Selected Topics in Electrical Engineering: Flow Cytometry Data Analysis

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

Outline

- Comparing univariate cell distributions
  - Earlier methods
  - Maximum positive difference method and Overton cumulative histogram subtraction
  - Super-enhanced Dmax subtraction
  - The Kolmogorov-Smirnov algorithm

Motivation

- Flow cytometry aims to characterize cells in a population that differ from one another in terms of their biomarker profiles
  - Different cells possess different biomarkers (receptors) suitable to their role in the larger organism
- A critical component to this aim is to identify the cells that possess a specific biomarker, termed as **positives**, against the others, termed as **negatives**
- Given two sample distributions where one is the control dataset of negatives and the other a test dataset, the question is :

  Can we identify the cells that are positive in the test dataset?
- Note that an answer to this question requires the delineation of a region on the fluorescence intensity scale associated with the positive cells
- A related, but simpler question is:

  Can we predict the fraction of positive cells in the test dataset?

Earlier Methods

- Adaptive thresholding at a fixed rate of background detection:
  - Tantamount to *constant false alarm rate* detection rule in detection
  - A threshold is determined on a control dataset of background fluorescence
    - Typically, the threshold “detects” 2% of the control cells as exhibiting positive fluorescence
  - The threshold is then applied to the dataset of interest to identify the positive cells
    - And the percentages thereof

Earlier Methods

- Adaptive thresholding at a fixed rate of background detection (continued):
  - Mathematically, using
    - $P_{cont}(i)$: The empirical cumulative distribution of the control dataset at the intensity level i
    - $P_{test}(i)$: The empirical cumulative distribution of the test dataset at the intensity level i
  - A threshold T is identified such that

