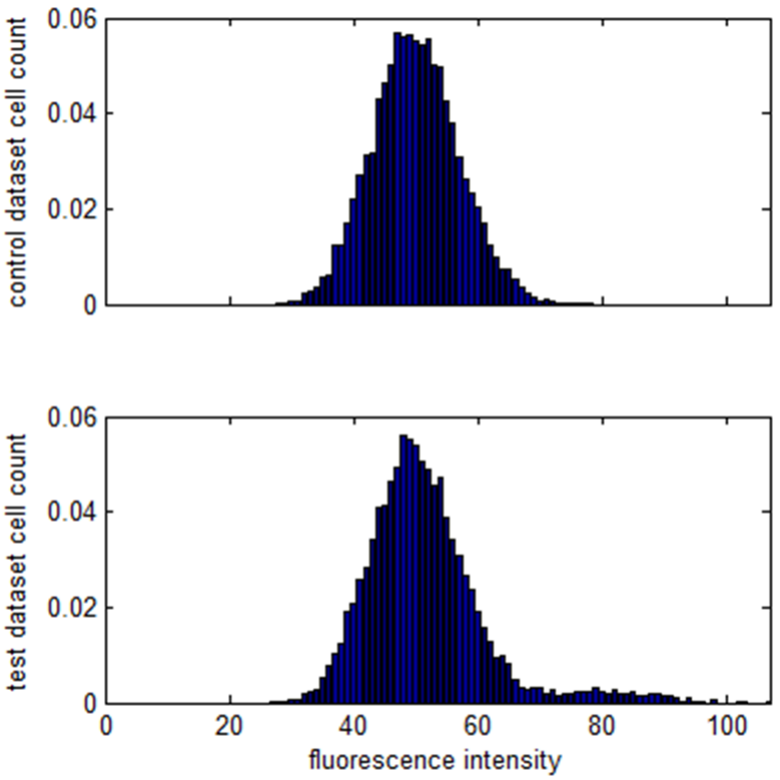
  is a realization of an approximately Gaussian distribution with zero mean and unit variance
    - Q: How?

Cox Method

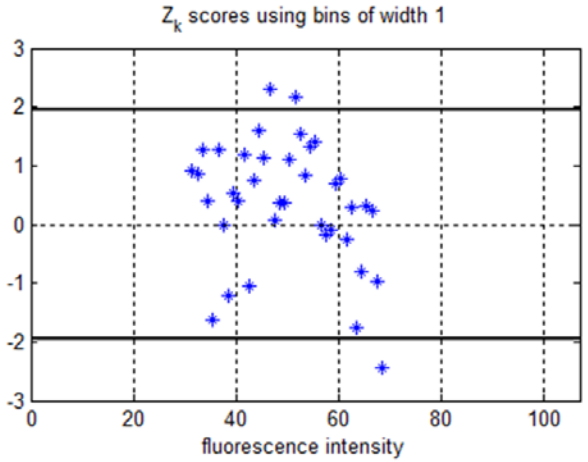
- Based on the observation, the Cox method prescribes the following:
  - Compute the histograms for the fluorescence intensities in the control as well as the test dataset
  - Using the cell counts for each bin B_k in the two datasets, calculate the scores Z_k
  - Plot the scores Z_k and identify the ones that are beyond ±1.96
    - The ±1.96 range calculates the 95% confidence interval for a Gaussian observation
  - The bins where the scores exceed the confidence interval indicate the fluorescence intensity regions where the two datasets differ
- Note:
  - The comparison of the two datasets is carried out with no reference to the underlying intensity distributions

