Foundation of Intelligent Systems, Part I

Statistical Learning Theory

mcuturi@i.kyoto-u.ac.jp

Previous Lecture : Probabilistic Setting, Loss, Risk

- We observe the outcomes of a pair of random variables (X, Y).
- **Probability** $\boldsymbol{P}$ for couples $(\mathbf{x}, y)$ on $\mathbb{R}^d \times \mathcal{S}$, with density $\boldsymbol{p}$

$$\boldsymbol{p}(X = \mathbf{x}, Y = y).$$

- **Loss** l to quantify by $\boldsymbol{l}(y, f(\mathbf{x}))$ the accuracy of a guess $f(\mathbf{x})$ for y, $e.g.$

$$\mathcal{S} = \{0, 1\} : \; l(a, b) = \delta_{a \neq b}, \quad \mathcal{S} = \mathbb{R} : \; l(a, b) = \|a - b\|^2$$

- **Risk**$(\boldsymbol{l}, \boldsymbol{p})$: average loss for a given function

$$\boldsymbol{R(f)} = \mathbb{E}_{\boldsymbol{p}}[\boldsymbol{l}(Y, f(X))] = \int_{\mathbb{R}^d \times \mathcal{S}} \boldsymbol{l}(y, f(\mathbf{x}))\, \boldsymbol{p}(\mathbf{x}, \boldsymbol{y}) d\mathbf{x} dy$$

Previous Lecture: Bayes Risk, Bayes Classifier/Estimator

- Bayes **Risk**: average loss for a given function

$$R^* = \inf_{\boldsymbol{f} \in (\mathbb{R}^d)^{\mathcal{S}}} R(\boldsymbol{f}) = \inf_{\boldsymbol{f} \in (\mathbb{R}^d)^{\mathcal{S}}} \mathbb{E}_p[l(Y, \boldsymbol{f}(X))]$$

- Bayes Classifier (when $\mathcal{S} = \{0, 1\}$)

$$f_B(\mathbf{x}) = \begin{cases} 1, & \text{if } p(Y = 1 | X = \mathbf{x}) \geq \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

- Bayes Estimator (when $\mathcal{S} = \mathbb{R}$)

$$f_B(\mathbf{x}) = \mathbb{E}[Y | X = \mathbf{x}] = \int_{\mathbb{R}} y\, p(Y = y, X = \mathbf{x}) dy$$

The **Bayes** classifier (estimator) achieves the **Bayes Risk** for classification with 1/0 loss (regression with squared error) $R(f_B) = R^*$

Previous Lecture: Empirical Risk

- In practice, no access to $\boldsymbol{P}$. The only thing we can use is a training set,

$$\{(\mathbf{x}_j, y_j)\}_{i=1,\cdots,n}.$$

- Assuming the sampling is i.i.d, a counterpart to the Risk is

$$\boldsymbol{R_n^{\mathrm{emp}}}(f) = \frac{\boldsymbol{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{l}(\boldsymbol{y_i}, f(\mathbf{x}_{\boldsymbol{i}}))\text{... compare with } \boldsymbol{R(f)} = \mathbb{E}_{\boldsymbol{p}}[\boldsymbol{l}(Y, f(X))]$$

- What is overfitting?
  - Choose $f^\star$, the best function in a class of functions $\mathcal{F}$ w.r.t R_n^{emp},

$$\boldsymbol{R_n^{\mathrm{emp}}}(f^\star) = \min_{f\in\mathcal{F}} \boldsymbol{R}(f),$$

  - find out that, unfortunately, $\boldsymbol{R_n^{\mathrm{emp}}}(f^\star) \ll \boldsymbol{R}(f^\star)$.

$$\boxed{\text{overfitting: rely too much on } R_n^{\mathrm{emp}} \text{ to look for a function with low } R.}$$

Previous Lecture: Empirical Risk

- remedies for overfitting:
  - **Restrict the set of candidate functions**

$$\min_{g\in\boldsymbol{\mathcal{G}}} R_n^{\text{emp}}(g).$$

  - **Penalize "undesirable" functions**

$$\min_{g\in\boldsymbol{\mathcal{G}}} R_n^{\text{emp}}(g) + \lambda\|\boldsymbol{g}\|^{\mathbf{2}}$$

  - **Penalize properly sets of functions $\mathcal{G}_d$ of increasing complexity**

$$\min_{d\in\mathbb{N},g\in\boldsymbol{\mathcal{G}_d}} R_n^{\text{emp}}(g) + \lambda\mathsf{pen}(d,\mathcal{G}_d)$$

Previous Lecture: Excess Risk

- For any candidate set of functions $\mathcal{G}$,
- We introduce $g^\star$ as a function achieving the lowest risk in $\mathcal{G}$,

$$R(g^\star) = \inf_{g \in \mathcal{F}} R(g),$$

we decompose

$$R(g_n) - R(f_B) = \underbrace{[\boldsymbol{R(g_n) - R(g^\star)}]}_{\textbf{Estimation Error}} + \underbrace{[\boldsymbol{R(g^\star) - R(f_B)}]}_{\textbf{Approximation Error}}$$

Bounds

Alleviating Notations

- More convenient to see a couple $(\mathbf{x}, y)$ as a realization of Z, namely

$$\mathbf{z}_i = (\mathbf{x}_i, y_i), Z = (X, Y).$$

- We define the *loss class*

$$\mathcal{F} = \{f : \mathbf{z} = (\mathbf{x}, y) \to \delta_{g(\mathbf{x}) \neq y},\ g \in \mathcal{G}\},$$

- with the additional notations

$$Pf = \mathbb{E}[f(X, Y)], P_n f = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}_i, y_i),$$

where we recover

$$P_n f = R_n^{\text{emp}}(g), \quad Pf = R(g)$$

Empirical Processes

For each $f \in \mathcal{F}$, $P_n f$ is a **random variable**
which depends on n **random** realizations of $Z = (X, Y)$.

- If we use P_n on **all** possible functions $f \in \mathcal{F}$, we obtain

The set of random variables $\{P_n f\}_{f \in \mathcal{F}}$ is called an
Empirical measure indexed by $\mathcal{F}$.

- A branch of mathematics studies explicitly the convergence of $\{Pf - P_n f\}_{f \in \mathcal{F}}$,

This branch is known as Empirical process theory.

Hoeffding's Inequality

- Recall that for a given g and corresponding f,

$$R(g) - R^{\mathrm{emp}}(g) = Pf - P_n f = \mathbb{E}[f(Z)] - \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i),$$

  which is simply the difference between the **expectation** and the empirical average of $f(Z)$.

- The **strong** law of large numbers says that

$$P\left(\lim_{n\to\infty}\left(\mathbb{E}[f(Z)] - \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i)\right) = 0\right) = 1.$$

Hoeffding's Inequality

- A more precise result is the

  Theorem 1 (Hoeffding). *Let* $Z_1, \cdots, Z_n$ *be* n *i.i.d random variables with* $f(Z) \in [a, b]$. *Then,* $\forall \varepsilon > 0$,

$$P\left(|P_n f - Pf| > \varepsilon\right) \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}.$$

- From

$$P\left(\lim_{n\to\infty}\left(\mathbb{E}[f(Z)] - \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i)\right) = 0\right) = 1.$$

  we get

$$P\left(\left|\mathbb{E}[f(Z)] - \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i)\right| > \varepsilon\right) \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}.$$

- Hoeffding's inequality is a **concentration inequality**.

Hoeffding's Inequality

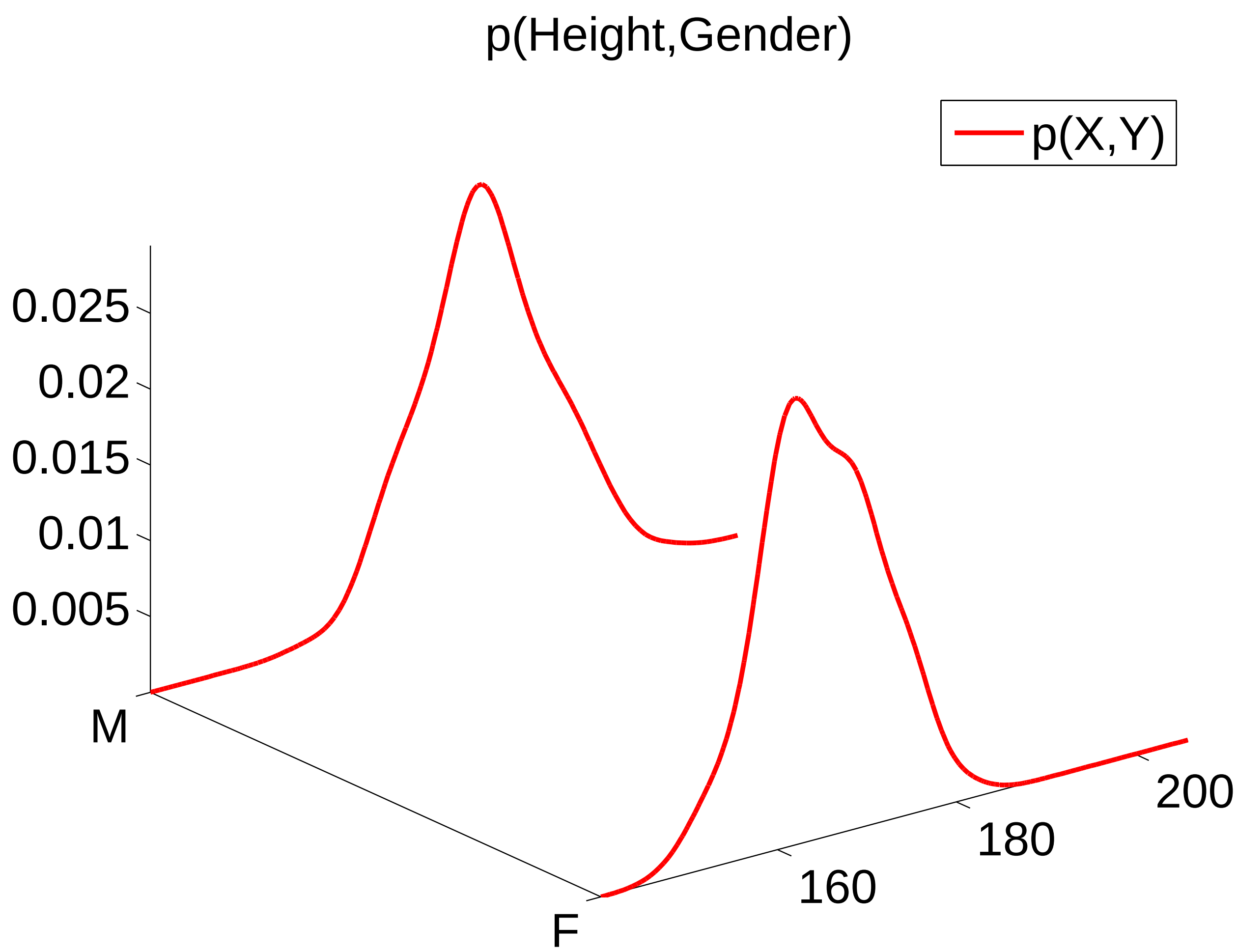
- We will need to prove it using another inequality,