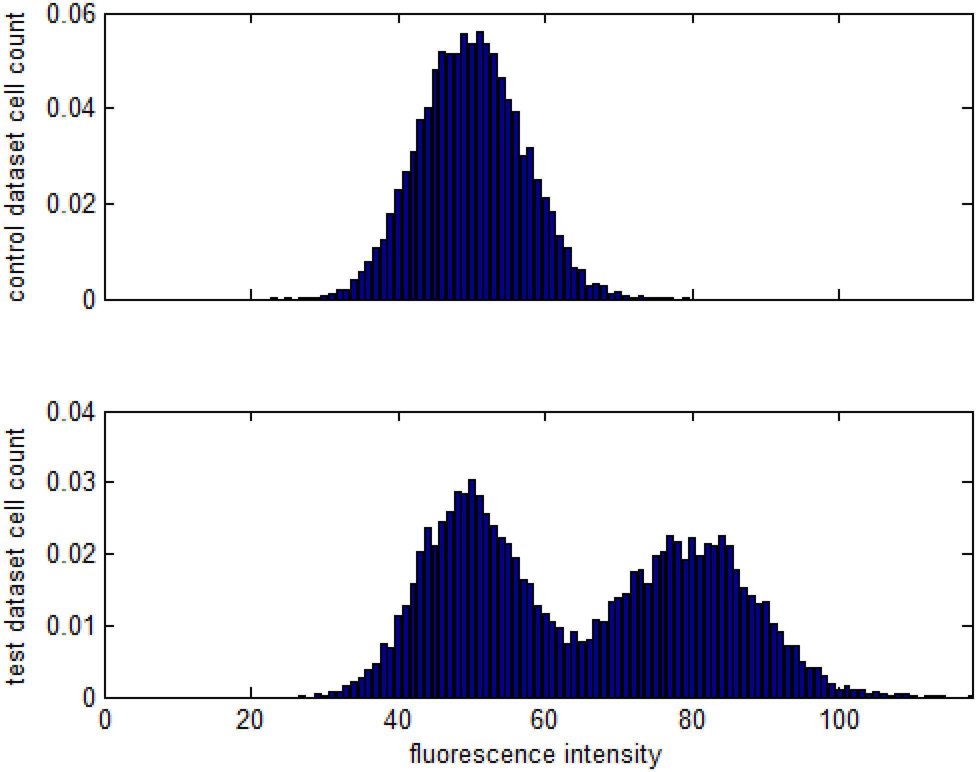
$$P_{cont}(T) = 0.98$$

  - The percentage of the positive cells in the test data is then given by

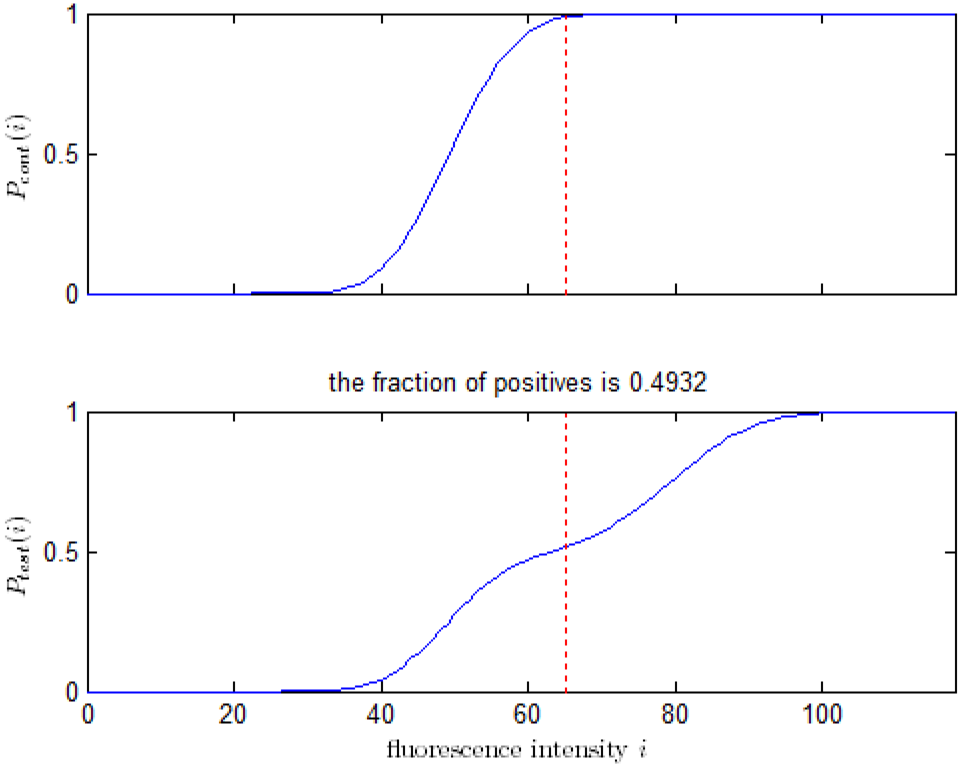
$$100\left(1 - P_{cont}(T)\right)$$

Earlier Methods

- Toy example:
  - Control dataset of 10000 cells
  - Test dataset of 10000 cells
  - A fraction of 0.50 of the test dataset drawn from the same distribution as the negatives of the control dataset
  - The remaining fraction of 0.50 drawn from a distinct distribution, and represent the positives



Earlier Methods



the fraction of positives is 0.4932

Earlier Methods

- Channel-by-channel subtraction:
  - Subtracts cell counts in each fluorescence channel (i.e. level) of a control histogram from those in a test histogram
    - The two histograms are normalized to have equal cell counts by a scalar normalizing factor
  - The channels with negative results are set to zero
  - The channels with positive counts characterize the fluorescence intensities with positive cells in the test histogram
  - The ratio of total (positive) differences to the test cell count calculates the percentage of positive cells

Earlier Methods

- Channel-by-channel subtraction (continued):
  - Mathematically, using
    - $p_{cont}(i)$ representing the normalized cell counts in the control dataset with intensity i
    - $p_{test}(i)$ representing the normalized cell counts in the test dataset with intensity i