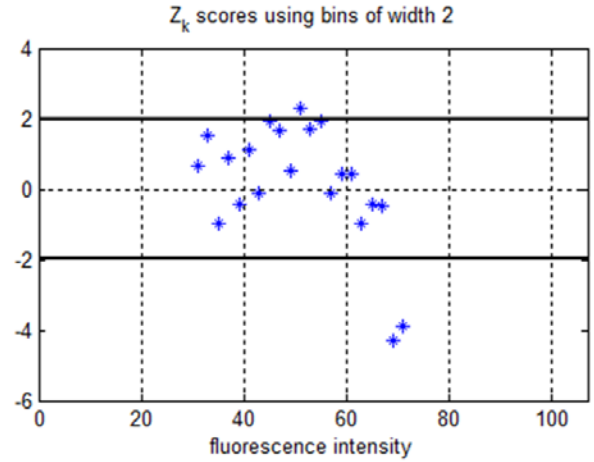
Cox Method

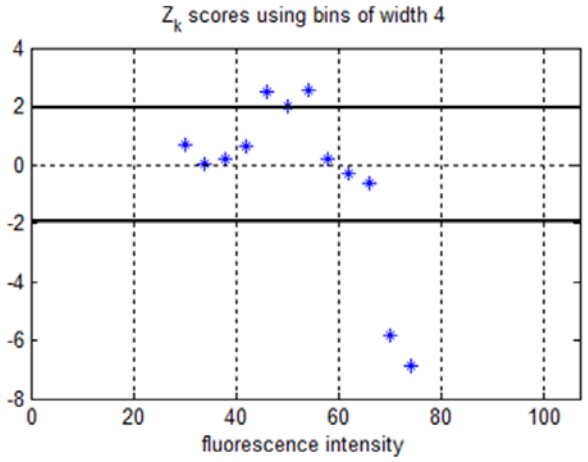
- Toy example:
  - Control and test datasets with 10000 cells each
    - 5% positives in the test dataset
  - Fluorescence intensities varying between 27 and 107
  - P values computed for the Z_k scores by squaring them and using the χ^2 distribution with one degrees of freedom