  Theorem 2 (Markov). *Let* $X \geq 0$ *be a non-negative random variable in* $\mathbb{R}$, *then*

  $$P(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

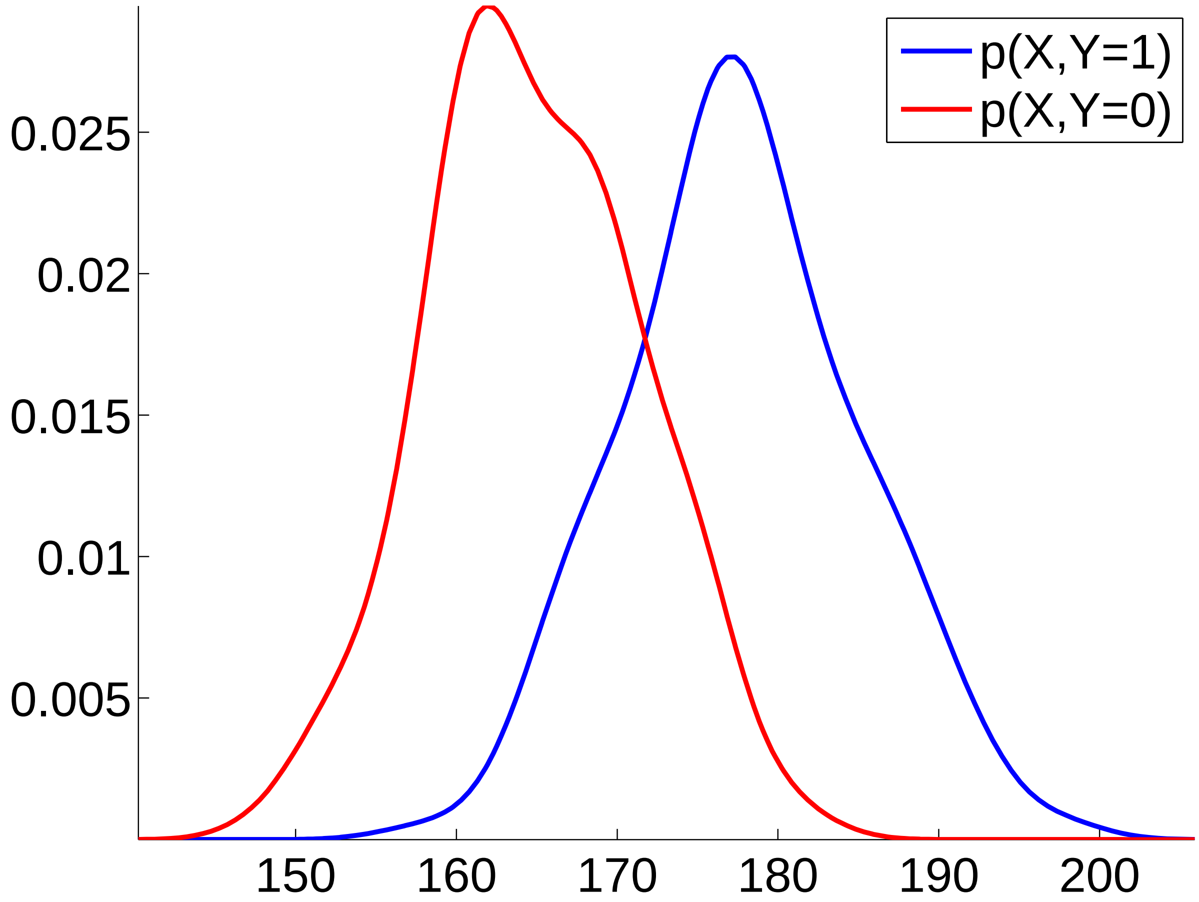
- Before getting to the proof, let's check the intuition behind it.

Height/Gender



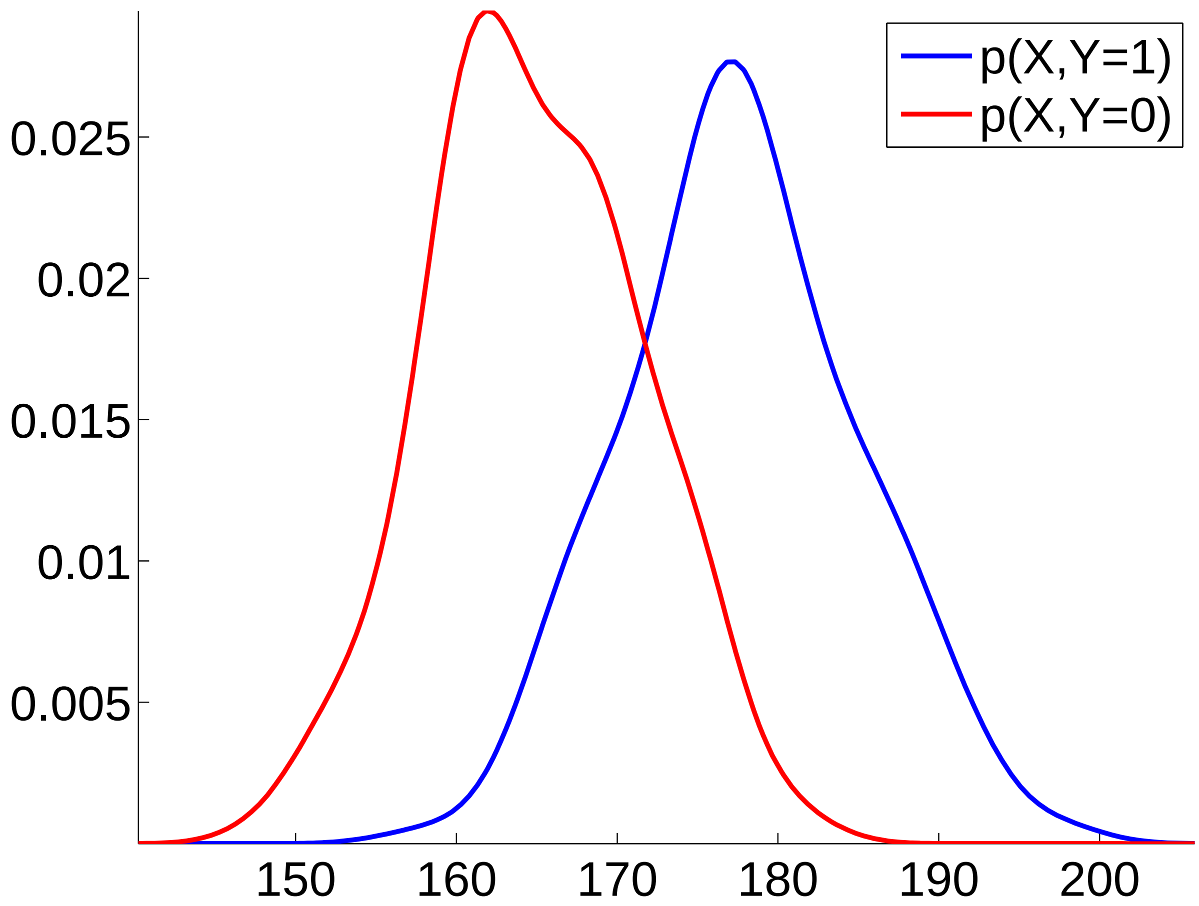
In 3 dimensions

Height/Gender



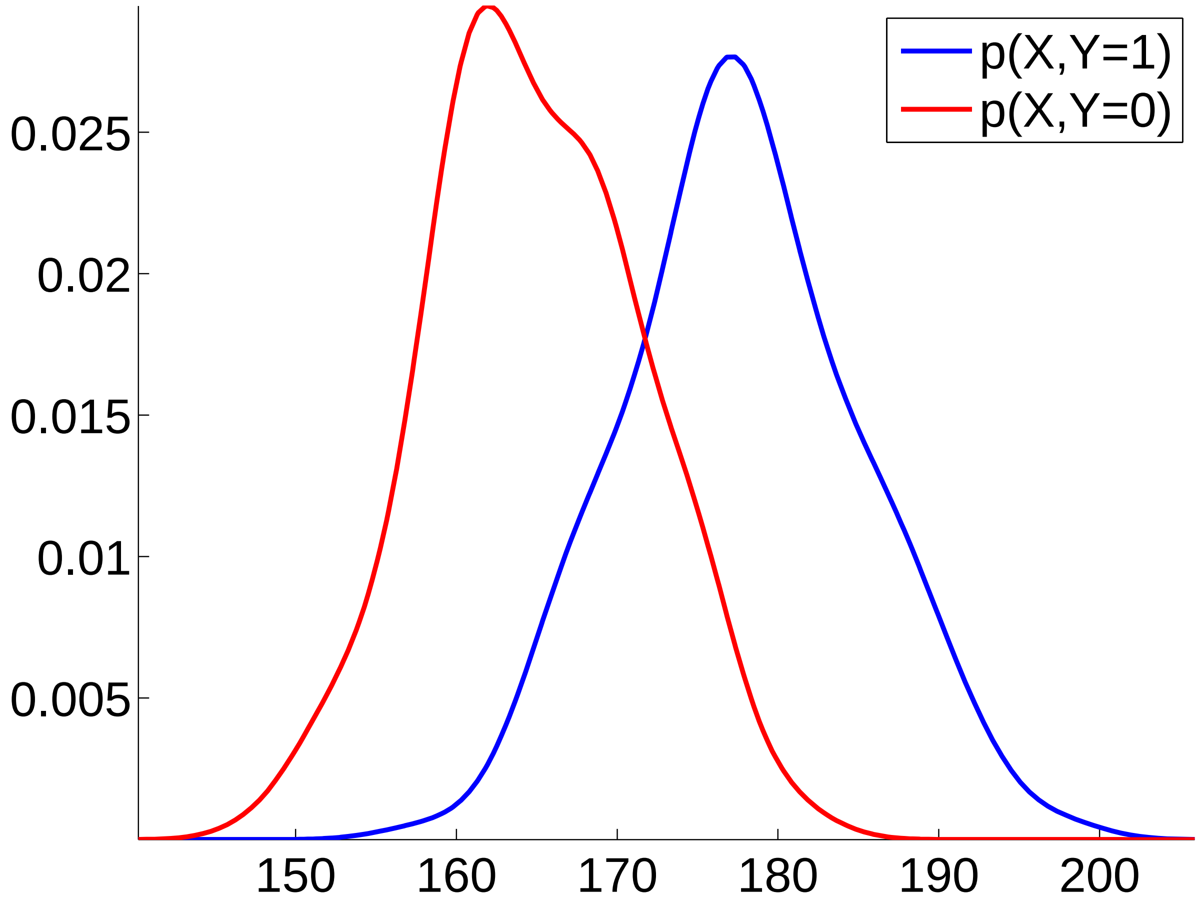
Easier to see in 2 dimensions, same content.

