Statistical Techniques for Data Science & Robotics

Week 6

Outline

- Finalize the Fisher's Discriminant
- Empirical CDF
- Non-parametric Tests
  - Kolmogorov-Smirnov test
  - Wilcoxon tests

Quiz

Fisher's discriminant

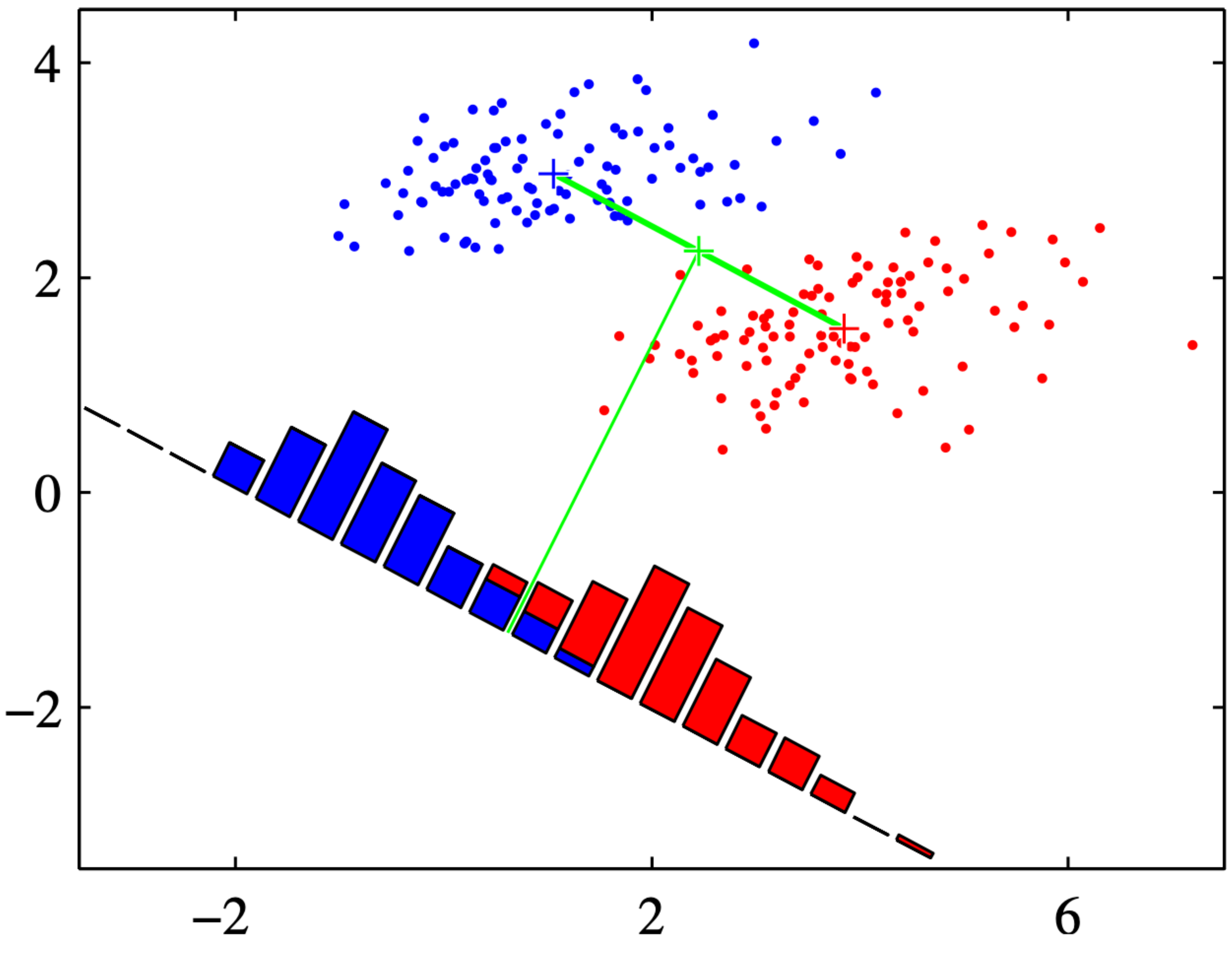
- In the Fisher's discriminant
  - $J(w) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$
  - Both numerator and denominator depend on w
  - **Task:** Write the explicit dependency for the numerator $(m_2 - m_1)^2$ as a function of w (Hint: use matrix notation)