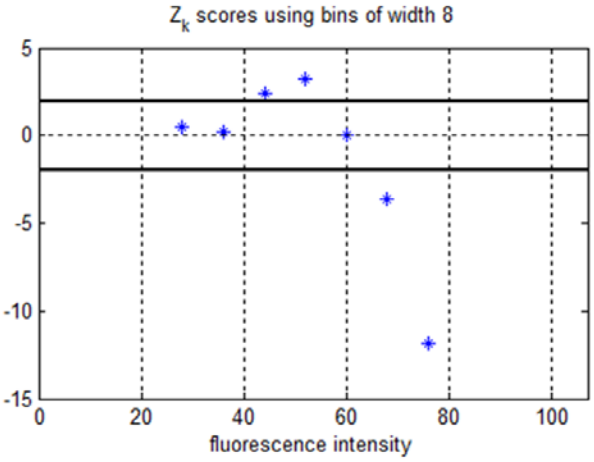


Cox Method

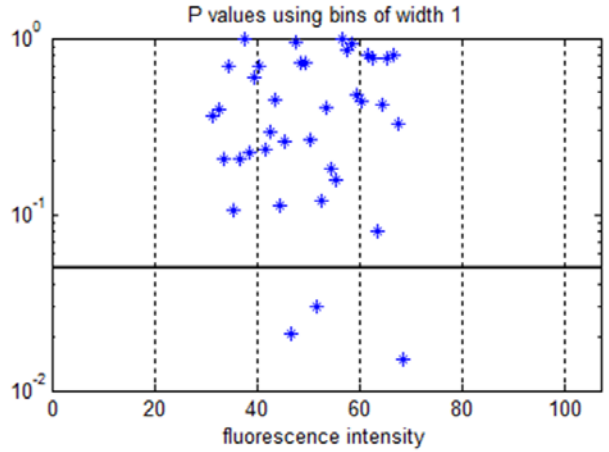

Z_k scores using bins of width 1
fluorescence intensity

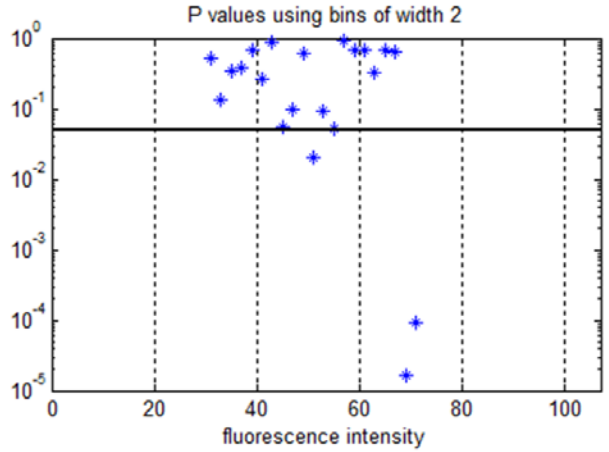

Z_k scores using bins of width 2
fluorescence intensity

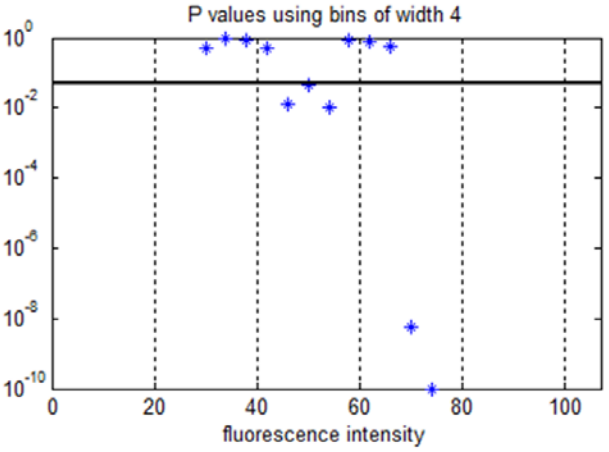

Z_k scores using bins of width 4
fluorescence intensity

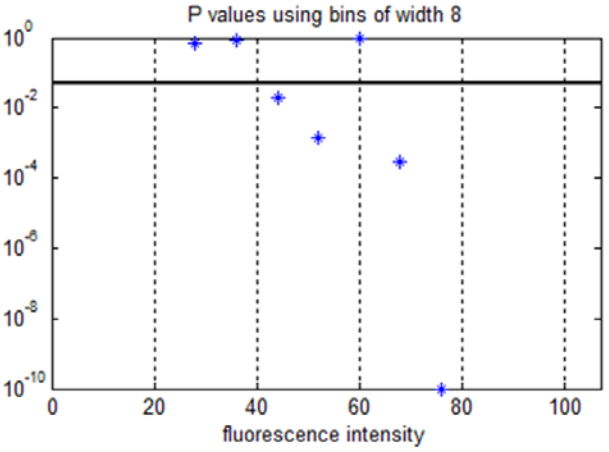

Z_k scores using bins of width 8
fluorescence intensity

Cox Method


P values using bins of width 1
fluorescence intensity


P values using bins of width 2
fluorescence intensity


P values using bins of width 4
fluorescence intensity


P values using bins of width 8
fluorescence intensity

Cox Method

- Remarks:
  - Regions on the intensity axis over which the two datasets differ are identified automatically
  - This can produce gates to separate the positives in the test dataset
    - As the union of the bins for which Z_k scores are below -1.96
  - However, the choice of the bins is arbitrary
    - Bins with smaller width have small cell counts, and the statistics is less powerful
      - Generally, bins with fewer than 20 cells from each of the two datasets are excluded from the analysis
    - Bins with larger width have larger cell counts for greater statistical power, but the intensity resolution is lost
    - This also implies that the statistical power between populated and unpopulated bins is not the same

Probability Binning

- The probability binning method modifies the Cox method to divide the fluorescence intensity data into bins that contain equal number of control dataset cells
  - resulting in bins of unequal sizes
- This ensures that
  - each bin is populated by a sufficient number of control cells, and
  - the maximum variance associated with the Poisson random variables governing the cell counts for each bin is capped

Probability Binning

- Mathematically, the probability binning method prescribes the following:
  - Given the control and the test datasets of fluorescence intensities, containing n^{cont} and n^{test} cells, respectively
  - For a number K of bins, define the bin B_k by

$$B_k = [I_{k-1}\ I_k]$$

where I_k denotes the intensity level of the $100k/K$'th percentile of the control dataset, for $k = 0,1, \ldots, K$