Height/Gender



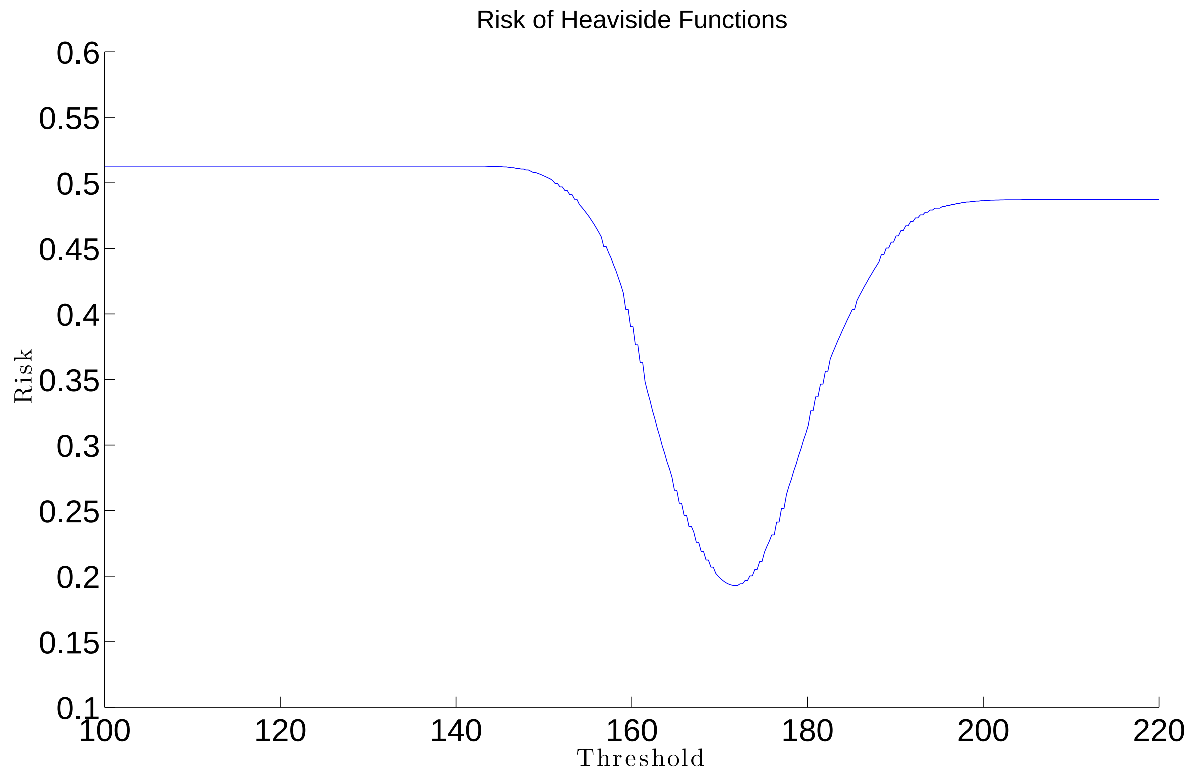
Assume for a minute that we **known** these two curves.

Height/Gender



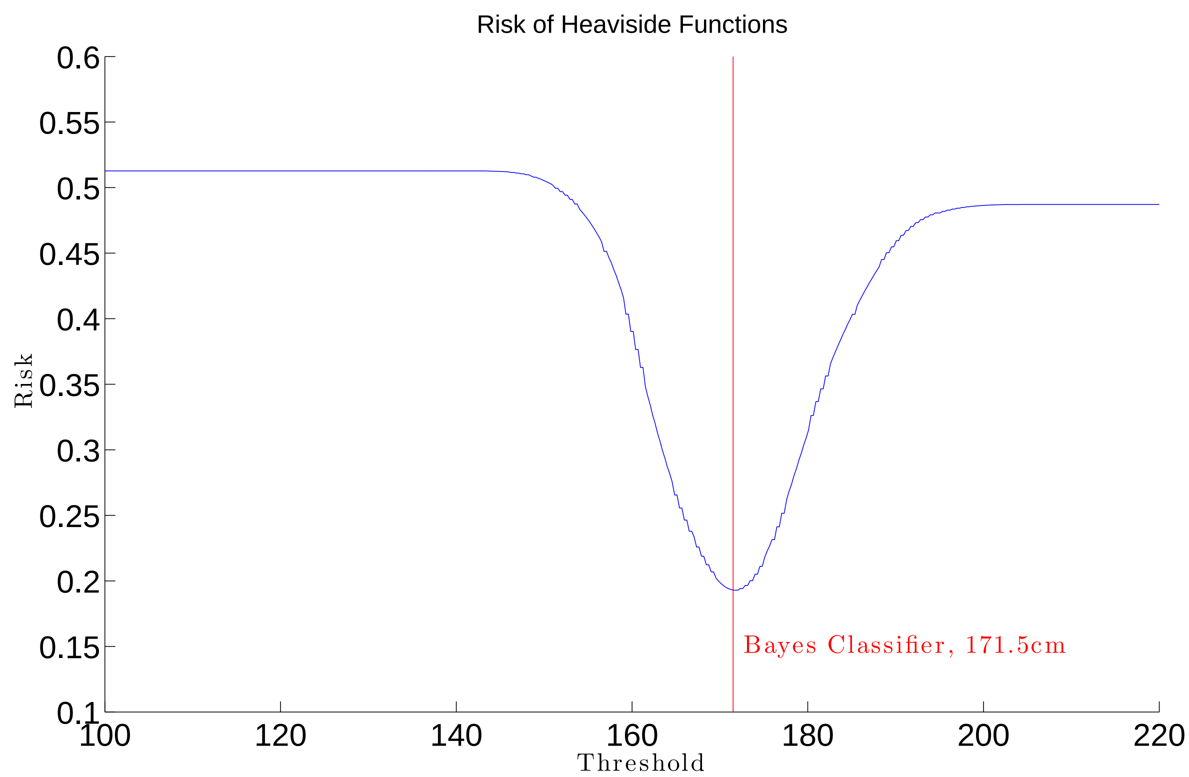
For any function f : Height $\mapsto$ Gender we can compute the risk

Height/Gender



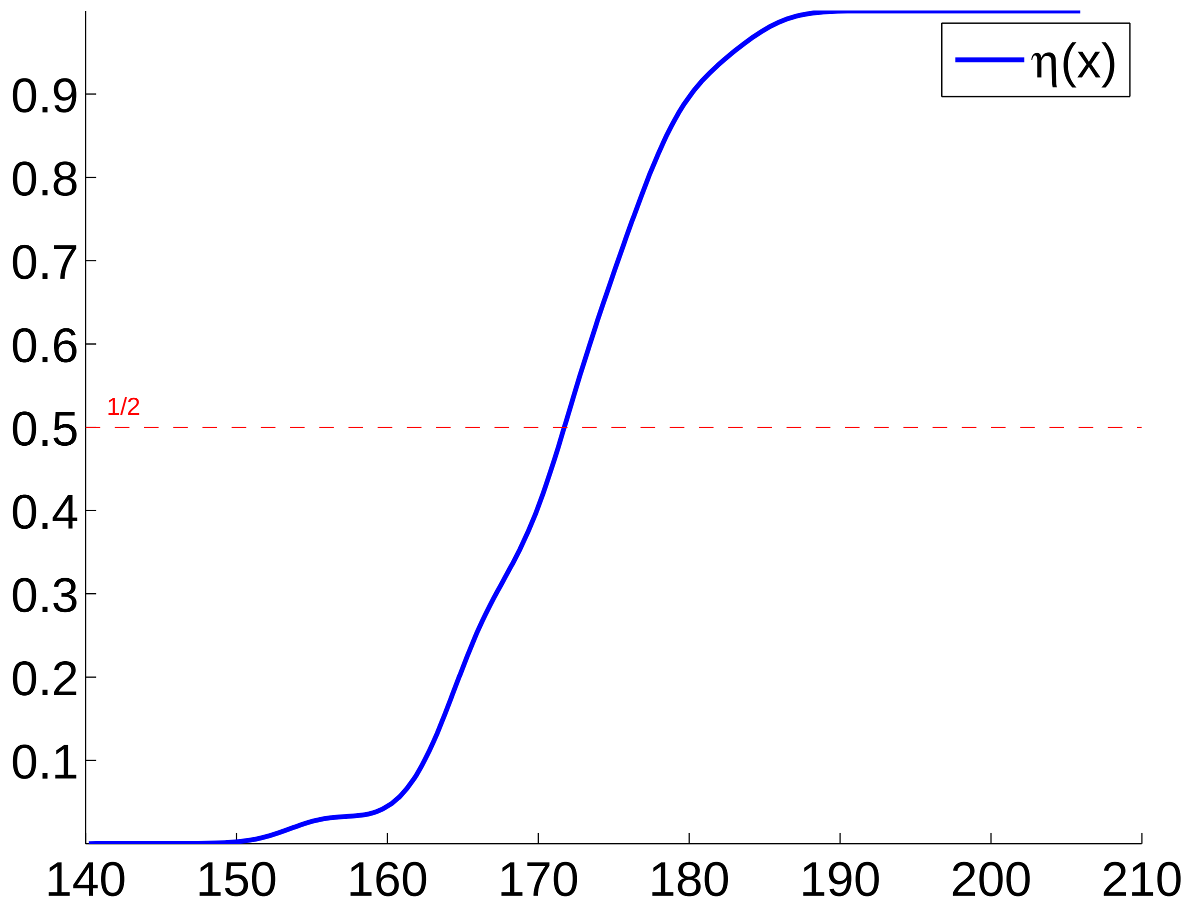
Risk for Heaviside functions $f(x) = \delta_{x>\tau}$