Fisher's linear discriminant

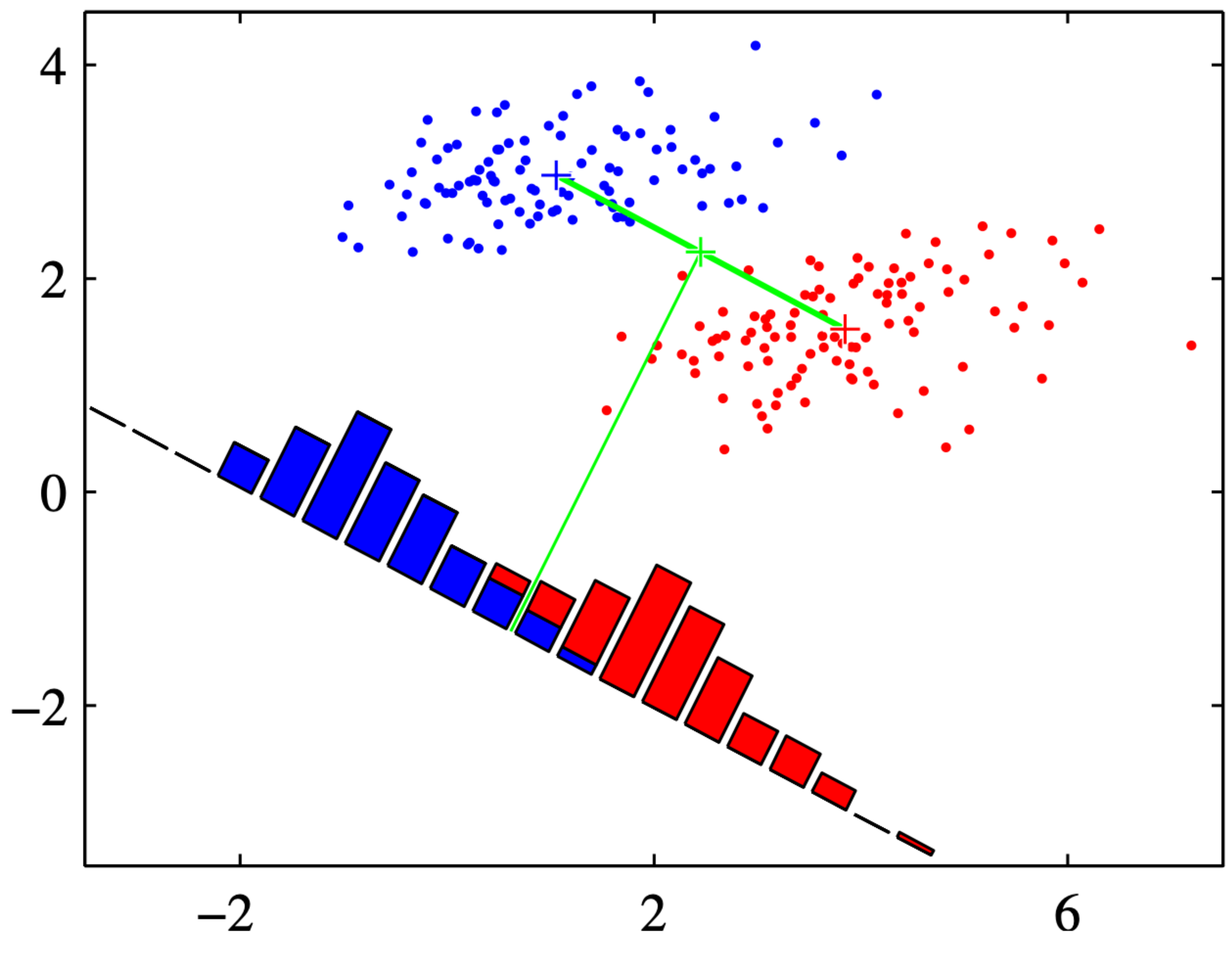
two classes, k = 2

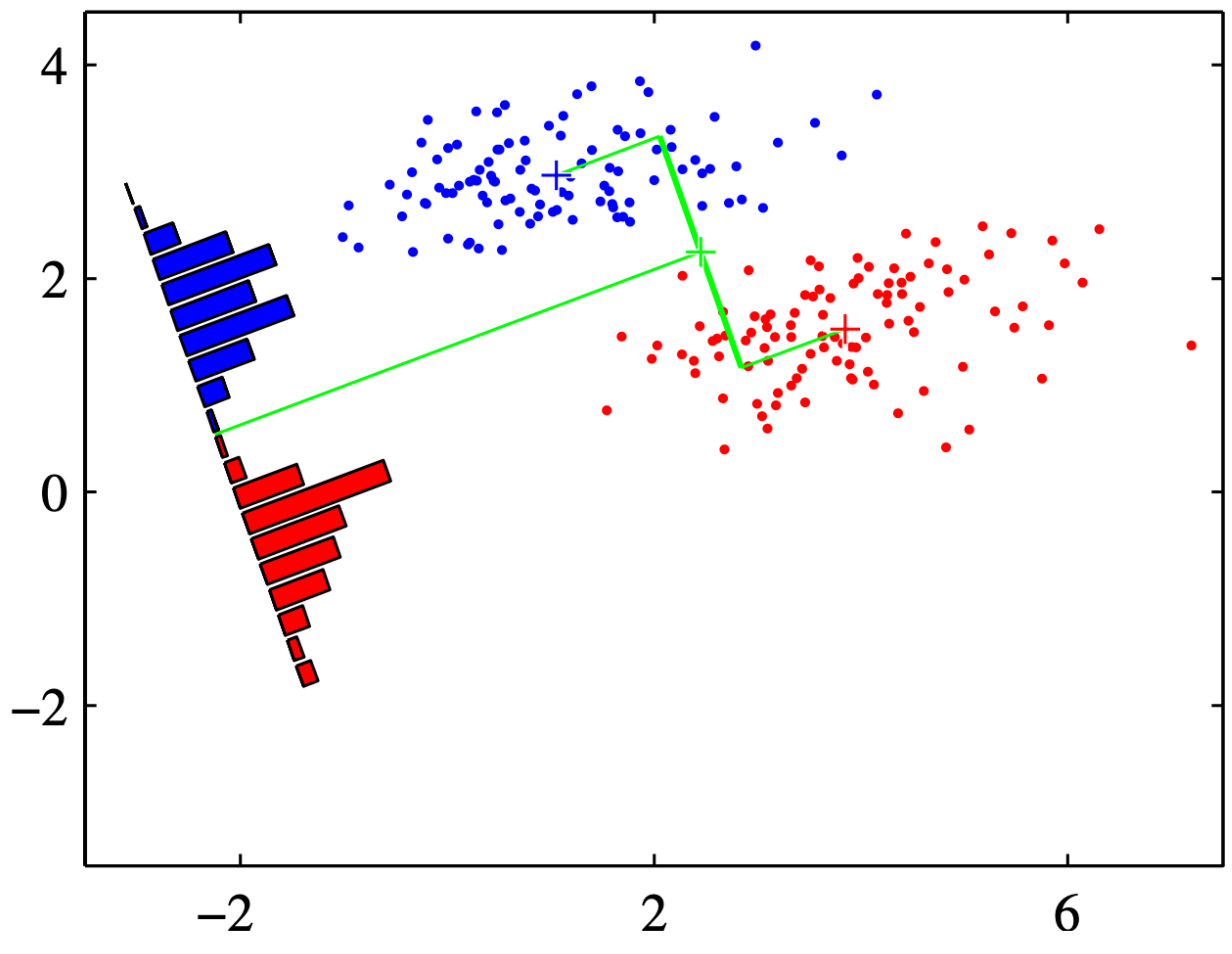
- Project high dimensional data to a line
  - and How to define that line?