    - Each bin now contains about n^{cont}/K cells, or a fraction $1/B$ of cells from the control dataset
  - Determine the number n_k^{test} of the test dataset cells contained in B_k
  - Finally, compute the measure

$$\chi^2_{PB} = \sum_{k=1}^{K} \frac{\left(\frac{n_k^{cont}}{n^{cont}} - \frac{n_k^{test}}{n^{test}}\right)^2}{\frac{n_k^{cont}}{n^{cont}} + \frac{n_k^{test}}{n^{test}}}$$

Probability Binning

- Observation:
  - Under the null hypothesis where both the control dataset and the test dataset have been drawn from the same (unknown) distribution, χ^2_{PB} is approximately Gaussian with

$$\overline{\chi^2_{PB}} = \frac{K}{n}$$

and

$$Var(\chi^2_{PB}) = \frac{K}{n^2}$$

where $n = \min(n^{cont}, n^{test})$

  - This then allows computing a P value for the null hypothesis via

$$1 - P_N\left(\frac{\chi^2_{PB} - \frac{K}{n}}{\frac{\sqrt{K}}{n}}\right)$$

where P_N denotes the Gaussian cumulative distribution function with zero mean and unit variance

Probability Binning

- Following the same reasoning, the method also constructs a metric $T(\chi)$ to compare the control and the test distributions by

$$T(\chi) = \max\left(0, \frac{\chi_{PB}^2 - \frac{K}{n}}{\frac{\sqrt{K}}{n}}\right)$$

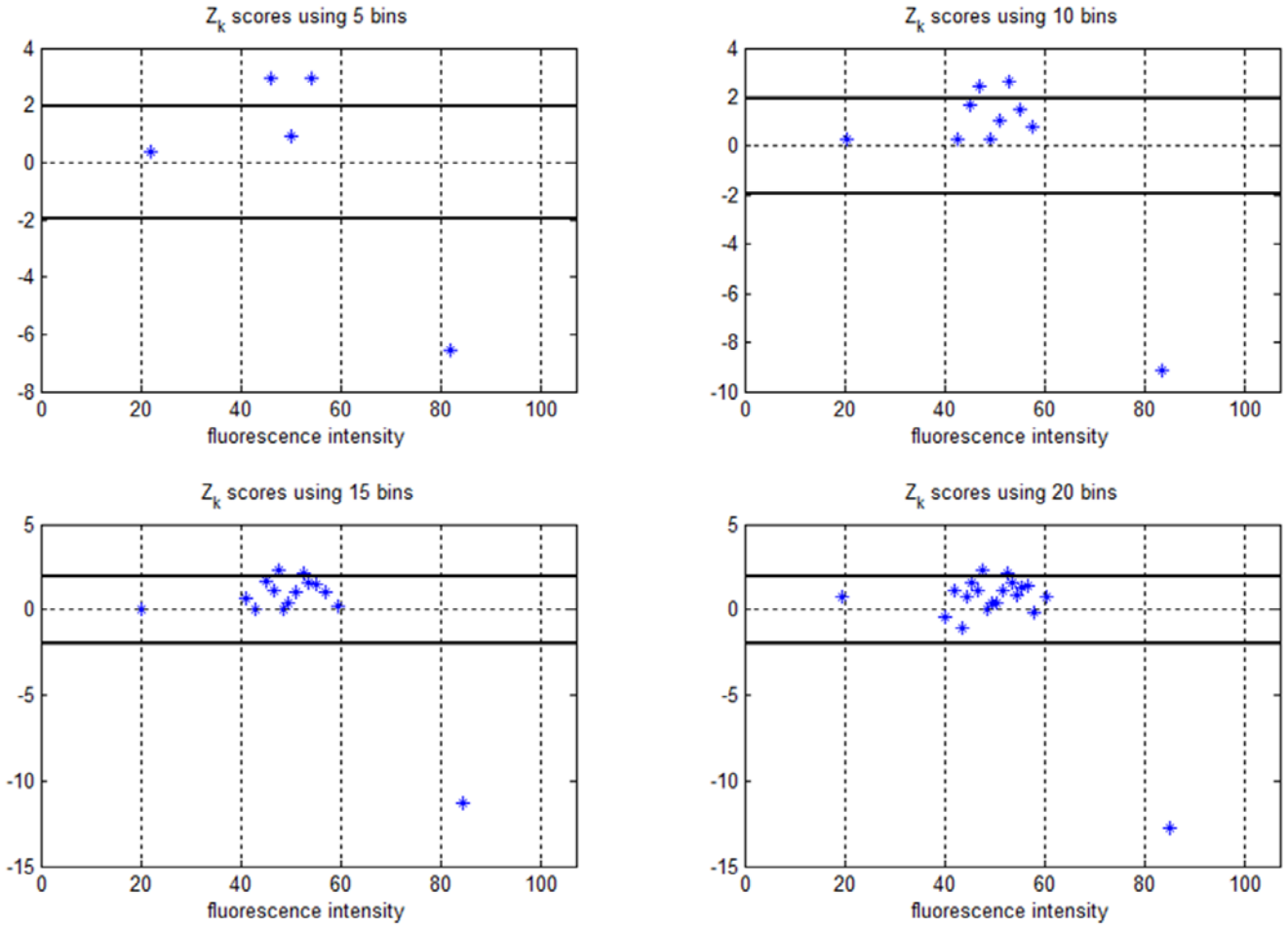
  - Small values of $T(\chi)$ indicate test datasets that are similar to the control dataset
- Notes:
  - An inverse monotonic relationship exists between the values of $T(\chi)$ and the P values
  - $T(\chi)$ can be used to rank several datasets in terms of similarity/dissimilarity to the control dataset