Height/Gender



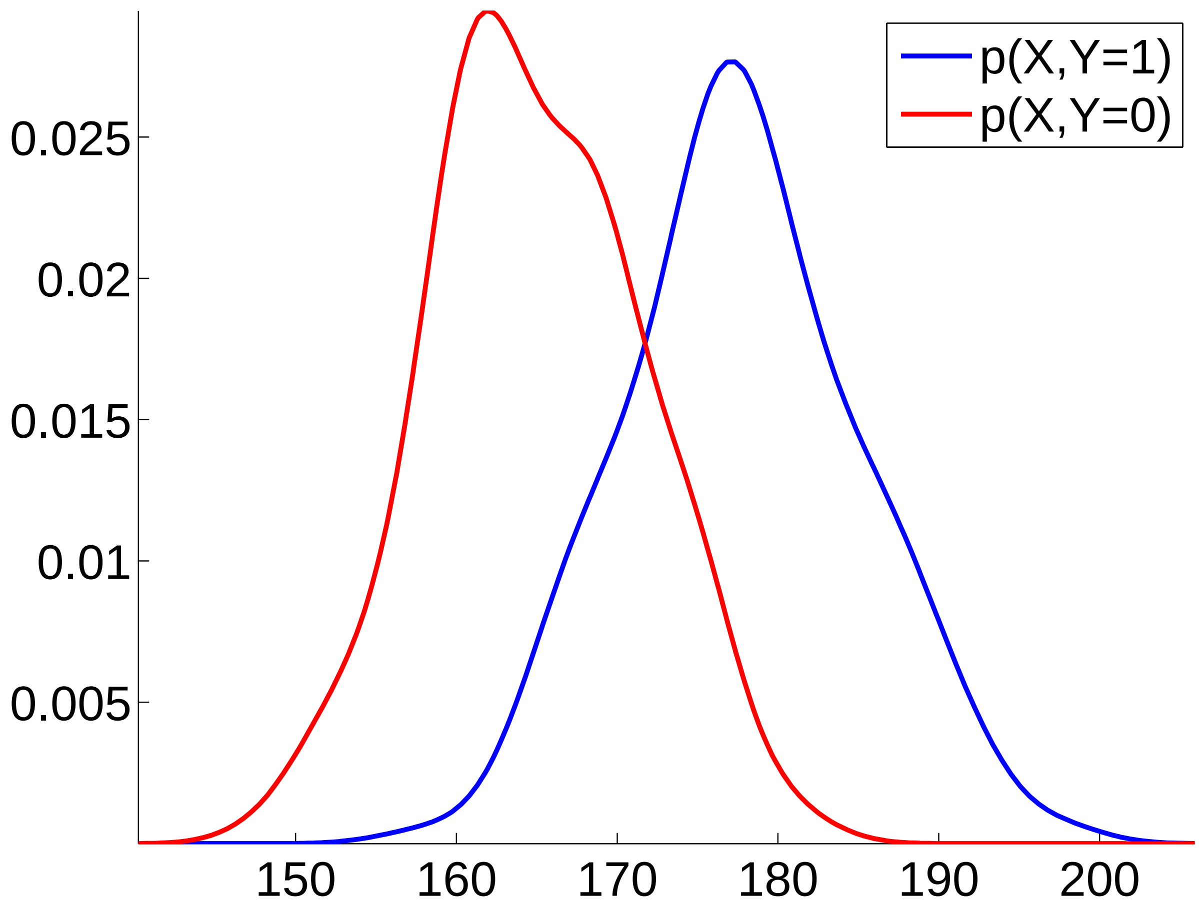
The risk is minimal for the thresholded function with $\tau \approx 171.5$

Height/Gender



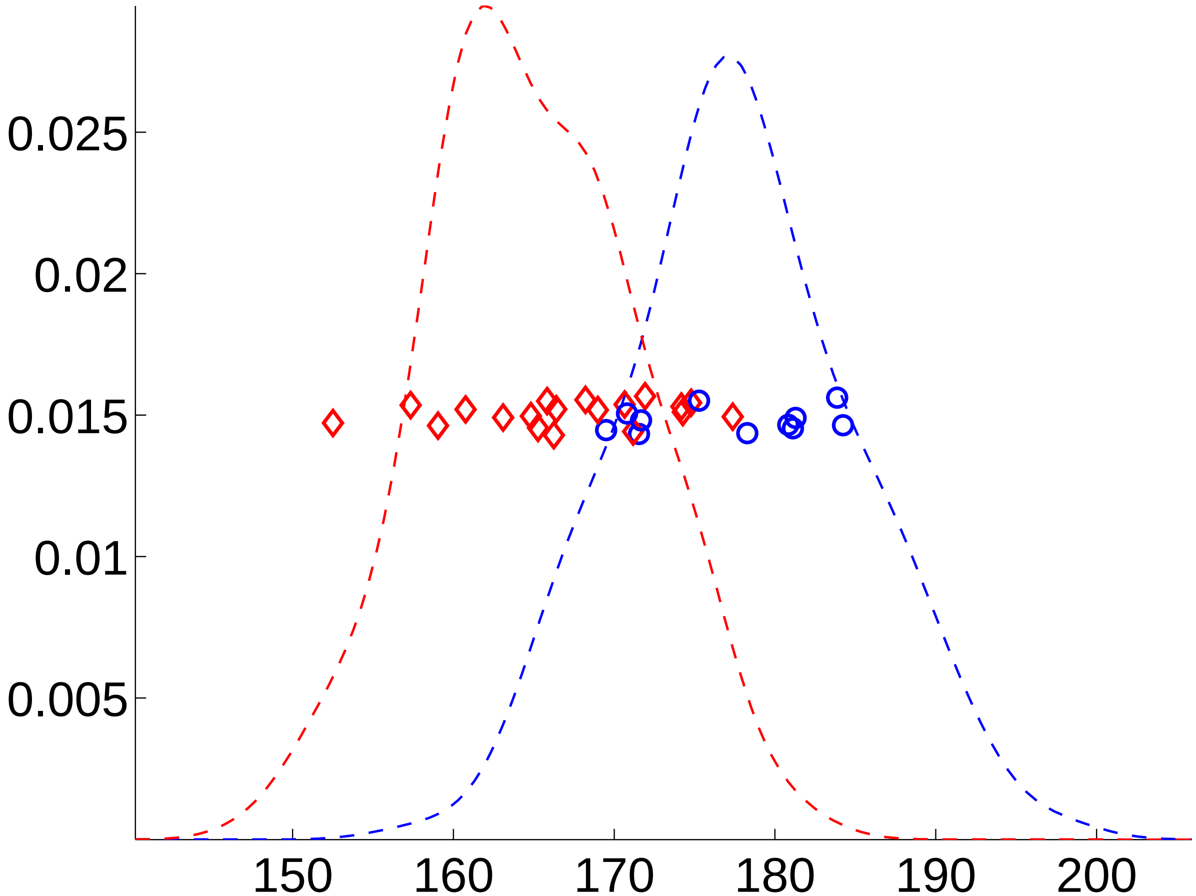
which matches our picture of the Bayes classifier and the $\eta(x) = P(Y = 1|X = \mathbf{x})$ function.

Height/Gender



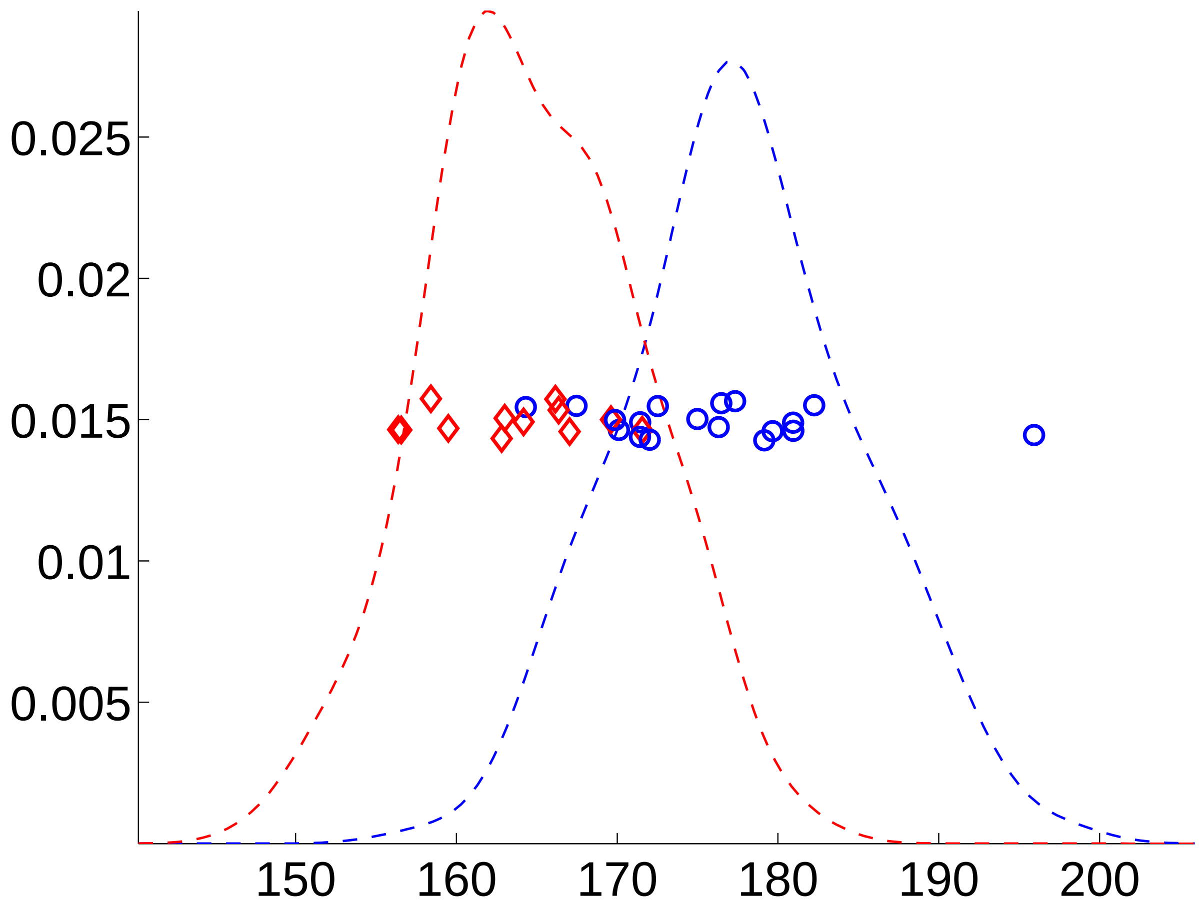
Unfortunately, we do not have access to this,