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$



Which line is better?





Derivation of Fisher's Linear Discriminant

actually, only a direction of the line, defined by the vector **w**

$$J(w) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

Break

Non-parametric Tests

Nonparametric methods

- In situations where the assumptions of distribution are not supported we use the so-called "nonparametric methods"
- Nonparametric methods require fewer assumptions about the underlying distribution
- They are, in general, robust to outliers
- Often the only methods available when the data consist of ranks, ordinal, or categorical data

Kolmogorov-Smirnov Test

Problem

- You have a sample X_n and you would like to check whether it came from some distribution with a CDF F, or not.

- (for continuous distributions F only)

Empirical CDF, ECDF, or $\hat{F}_n$

- Suppose that the CDF (F) of the population is unknown
  - Can we estimate it?
- Why not use an estimate of F based on the sample $(x_1, \ldots, x_n)$ at hand?
- The most well known estimate of F is the empirical distribution function $\hat{F}_n$

  - $\hat{F}_n(x) = \dfrac{\sum_{i=1}^{n} I(x_i \leq x)}{n}$

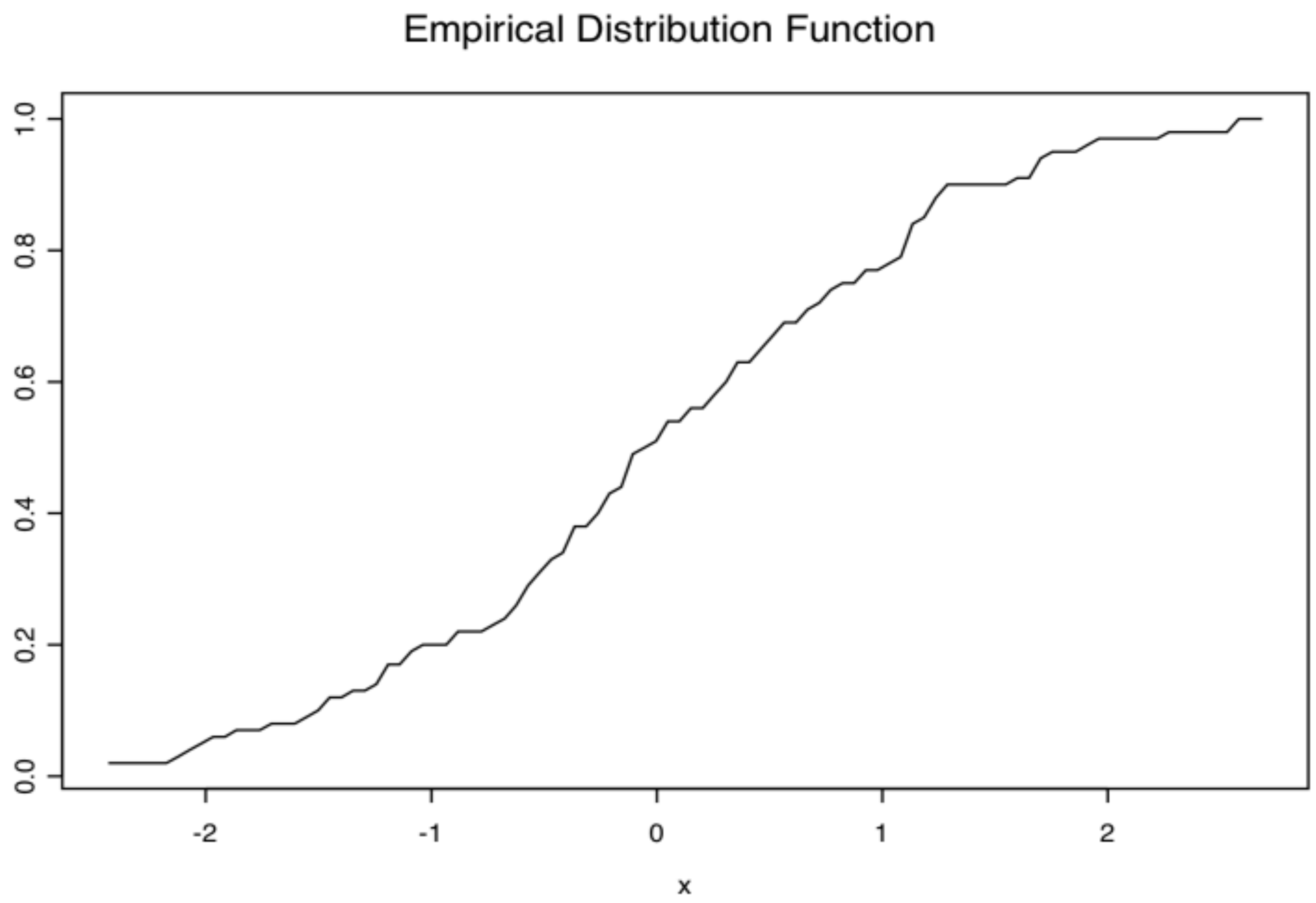
ECDF

- So, we use a sample to model the CDF

- $\hat{F}_n \to F, \quad n \to \infty$

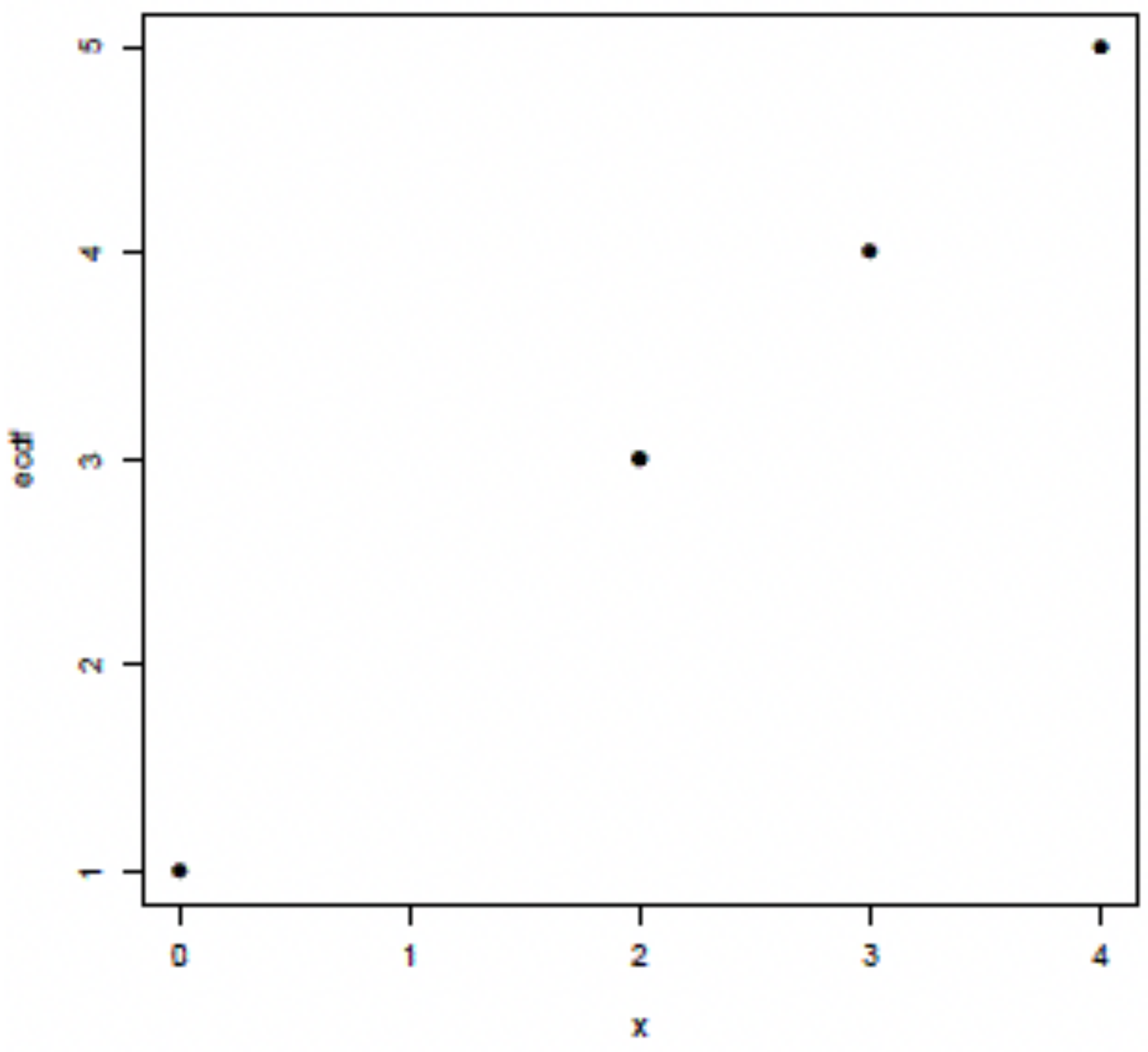
$$n\hat{F}_n(x) = \sum I(x_i \leq x) \sim Bin(n, F(x))$$

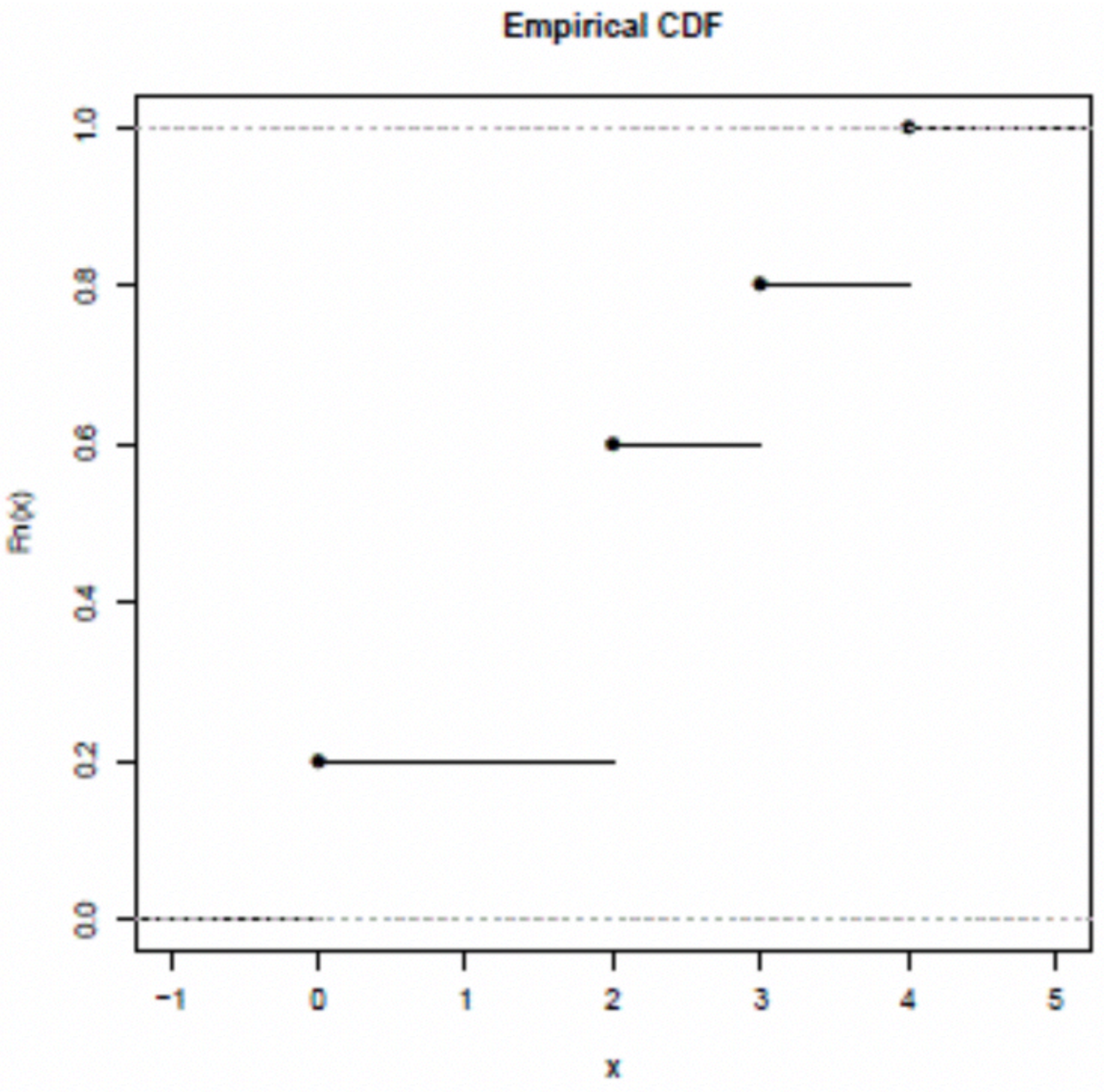
$$\sup |\hat{F}_n(x) - F(x)| \to 0, \quad n \to \infty$$