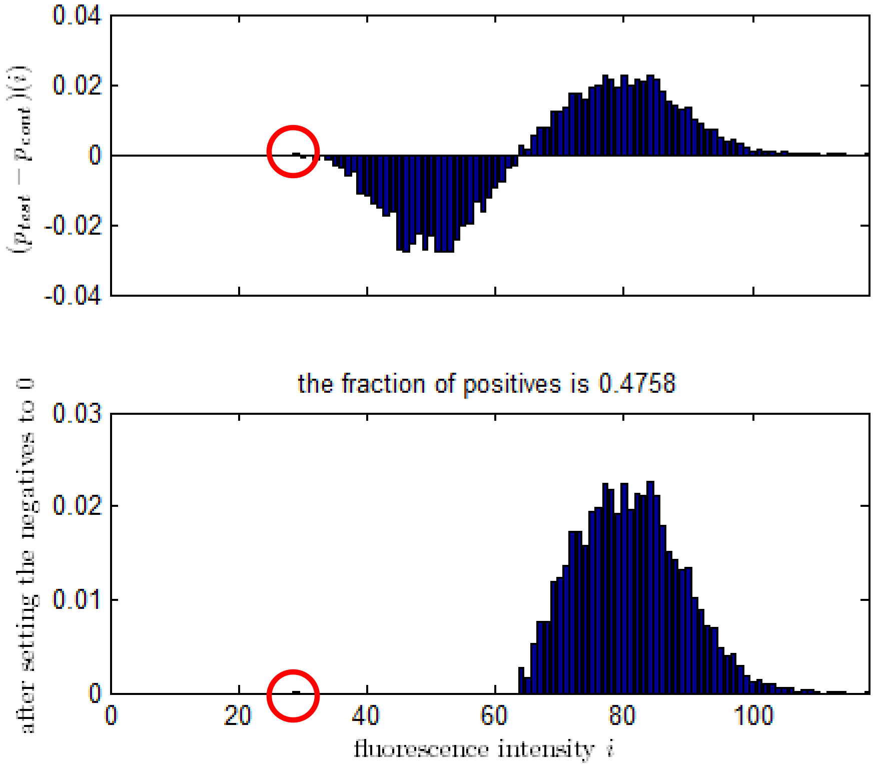
    such that

    $$P_{cont}(i) = \sum_0^i p_{cont}(j) \text{ and } P_{test}(i) = \sum_0^i p_{test}(j)$$
  - Letting

    $$R = \{i | p_{test}(i) > p_{cont}(i)\}$$
  - The percentage of positive cells is given by

    $$100 \sum_{i \in R} (p_{test}(i) - p_{cont}(i))$$

Earlier Methods



Method of Maximum Positive Difference

- This method identifies the largest difference between the control and test cell counts with intensities greater than equal to a threshold
  - Given a threshold intensity level, the positive cells are those that have fluorescence intensity greater than or equal to that level
  - The difference between the positive cell percentages between the test dataset and the control dataset can be computed for each threshold
  - Varying the threshold, the level providing the largest difference can be identified

Method of Maximum Positive Difference

- Mathematically,
  - For a given threshold T, the difference in consideration is

$$\left(1 - P_{test}(T)\right) - \left(1 - P_{cont}(T)\right) = P_{cont}(T)\text{-}P_{test}(T)$$