Height/Gender



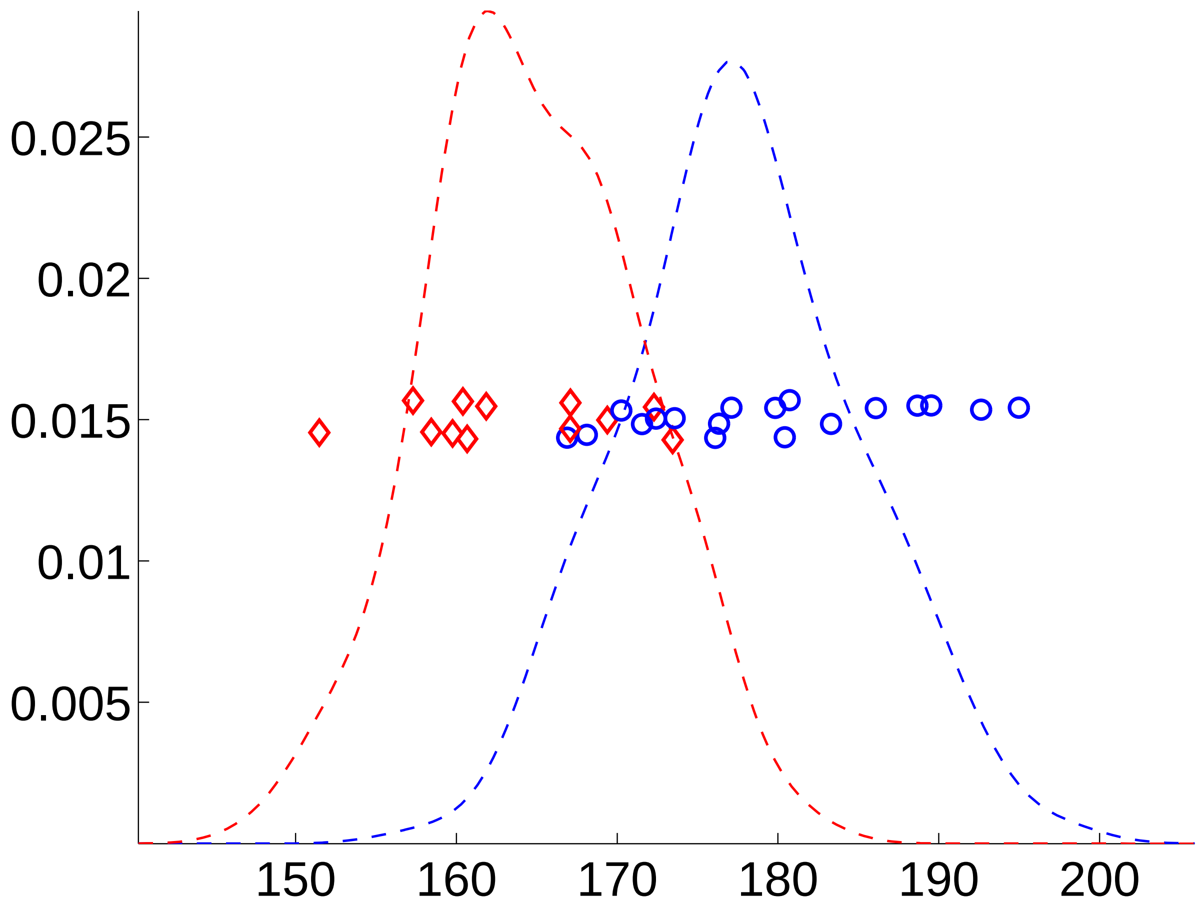
But rather this...

Height/Gender



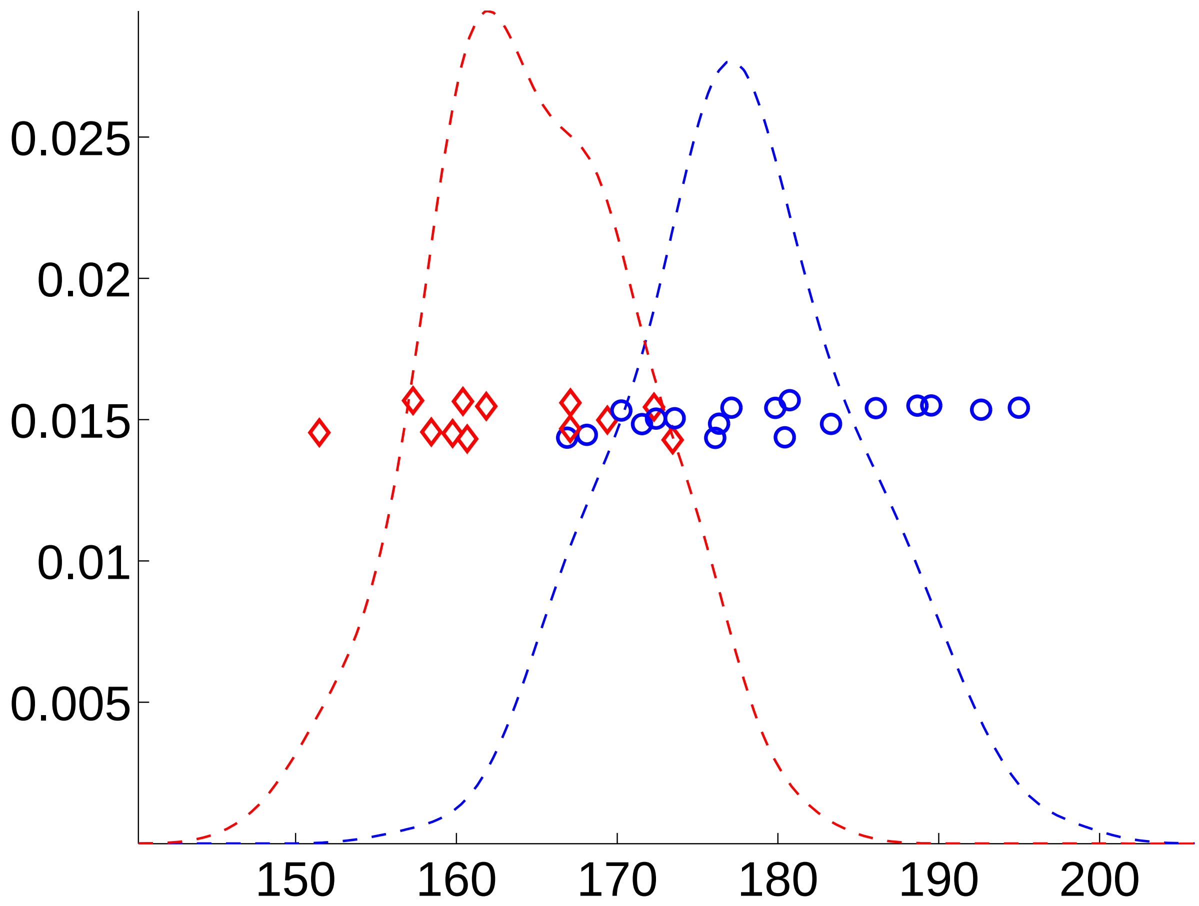
or this...

Height/Gender



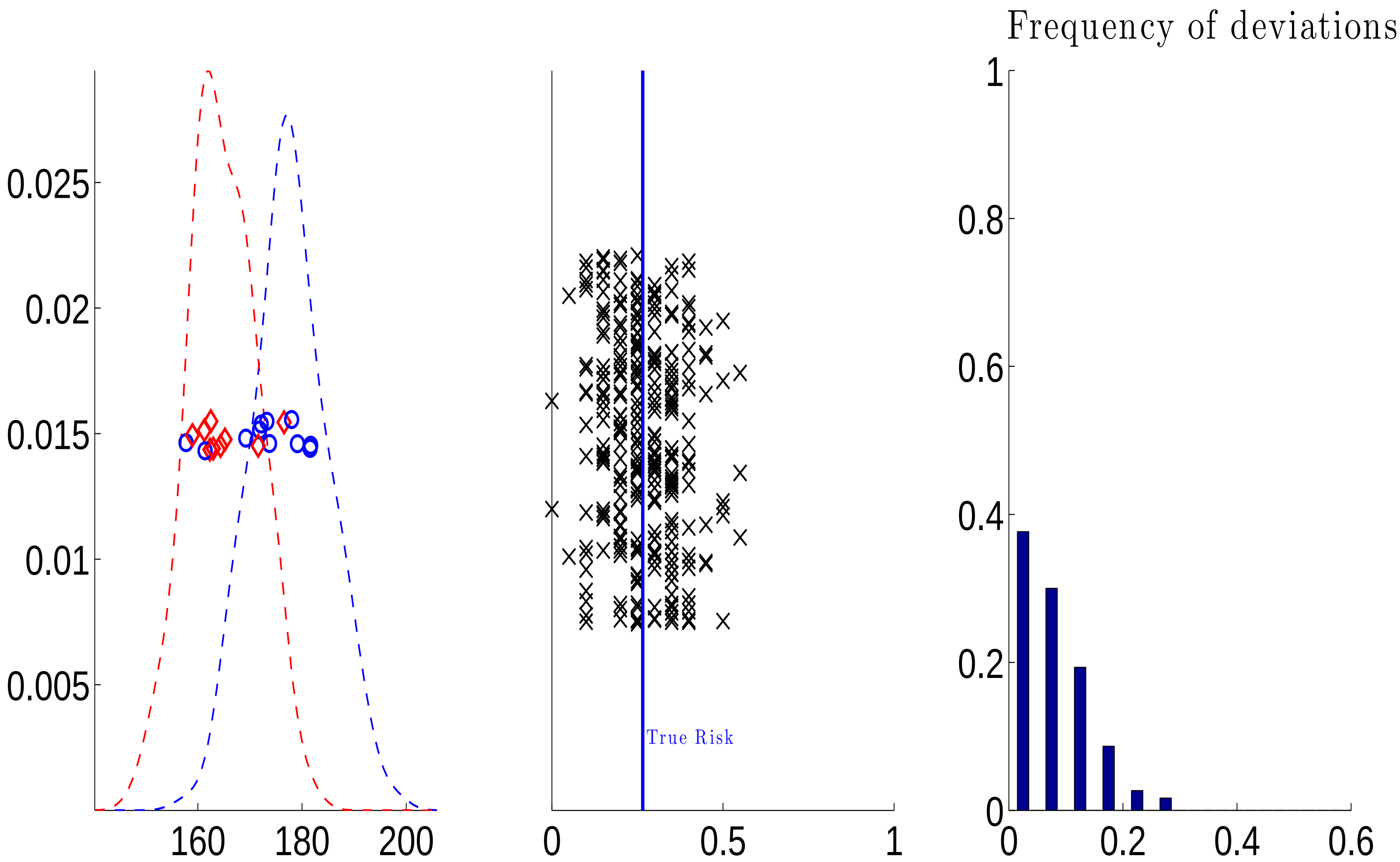
or even this... we assume our samples are **random**.