Example of ECDF

Dataset contains 5 values: (4, 0, 2, 3, 2)





Exercise

Check whether $\mathbb{E}(\hat{F}_n(x)) = F(x)$ for a fixed value x

Solution

- $Y = I(x_i \leq x)$ is a Bernoulli r.v. (either 1 or 0)

- $Y \sim Bern(p)$, where $p = F(x)$

- $n\hat{F}_n(x) = \sum_{i=1}^{n} I(x_i \leq x)$ is Binomial, $Bin(n, p)$

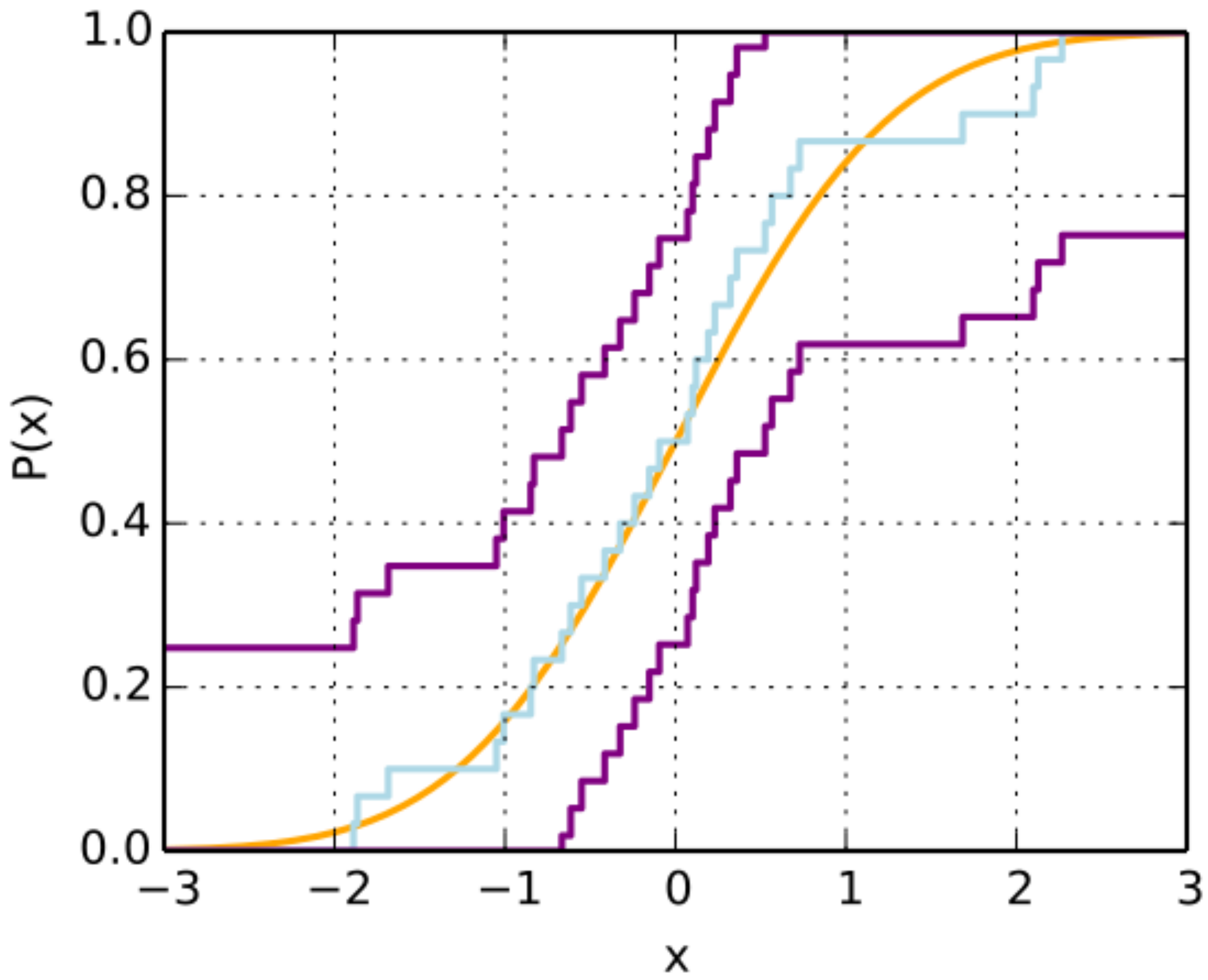
- $\mathbb{E}(n\hat{F}_n(x)) = np = nF(x)$ => $\mathbb{E}(\hat{F}_n(x)) = F(x)$