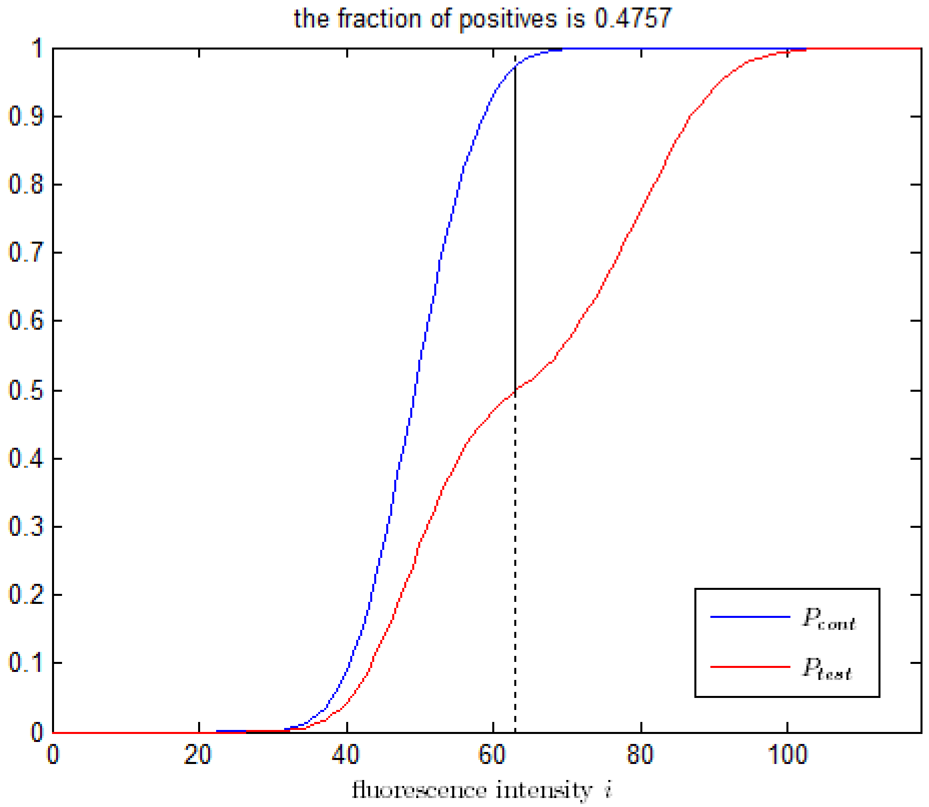
  - The maximum is obtained at the threshold T^* defined by

$$T^* = \underset{T}{\operatorname{argmax}}(P_{cont}(T) - P_{test}(T))$$

  - The percentage of the positive cells is then given by

$$100(P_{cont}(T^*) - P_{test}(T^*))$$

Method of Maximum Positive Difference



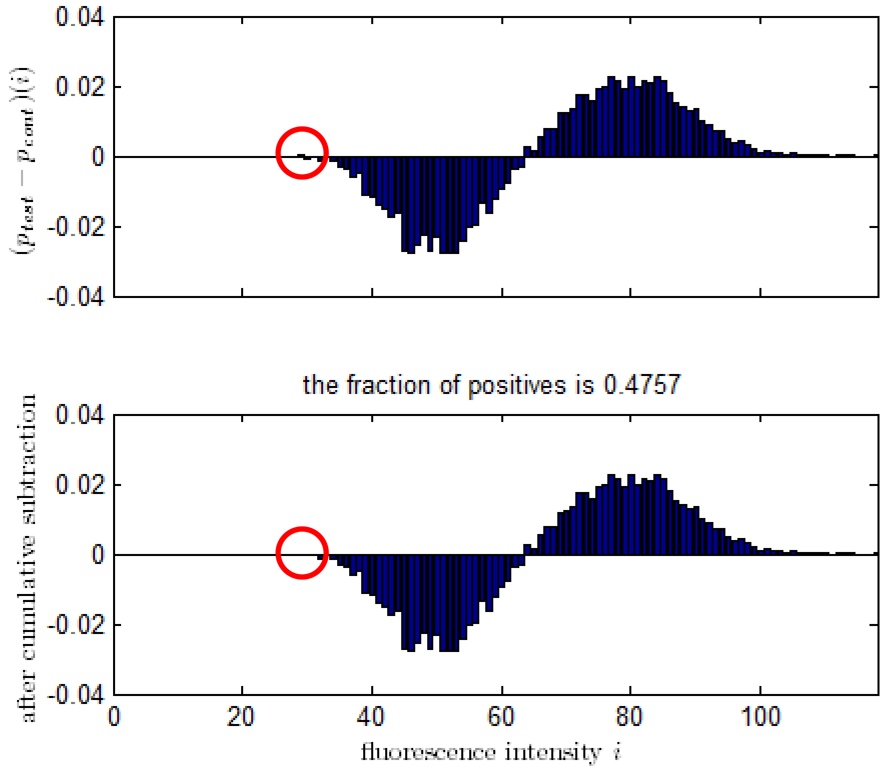
Overton Cumulative Histogram Subtraction