Height/Gender



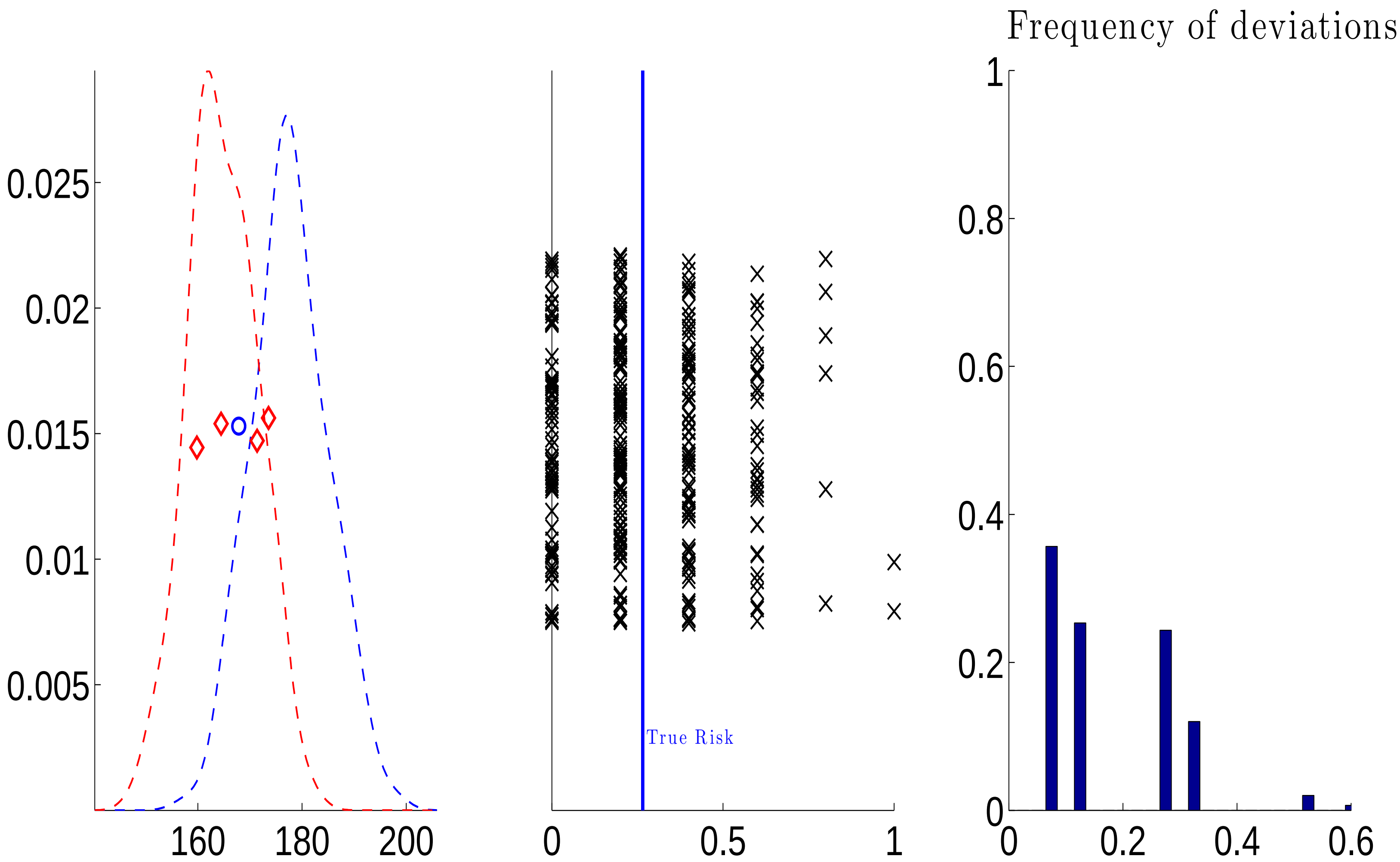
Hoeffding's Inequality: $P\left(|P_n f - Pf| > \varepsilon\right) \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}$.