Dvoretzky-Kiefer-Wolfowitz (DKW) inequality

$$\sup |\hat{F}_n(x) - F(x)| \to 0, \quad n \to \infty$$

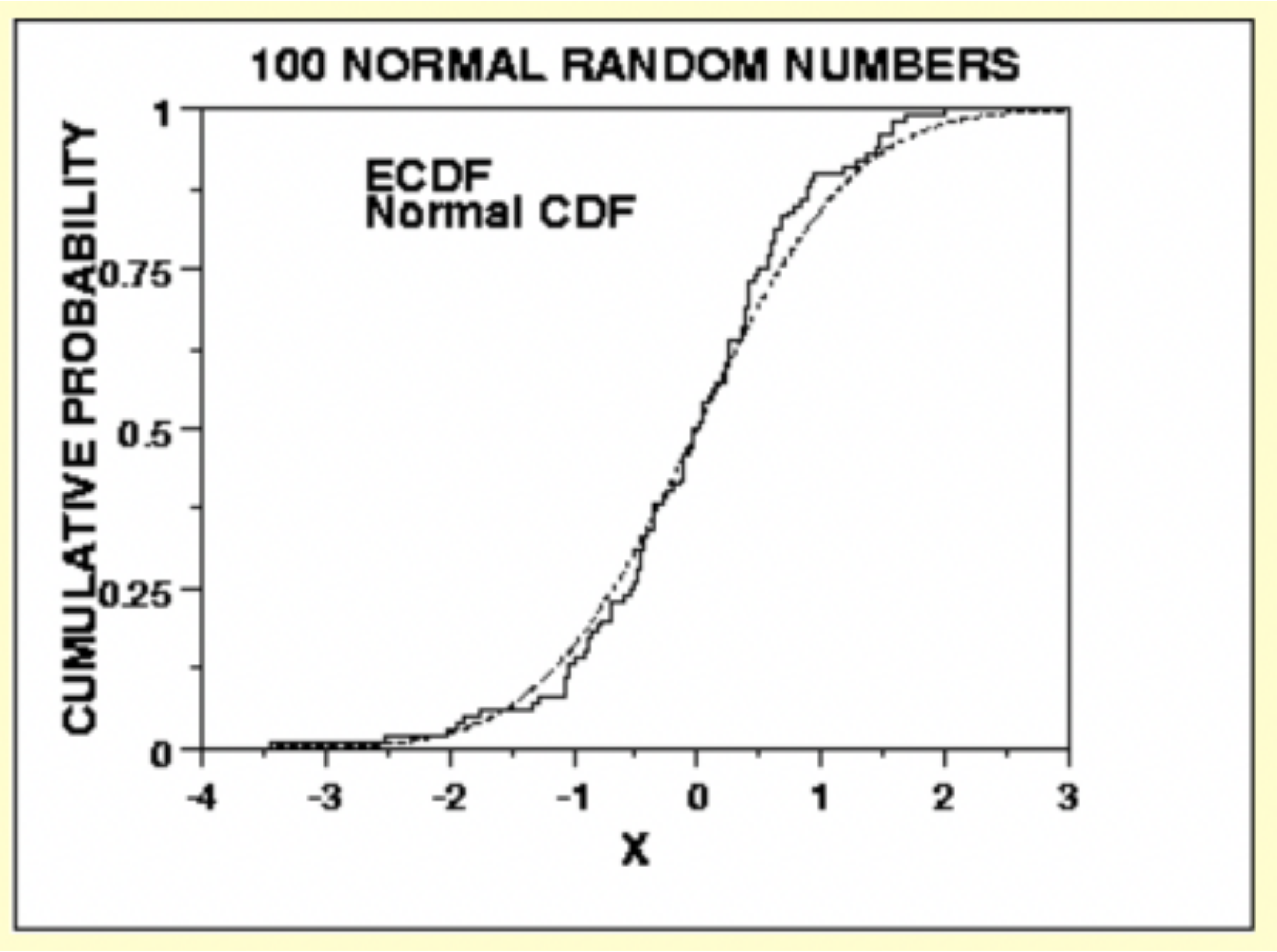
- for every $\varepsilon > 0$
- $\mathbb{P}\left(\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}$
- $\hat{F}_n(x) - \varepsilon \leq F(x) \leq \hat{F}_n(x) + \varepsilon$