- This method refines the channel-by-channel subtraction
  - Straightforward subtraction finds the channels with positive or negative differences
    - Though the negative differences are replaced by zeros
  - But the ultimate goal is to find a threshold fluorescence intensity level (i.e. channel) to identify the fluorescence region associated with positive cells
    - as distinct from the negatives
  - So the method packs the negative differences onto the positive differences observed in the lower channels
  - Once finished,
    - residual negatives are set to zero, and
    - the sum of the positive differences computes the fraction of positives in the test dataset

Overton Cumulative Histogram Subtraction

- Mathematically:
  - The original difference $p_{test}(i) - p_{cont}(i)$ is modified so that
    - $p_{test}(i) - p_{cont}(i)$ is zero for $i < T$ for some value T, and
    - $p_{test}(i) - p_{cont}(i)$ is positive for $i \geq T$
  - This provides a best-guess estimate for the threshold T:
    - In the original case, earlier positive differences can be followed by negative differences due to noise
    - After the "correction," the differences are idealized so that the difference is always positive for $i \geq T$
      - No positive differences are followed by negatives

Overton Cumulative Histogram Subtraction



Super-Enhanced $D_{\max}$ Subtraction

- The $D_{\max}$ method:
  - Technically, $D_{\max}$ is defined as

$$D_{\max} = \max_{i}(P_{cont}(i) - P_{test}(i))$$

where $P_{test}(i)$ and $P_{cont}(i)$ are the cumulative distributions of the test and the control datasets, respectively, as before

  - The idea is based on the observation that $D_{\max}$ estimates the fraction of positive cells in the test dataset
    - Assuming that the positive and negative cell fluorescence distributions are distinct, $P_{cont}(i) - P_{test}(i)$ is maximal when all the negatives and none of the positives are covered in the interval $[0, i]$
    - Errors accumulate when the distributions overlap

Super-Enhanced $D_{\max}$ Subtraction

- The enhanced $D_{\max}$ method:
  - The original $D_{\max}$ method tends to underestimate the actual positive percentage in the test dataset, especially with non-zero overlap between the positives and the negatives
    - Q: Why?

    (Hint: Consider what $D_{\max}$ corresponds to in a plot of $P_{test}(i)$ versus $P_{cont}(i)$)
  - A correction can be obtained by scaling it using the value of the cumulative distribution of the control dataset at the corresponding fluorescence intensity
  - Mathematically, this prescribes using

$$100\,\frac{D_{\max}}{P_{cont}(T)}$$

to compute the positive percentage where

$$T = \underset{i}{\operatorname{argmax}}(P_{cont}(i) - P_{test}(i))$$

Super-Enhanced $D_{\max}$ Subtraction

- The super-enhanced $D_{\max}$ subtraction:
  - It can be shown that the actual expression for the positive fraction is equal to

$$\frac{D_{\max} + P_{pos}(T)}{P_{cont}(T)}$$

  where

$$P_{test}(T) = P_{pos}(T) + P_{neg}(T)$$

    - Hence, as T grows large, $D_{\max}$ goes to zero, $P_{cont}(T)$ goes to one, and the ratio above converges to the fraction of positives in the test dataset
  - Further correction on the enhanced $D_{\max}$ subtraction method entail estimating $P_{pos}(T)$

Super-Enhanced $D_{\max}$ Subtraction

- The super-enhanced $D_{\max}$ subtraction (continued):
  - Given the fluorescence intensity T at the maximum difference
  - Suppose new cumulative distributions are formed by limiting the range of fluorescence intensities to within $[0, T]$

$$P'_{cont}(i) = \frac{P_{cont}(i)}{P_{cont}(T)}$$

and

$$P'_{test}(i) = \frac{P_{test}(i)}{P_{test}(T)}$$

  - Now, repeating the enhanced $D_{\max}$ subtraction method using $P'_{cont}(i)$ and $P'_{test}(i)$ provides a maximum difference of $D'_{\max}$ at T'