Probability Binning

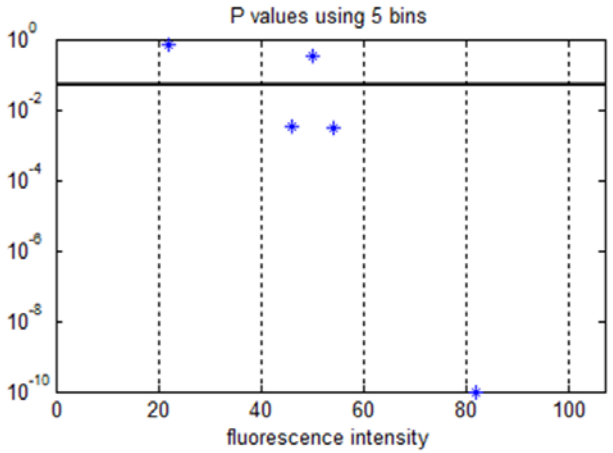
- For the toy problem:
  - Computed the histograms for varying number of bins $K = 5, 10, 15, 20$
  - Calculated the χ^2_{PB} score
  - Computed the P value using the corresponding mean $\overline{\chi^2_{PB}}$ and variance $Var(\chi^2_{PB})$
  - Calculated the metric $T(\chi)$

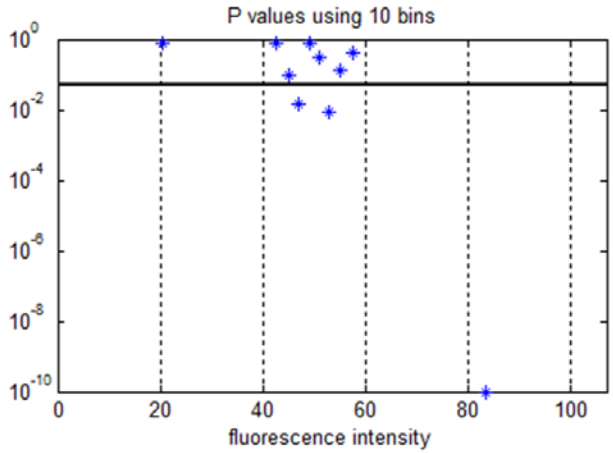
K	$\overline{\chi^2_{PB}}$	$Var(\chi^2_{PB})$	χ^2_{PB}	P value	$T(\chi)$
5	0.0005	5.0 10-8	0.0061	≈ 0	25.20
10	0.0010	1.0 10-7	0.0103	≈ 0	29.57
15	0.0015	1.5 10-7	0.0149	≈ 0	34.61
20	0.0020	2.0 10-7	0.0189	≈ 0	37.89