Hoeffding's Inequality



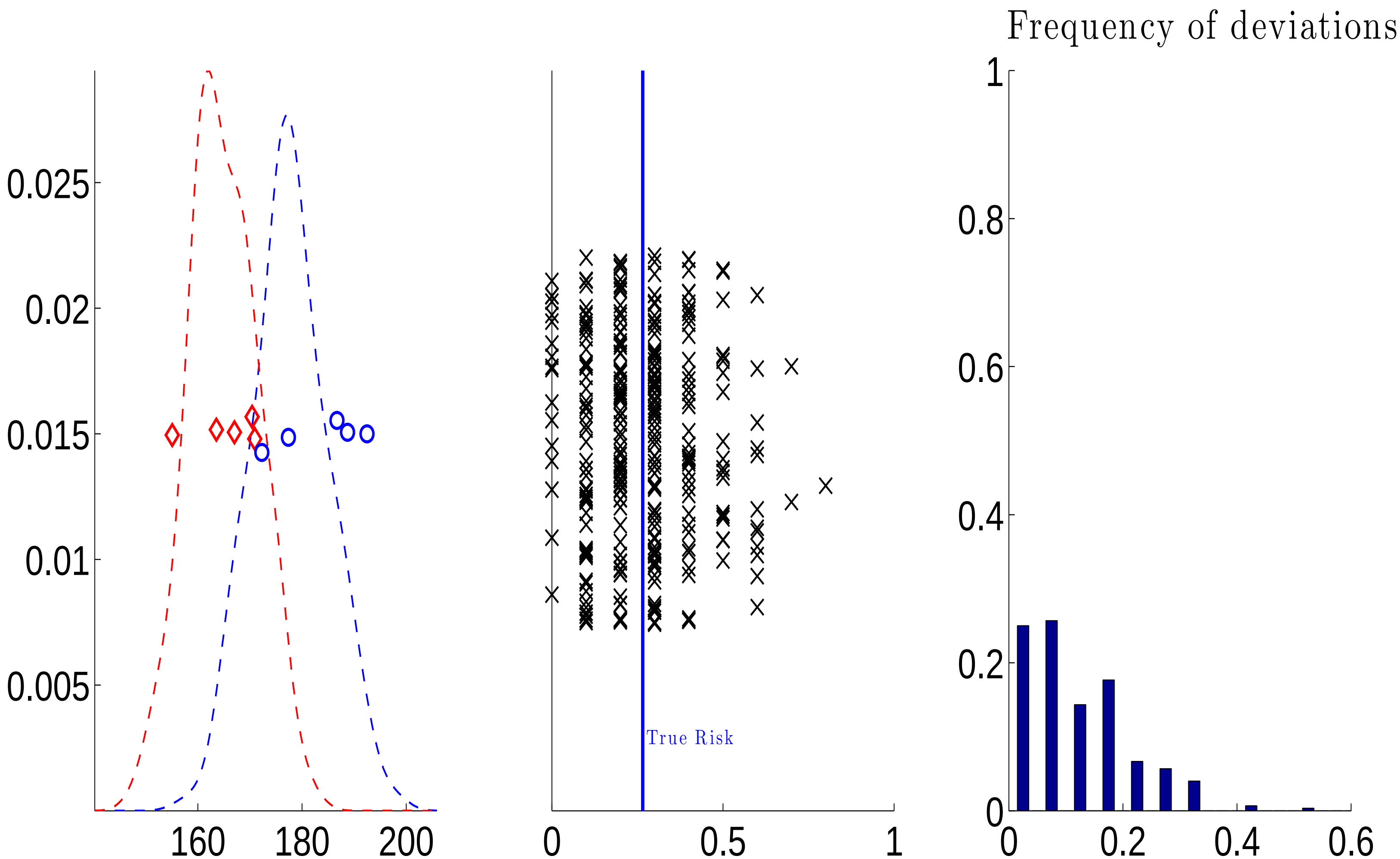
Let's check on Matlab what this means

Hoeffding's Inequality



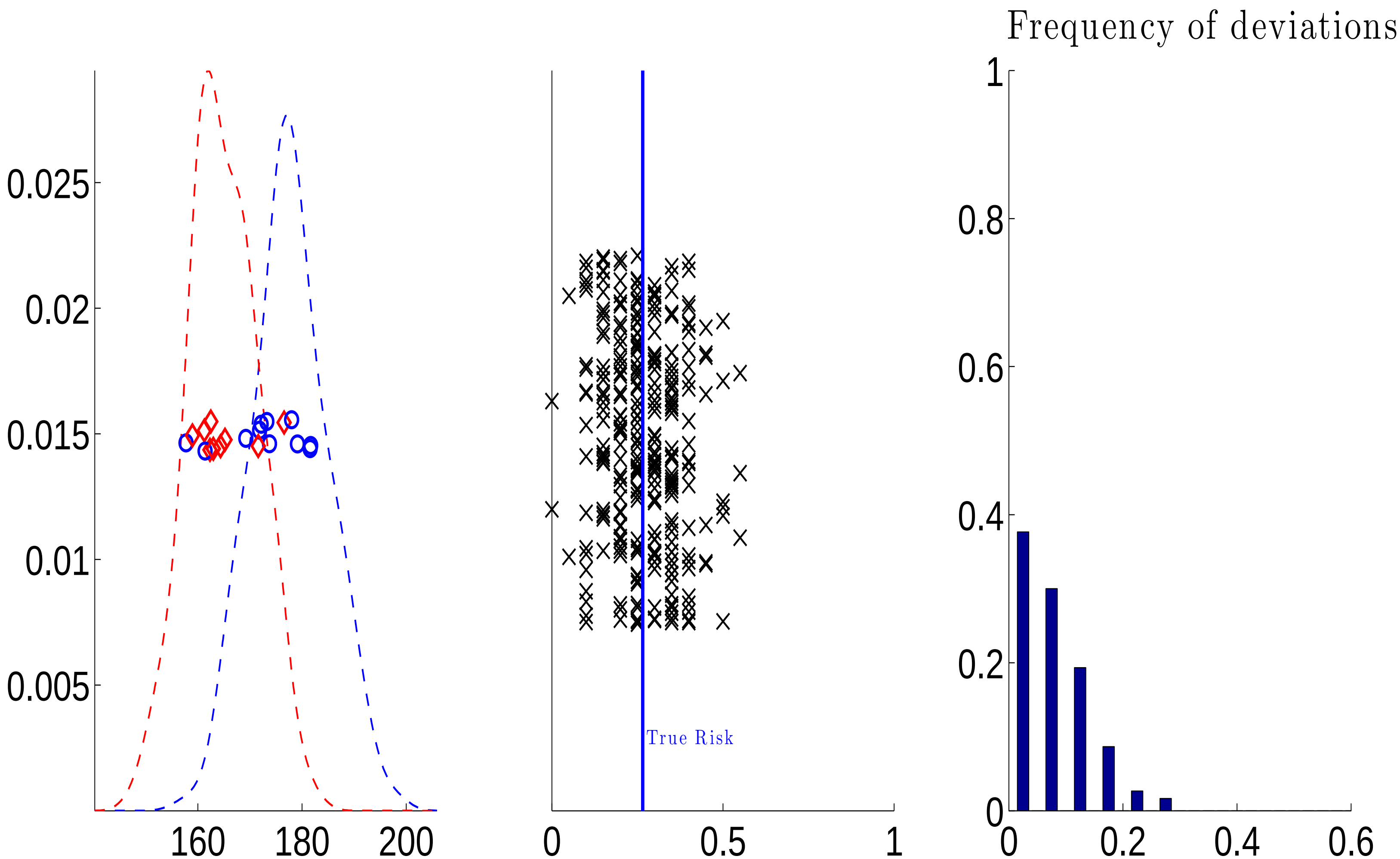
with $n = 5$ resampled 300 times

Hoeffding's Inequality



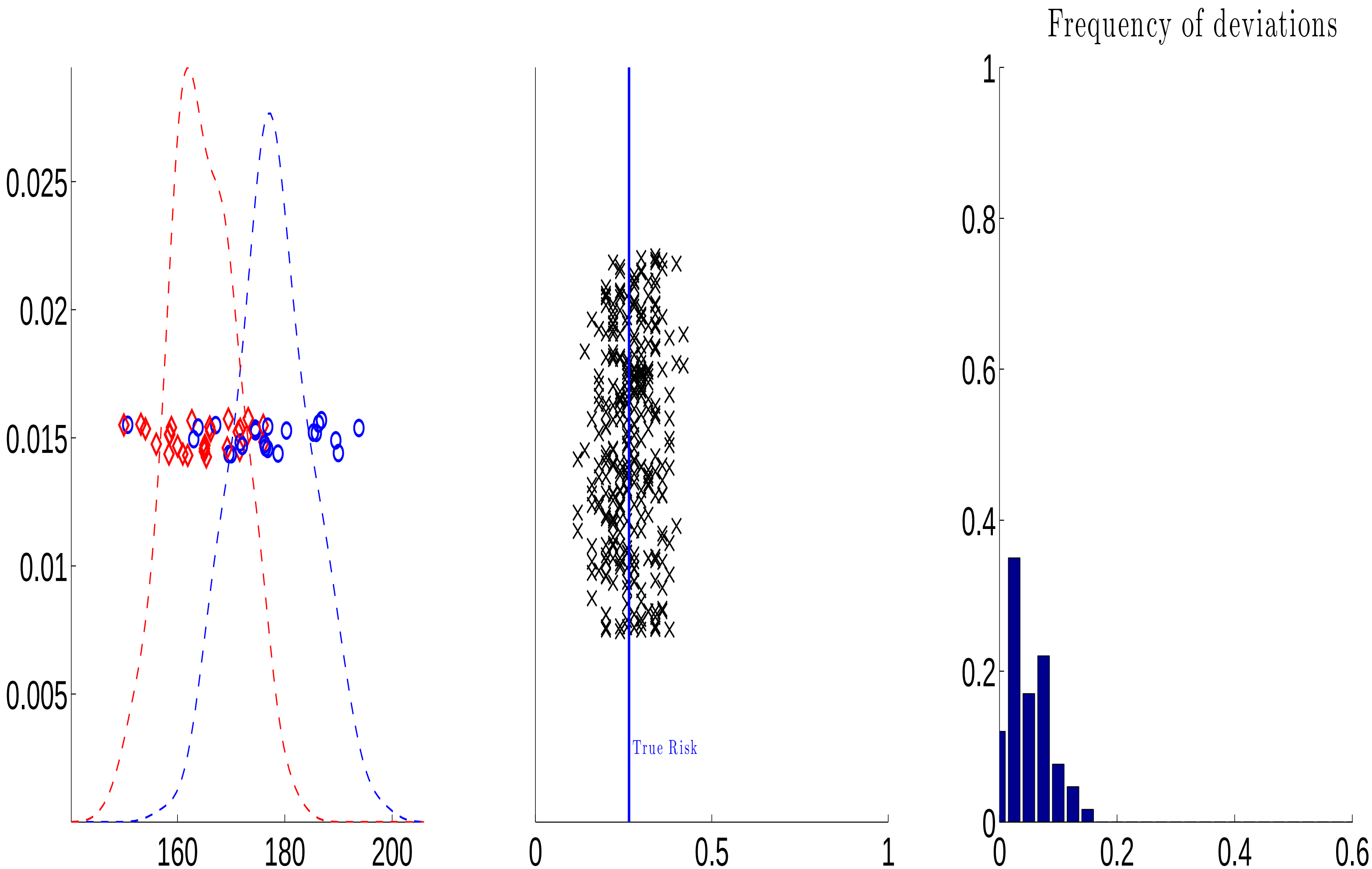
with $n = 10$ resampled 300 times

Hoeffding's Inequality



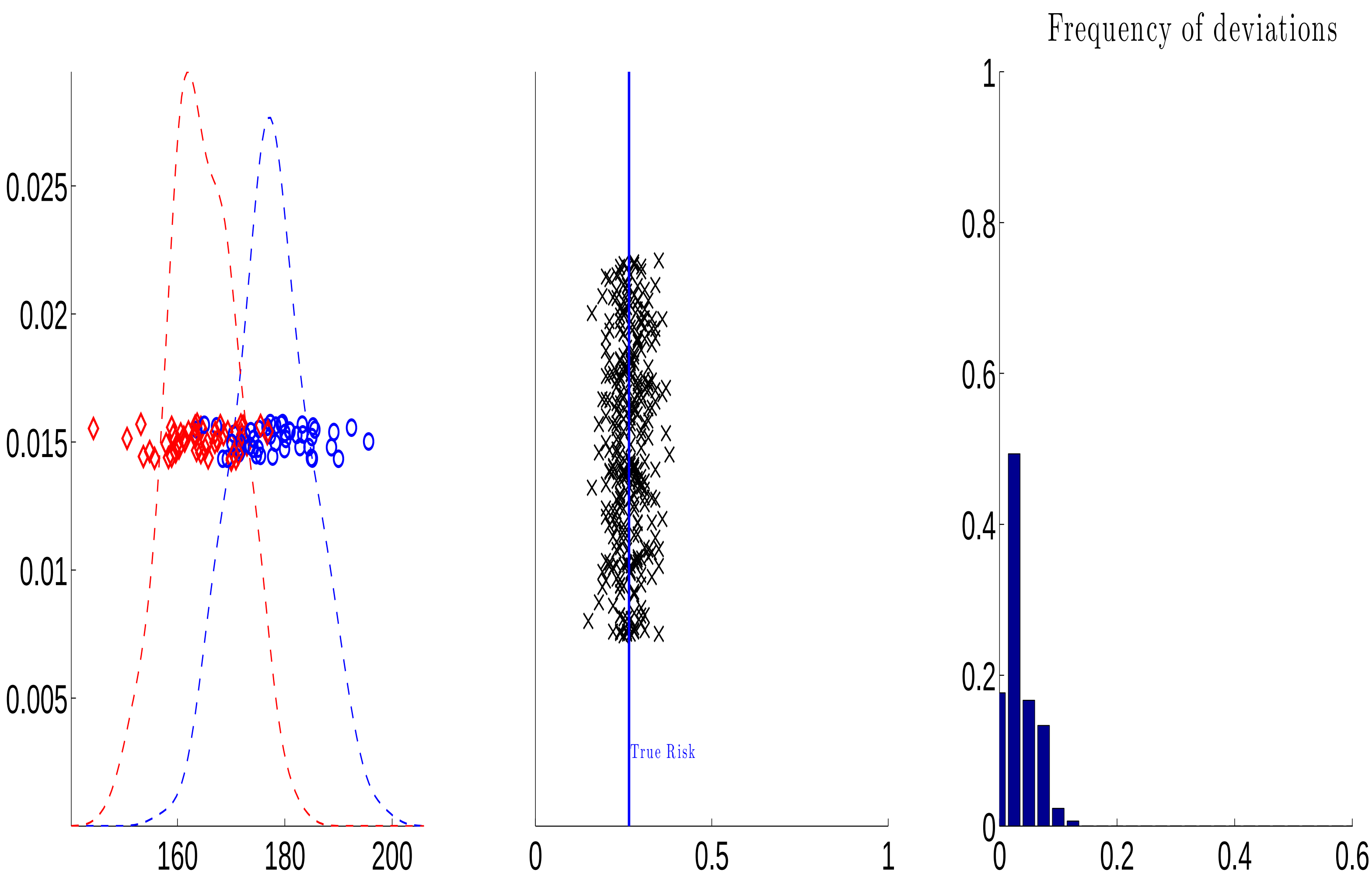
with $n = 20$ resampled 300 times

Hoeffding's Inequality



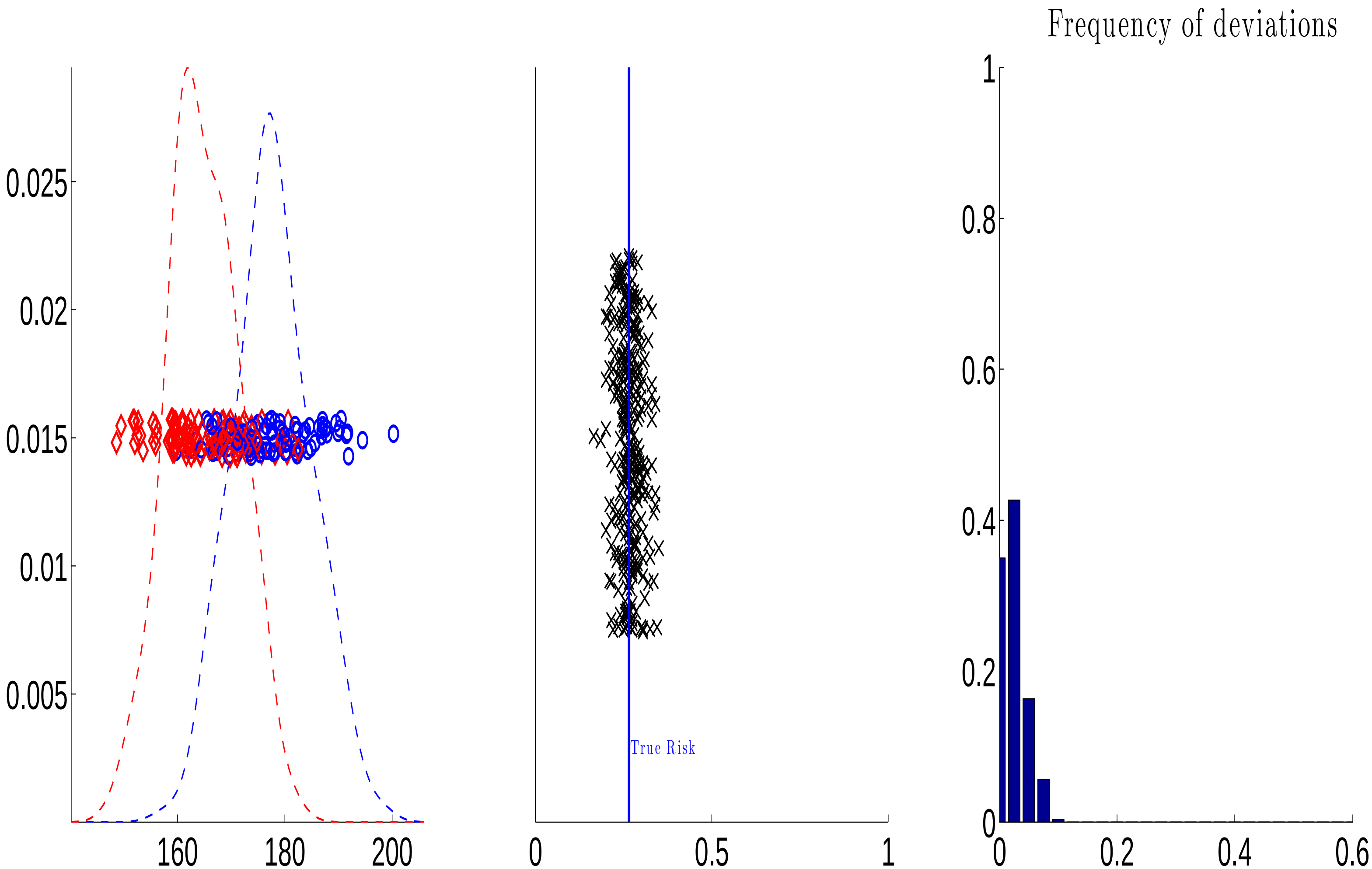
with $n = 50$ resampled 300 times

Hoeffding's Inequality



with $n = 100$ resampled 300 times

Hoeffding's Inequality



with $n = 200$ resampled 300 times

Some Proofs

Theorem 3 (Hoeffding). *Let $Z_1, \cdots, Z_n$ be n i.i.d random variables with $f(Z) \in [a, b]$. Then, $\forall \varepsilon > 0$,*

$$P\left(|P_n f - Pf| > \varepsilon\right) \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}.$$

Theorem 4 (Markov). *Let $X \geq 0$ be a non-negative random variable in $\mathbb{R}$, then*

$$P\left(X \geq t\right) \leq \frac{\mathbb{E}[X]}{t}.$$

Inverting Hoeffding's Inequality

- Naturally, if

$$P\left(|P_n f - Pf| > \varepsilon\right) \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}.$$

- then for $\delta > 0$,

$$P\left(|P_n f - Pf| > (b-a)\sqrt{\frac{\log \frac{2}{\delta}}{2n}}\right) \leq \delta.$$

- which is also interpreted as, with probability at least $1 - \delta$,

$$|P_n f - Pf| \leq (b-a)\sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

Interpretation in terms of Risk

- Functions f take values between $a = 0$ and $b = 1$. $b - a = 1$ for all inequalities.
- For any function g, and any δ, with probability at least $1 - \delta$,

$$R(g) \leq R_n^{\text{emp}}(g) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

- Note that the ***probability at least*** statement refers to **samples of size** n.
- This result, seemingly nice, is not very useful... why?
  - Get data first, estimate g_n... gap between $R(g_n)$ and $R_n(g_n)$?
  - Define $\hat{g}$ as $\hat{g}(\mathbf{x}_i) = y_i$ and $\hat{g} = 0$ everywhere else. Of course, $R(\hat{g}) \gg R_n^{\text{emp}}(\hat{g}) \stackrel{\text{def}}{=} 0$.

Interpretation in terms of Risk

- This is why we focus now on **uniform** deviations on the function class,

$$\sup_{f \in \mathcal{F}} Pf - P_n f,$$

since we know that whatever the function g_n we choose with the sample,

$$R(g) - R_n(g_n) \leq \sup_{g \in \mathcal{G}} R(g) - R_n(g) = \sup_{f \in \mathcal{F}} Pf - P_n f,$$

Obtaining Uniform Bounds

- Simple example with two functions f_1 and f_2.
- Define the two sets of n-uples,

$$C_{\mathbf{1}} = \{\{(\mathsf{x}_1, y_1), \cdots, (\mathsf{x}_n, y_n)\} \mid Pf_{\mathbf{1}} - P_n f_{\mathbf{1}} > \varepsilon\}$$

  and

$$C_{\mathbf{2}} = \{\{(\mathsf{x}_1, y_1), \cdots, (\mathsf{x}_n, y_n)\} \mid Pf_{\mathbf{2}} - P_n f_{\mathbf{2}} > \varepsilon\}$$

- These sets are the "bad" sets for which empirical risk is much lower than the real risk.

Obtaining Uniform Bounds