Super-Enhanced $D_{\max}$ Subtraction

- The super-enhanced $D_{\max}$ subtraction (continued):
  - Furthermore, the fraction

$$\frac{{D'}_{\max}}{{P'}_{cont}(T')}$$

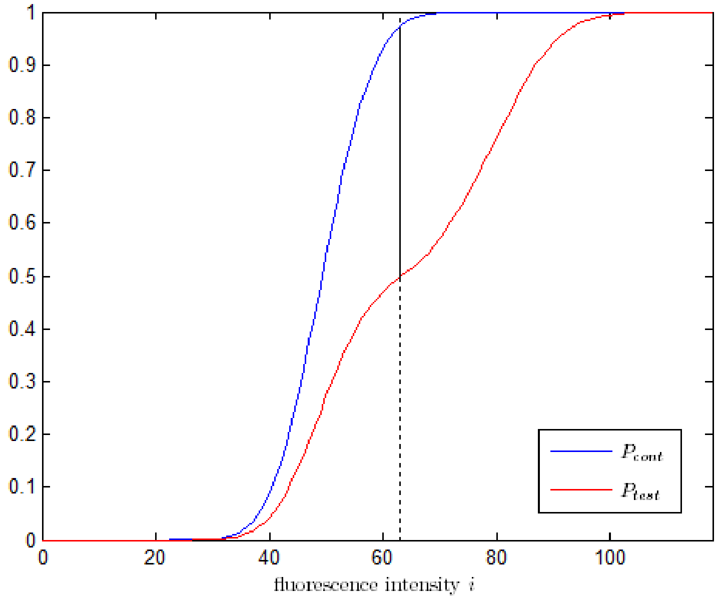
estimates $P_{pos}(T)$

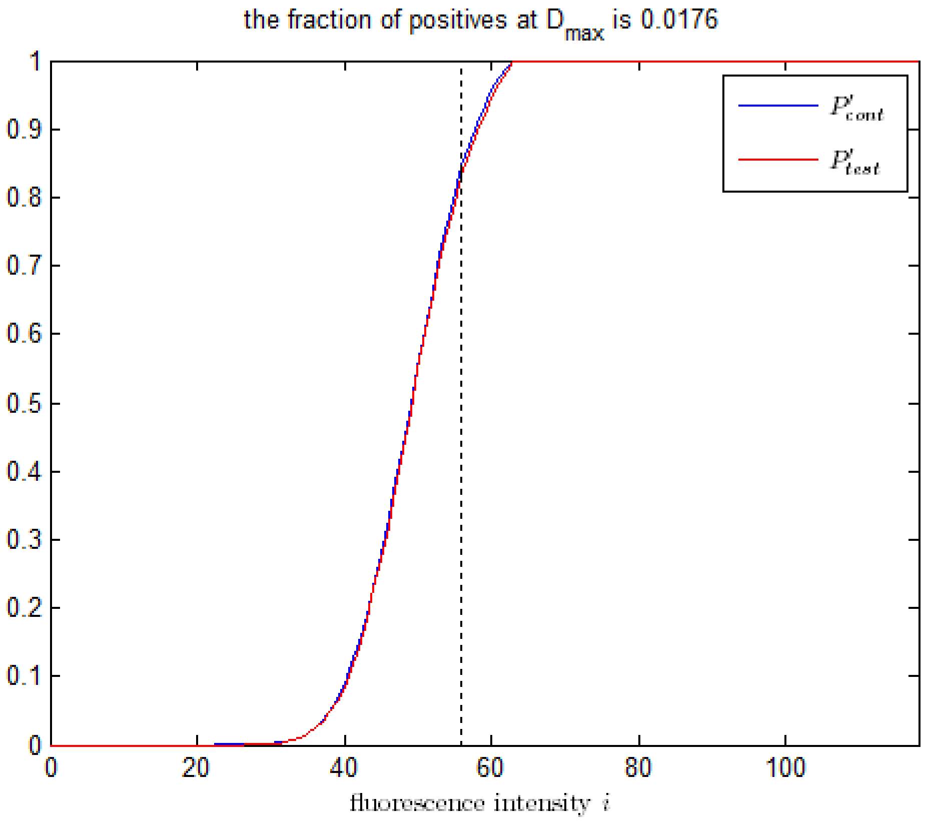
  - Using this estimate in the earlier expression provides