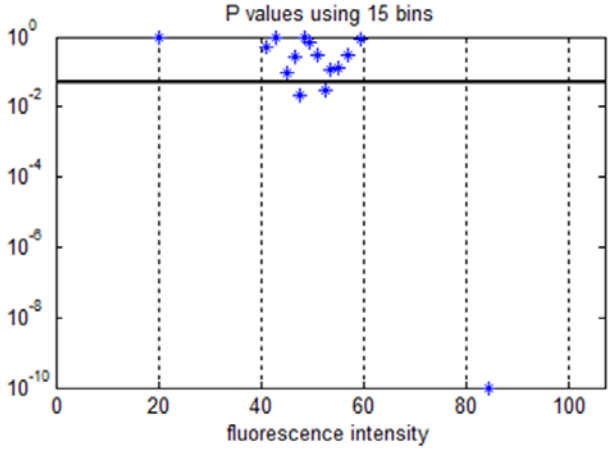
Probability Binning

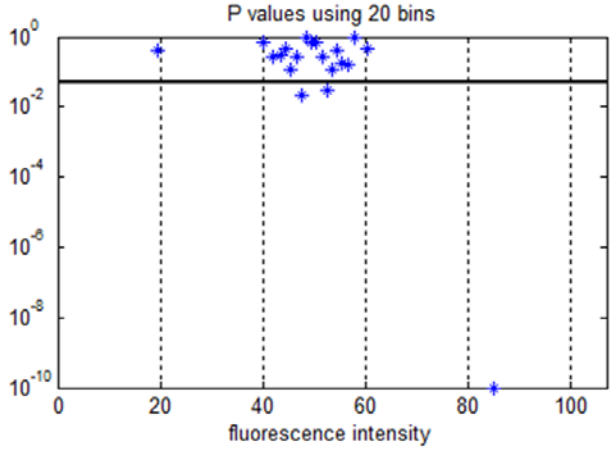


Probability Binning









Probability Binning

- Remarks:
  - The probability binning method detects the differences between the control and the test datasets with just 5% positives in the test dataset
  - While the bins are designed to hold a fixed fraction of the control dataset cells, the actual counts vary due to the integer nature of the fluorescence intensity data

Probability Binning on Multivariate Data

- The basic idea in Cox method as well as in probability binning is to group cells in bins according to their fluorescence intensities
- The number of cells in each bin from the control and the test datasets is then used to derive statistical significance measures related to the differences between the respective datasets for the corresponding bin
- The bins over which the differences are significant then indicate the regions of difference between the two datasets
  - In terms of the fluorescence intensities
  - That can be used for gating to label the positives in the test dataset
- The same idea can be used to multicolor datasets partitioned into high-dimensional bins
  - The underlying statistics is blind to the dimensionality of the data

Probability Binning on Multivariate Data

- Probability binning strategy to partition multicolor flow cytometry data:
  - Starting with one bin containing all cells
  - Carry out dyadic partitioning on the bin by:
    - Finding the fluorescence channel with the maximal intensity variance
    - Dividing the bin into two bins by the median intensity value along the fluorescence channel with the greatest variance
  - Carry out dyadic partitioning separately for each of the two daughter bins
- Notes:
  - The process is continued recursively until a stopping condition is met
    - For a given number of bins/divisions/…
  - This process produces bins hosting roughly equal number of cells
    - Up to some cumulative effects of integer-valued intensities