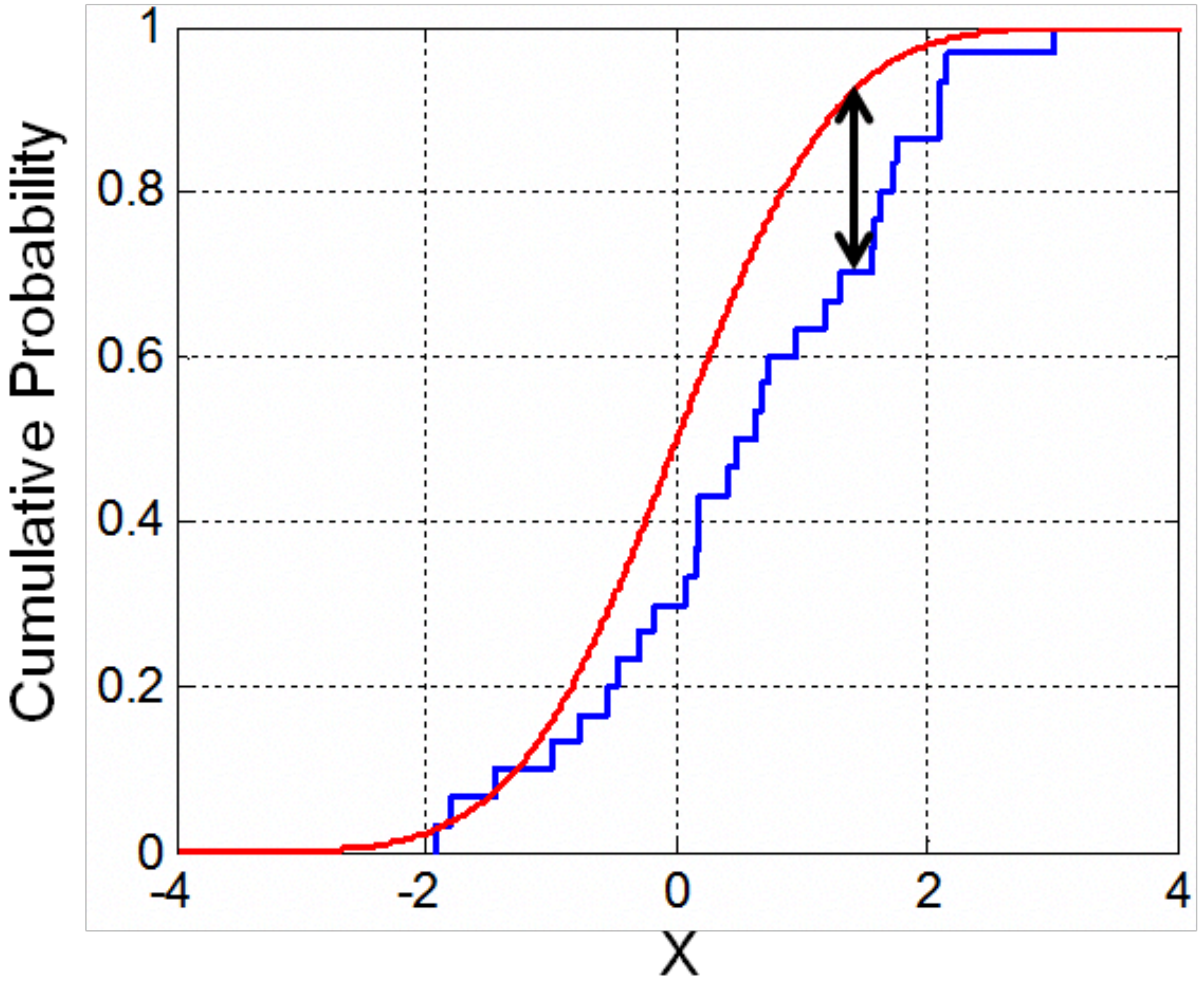
where $\varepsilon = \sqrt{\frac{\ln \frac{2}{\alpha}}{2n}}$, given a value of alpha



KS test idea

- The K-S test is used to decide if a sample comes from a population with a specific distribution.
- Given a sample, build a ECDF $\hat{F}_n$
- Compare the ECDF and the CDF (F)





Definition and Statistic

- H_0 : the data follow a specified distribution
- H_a : the data do not follow a specified distribution
- $D = \sup_x |F_n(x) - F(x)|$

Exercise (Homework)

Check whether for a fixed value x

$$Var(\hat{F}_n(x)) = \frac{F(x)(1 - F(x))}{n}$$

Break

Kolmogorov-Smirnov

(two samples)