$$100\,\frac{D_{\max}+\dfrac{{D'}_{\max}}{{P'}_{cont}(T')}}{P_{cont}(T)}$$

to compute the fraction of positive cells in the test dataset

Super-Enhanced $D_{\max}$ Subtraction





The final estimate of the fraction of positive cells in the test dataset is:

$$\frac{0.4757 + 0.0176}{0.9733} = 0.5068$$

The Kolmogorov-Smirnov Algorithm

- This method is based on the Kolmogorof-Smirnov test to see if two samples are drawn from the same distribution:
  - Two datasets are given, one control and the other test, with n_{cont} and n_{test} samples respectively
  - Calculate the KS statistic

$$K = \sqrt{\frac{n_{cont} n_{test}}{n_{cont} + n_{test}}} D_{\max}$$

  - Under the null hypothesis where the samples in both datasets are drawn from the same distribution, K is governed by the Kolmogorov distribution with

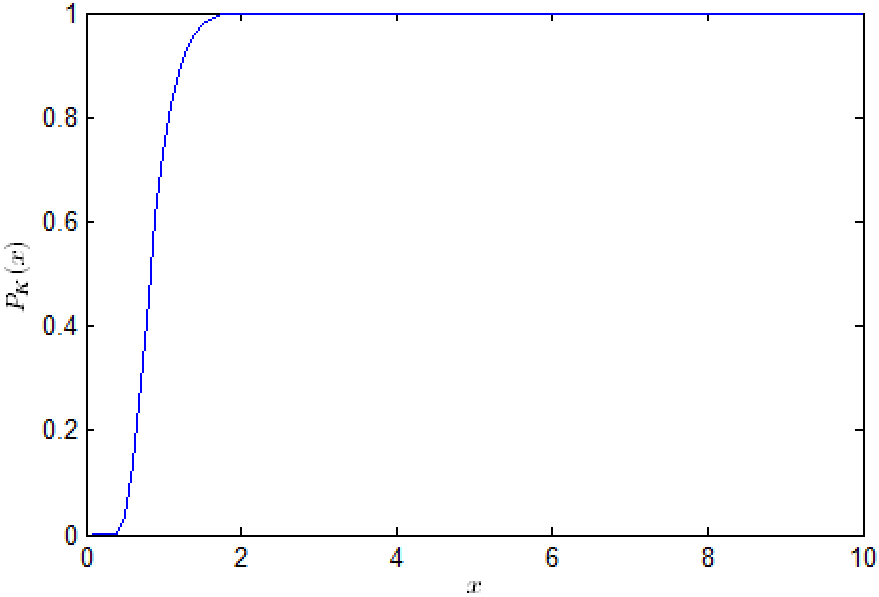
$$P_K(x) = \Pr\{K \leq x\} = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1} \exp(-2k^2x^2)$$

for large n_{cont} and n_{test}

  - The null hypothesis is rejected if $P_K(K) > 1 - \alpha$ for the observed K, where α represents a desired level of statistical significance

The Kolmogorov-Smirnov Algorithm

- For the toy example:
  - $n_{cont} = n_{test} = 10000$
  - $K = \sqrt{\frac{n_{cont} n_{test}}{n_{cont}+n_{test}}} D_{\max} = \sqrt{5000} \cdot 0.4757 = 33.6371$
  - $P_K(33.6371) \cong 1$ ➔ the P-value is practically zero!!



The Kolmogorov-Smirnov Algorithm

- Remarks:
  - The Kolmogorov-Smirnov algorithm carries out a statistical test to determine the confidence interval at which two cell distributions are different
  - It does not, in essence, delineate a region of fluorescence intensities over which they differ
  - On the other hand, it uses D_{max} to determine the confidence interval, that can be used to identify the fraction of positive cells in the test dataset

Summary

- Many different but related methods exist to predict the fraction of positive cells in a test dataset in contrast to a control dataset of all-negative cells
- While these methods compute parameters linked to critical fluorescence intensity levels, they do not directly delineate the regions in the fluorescence intensity scale associated with the positive cells
  - Though it is clear that they are the cells in the test dataset with greater fluorescence intensity
- Regions of difference between the fluorescence intensity distributions of two samples, or gates, can be identified using the alternative method of probability binning