Probability Binning on Multivariate Data

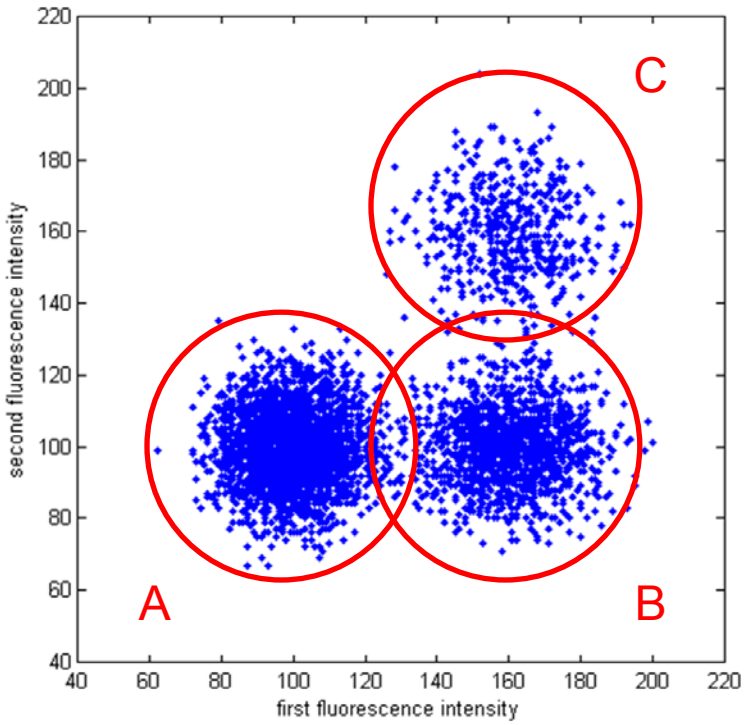
- On multicolor data, the probability binning method prescribes the following:
  - Given a control and a test dataset of multicolor fluorescence intensities
  - Partition the control dataset into dyadic bins
  - Carry out the same statistical analyses prescribed for univariate data to compute
    - the χ^2_{PB} scores,
    - the P values, and
    - the $T(\chi)$ metric
- Note:
  - Statistical analysis requires that bins have a sufficient number of cells from both datasets
  - Hence, the comparison will lack statistical power to distinguish the two datasets over bins that are poorly populated by either dataset

Probability Binning on Multivariate Data

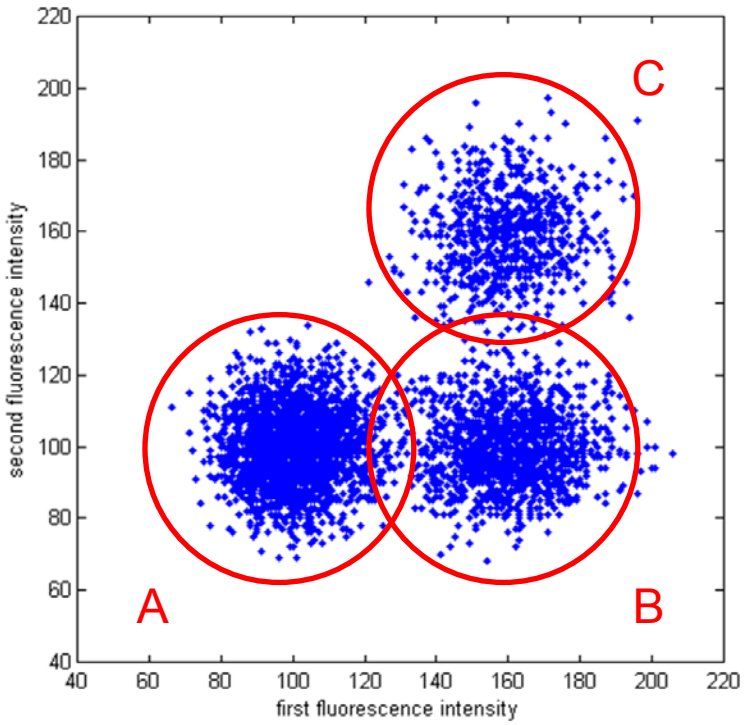
- Toy Problem:
  - Control and test datasets containing 2000 cells drawn from three different subpopulations
    - Control dataset:
      - 60% from subpopulation A
      - 30% from subpopulation B
      - 10% from subpopulation C
    - Test dataset:
      - 55% from subpopulation A
      - 30% from subpopulation B
      - 15% from subpopulation C
  - Bins determined on the control dataset
  - Statistical comparison carried out via Z_k scores
    - and the corresponding P values