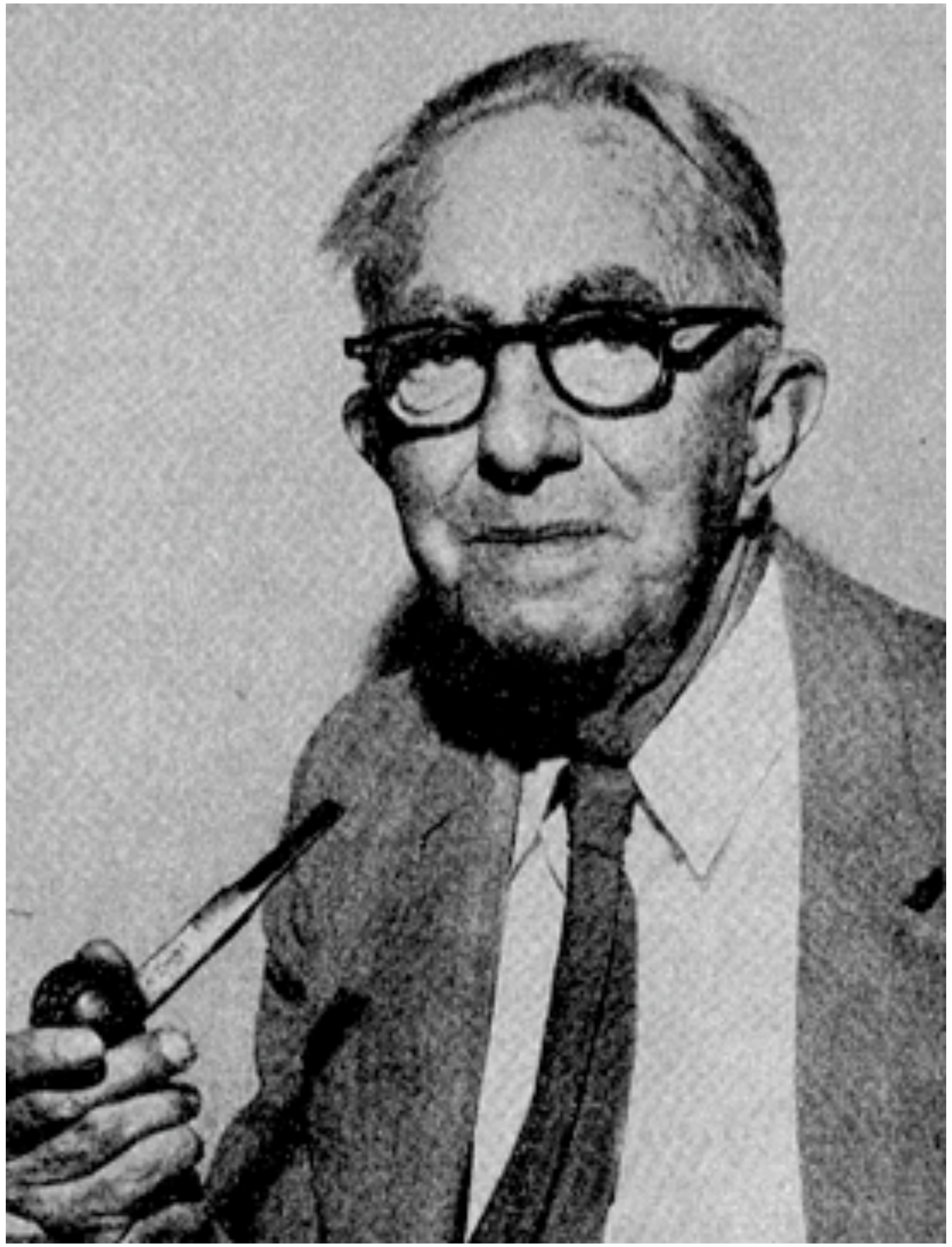
Two sample KS test

- There is also a two sample version of the test that checks that samples follow the same distribution
  - $D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$
- Null Hypothesis is rejected when
  - $D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}}$
- where
  - $c(\alpha) = \sqrt{-\frac{1}{2}\ln(\alpha/2)}$

Homework

- Perform Monte-Carlo simulation, and see that the probability of the event "true CDF falls outside the confidence band" is less than α:
  - Pick true underlying distribution
  - Generate a sample of size n
  - Compute ECDF and a confidence band (using DKW)
  - Test whether true CDF falls outside confidence band for any x (event A)
  - Repeat m times, calculate the frequency of the "event A"

Wilcoxon tests



The sign test

The sign test

- Let X be a continuous-type random variable and let m denote the median of X.

- To test the hypothesis
- H_0: $m = m_0$ against an appropriate alternative hypothesis,
- we can use **a sign test**.

- Let Y equals to the number of negative values among differences:
  - $X_1 - m_0, X_2 - m_0, \ldots, X_n - m_0$
- Y has a **binomial distribution** $Bin(n, 1/2)$ under H_0 and is the statistic for the sign test

Wilcoxon signed rank test

Wilcoxon signed rank test. Definition

- The test takes into account the magnitude of the differences
  - $X_1 - m_0, X_2 - m_0, \dots, X_n - m_0$
- We rank the absolute values: $|X_1 - m_0|, |X_2 - m_0|, \dots, |X_n - m_0|$
- R_k denotes the rank of $|X_k - m_0|$ in the sorted list. Note, that $R_k \in \{1,\dots,n\}$
- If the difference $X_k - m_0$ is negative, then multiply R_k by -1
- The Statistic of the test (sum of signed ranks):
  - $W = \sum_{k=1}^{n} R_k$

Wilcoxon signed rank test. Requirements

- Test requires:
  - continuous data;
  - univariate;
  - symmetric distribution of the differences around the median

Example

Suppose the lengths of $n = 10$ sunfish are

$$x_i:\ 5.0\ \ 3.9\ \ 5.2\ \ 5.5\ \ 2.8\ \ 6.1\ \ 6.4\ \ 2.6\ \ 1.7\ \ 4.3$$

We shall test H_0: $m = 3.7$ against the alternative hypothesis H_1: $m > 3.7$. Thus, we have

$x_k - m_0$:	1.3,	0.2,	1.5,	1.8,	−0.9,	2.4,	2.7,	−1.1,	−2.0,	0.6
$\lvert x_k - m_0\rvert$:	1.3,	0.2,	1.5,	1.8,	0.9,	2.4,	2.7,	1.1,	2.0,	0.6
Ranks:	5,	1,	6,	7,	3,	9,	10,	4,	8,	2
Signed Ranks:	5,	1,	6,	7,	−3,	9,	10,	−4,	−8,	2

- W = ?
  - the closer W to zero, the more chance for failing to reject hull hypothesis

Normal approximation of Wilcoxon distribution

- The sampling distribution of W converges to a normal distribution.
- As
  - $Var(W) = \dfrac{n(n+1)(2n+1)}{6}$
- Thus, if n > 20:
  - $Z = \dfrac{W}{\sqrt{n(n+1)(2n+1)/6}}$
- And we can use Z-table for p-values

Wilcoxon rank-sum test

Rank-sum test. Definition

we are interested in question: If two means of two distributions are equal?

so, we have 2 samples with different sizes (n_1, n_2)

$$H_0 : \tilde{\mu}_1 = \tilde{\mu}_2$$

We arrange the $(n_1 + n_2)$ observations and assign them ranks from 1 to $(n_1 + n_2)$

w_1 - sum of ranks in the smaller sample

w_2 - sum of ranks in the bigger sample

Statistics: $u_1 = w_1 - \dfrac{n_1(n_1+1)}{2}$; and $u_2 = w_2 - \dfrac{n_2(n_2+1)}{2}$; or $u = \min(u_1, u_2)$

Rank-sum test. Decision making

H_0	H_1	Compute
$\tilde{\mu}_1 = \tilde{\mu}_2$	$\tilde{\mu}_1 < \tilde{\mu}_2$	u_1
	$\tilde{\mu}_1 > \tilde{\mu}_2$	u_2
	$\tilde{\mu}_1 \neq \tilde{\mu}_2$	u

- If statistic (for a corresponding case) is small enough, then reject the H_0

Case Study Discussion (aka Midterm)