Probability Binning on Multivariate Data

control dataset

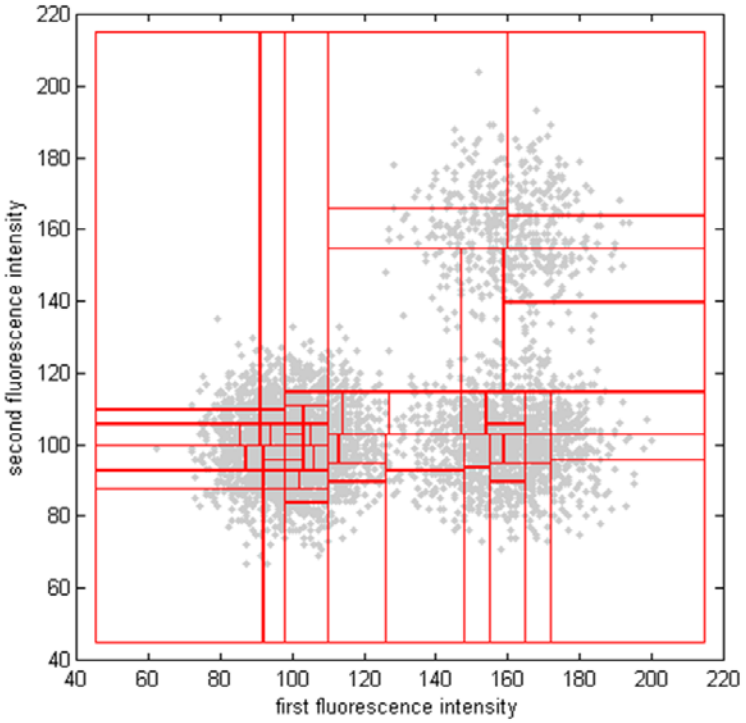


test dataset

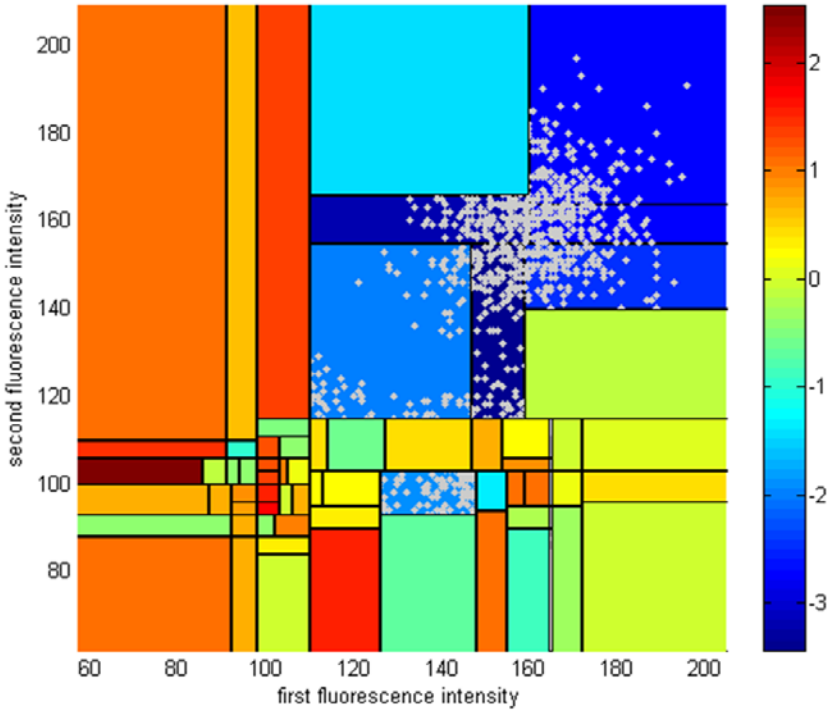


Probability Binning on Multivariate Data

bins on the control dataset



the Z_k scores



Summary

- Comparison of flow cytometry datasets relies on statistical test to identify
  - the fraction of “positive” cells that exist in the test dataset in contrast to a control dataset
  - the region marking the differences on the fluorescence intensity scale
- The statistical tests operating on binned intensities require bins defined in a way to contain a sufficient number of cells
  - for statistical power
- This, in turn, makes the analysis dependent on and exclusive to the bins themselves
  - The results depend on the way the bins are defined
  - Differences, when identified, apply to the whole bin, regardless of which portion of the bin contains cells of the contrasted datasets
- Automated methods that can contrast different flow cytometry datasets on the basis of individual cells is needed for
  - contrasting two datasets with high resolution
  - identifying the rare cells that appear in one dataset but not in the other