- For each, we have following Hoeffing's inequality **(no absolute value)**, that

$$P(C_1) \leq \delta, P(C_2) \leq \delta \text{ where } \delta = e^{-2n\varepsilon^2}.$$

- Note that whenever a n-uple is in $C_1 \cup C_2$, then either

$$Pf_{\mathbf{1}} - P_n f_{\mathbf{1}} > \varepsilon \text{ or } Pf_{\mathbf{1}} - P_n f_{\mathbf{1}} > \varepsilon.$$

- Of course, $P(C_1 \cup C_2) \leq P(C_1) + P(C_2) \leq 2\delta$.
- Thus, with probability smaller than 2δ at least one of f_1 or f_2 will be such that $Pf_{\mathbf{1}} - P_n f_{\mathbf{1}} > \varepsilon$.

Generalizing to N functions

- Consider $f_1, \cdots, f_N$ functions.
- Define the corresponding sets of n-uples, $C_1, \cdots, C_N$ with ε fixed.
- Of course,
$$P(C_1 \cup C_2 \cup \cdots \cup C_N) \leq \sum_{i=1}^{N} P(C_i)$$
- Use now Hoeffding's inequality

$$\begin{aligned} P(\exists f \in \{f_1, \cdots, f_N\} \,|\, Pf - P_n f > \varepsilon) &= P\left(\bigcup_{i=1}^{N} C_i\right) \\ &\leq \sum_{i=1}^{N} P(C_i) \leq N\delta = Ne^{-2n\varepsilon^2} \end{aligned}$$

Hoeffding's bound for finite families of functions

- We thus have that for **any** family of N functions,

$$P(\sup_{Pf - P_n f} \geq \varepsilon) \leq N e^{-2n\varepsilon^2},$$

- or equivalently, that if $\mathcal{G} = \{g_1, \cdots, g_N\}$, with probability at least $1 - \delta$,

$$\forall g \in \mathcal{G}, \quad R(g) \leq R_n(g) + \sqrt{\frac{\log N + \log \frac{1}{\delta}}{2n}}$$

Hoeffding's bound for countable families of functions

- Suppose now that we have a countable family $\mathcal{F}$
- Suppose that we assign a number $\delta(f) > 0$ to each $f \in \mathcal{F}$, which we use to set

$$P\left(|Pf - P_n f| > \sqrt{\frac{\log \frac{2}{\delta(f)}}{2n}}\right) \leq \delta(f),$$

- Using the union bound on a **countable set** (basic probability axiom),

$$P\left(\exists f \in \mathcal{F} : |P_n f - Pf| > \sqrt{\frac{\log \frac{2}{\delta(f)}}{2n}}\right) \leq \sum_{f \in \mathcal{F}} \delta(f).$$

- Let us set $\delta(f) = \rho p(f)$ with $\rho > 0$ and $\sum_{f \in \mathcal{F}} p(f) = 1$.
- Then with probability $1 - \delta$,

$$\forall f \in \mathcal{F}, Pf \leq P_n f + \sqrt{\frac{\log \frac{1}{p(f)} + \log \frac{1}{\rho}}{2n}}.$$

Hoeffding's bound for general families of functions

- Two problems:
  - Most interesting families of functions are not countable.
  - Defining the weights $p(f)$ is not so obvious.
- However, what really matters for a sample $\mathbf{z}_1, \cdots, \mathbf{z}_n$ is

$$\mathcal{F}_{\mathbf{z}_1,\cdots,\mathbf{z}_n} = \{(f(\mathbf{z}_1), f(\mathbf{z}_2), \cdots, f(\mathbf{z}_n)),\ f \in \mathcal{F}\}$$

- $\mathcal{F}_{\mathbf{z}_1,\cdots,\mathbf{z}_n}$ is a large set of binary vectors $\subset \{0,1\}^N$
- The more complex $\mathcal{F}$, the larger $\mathcal{F}_{\mathbf{z}_1,\cdots,\mathbf{z}_n}$ with maximum 2^n possible elements.

Definition 1 (Growth Function). *The growth function of $\mathcal{F}$ is equal to*

$$S_{\mathcal{F}}(n) = \sup_{(\mathbf{z}_1,\cdots,\mathbf{z}_n)} |\mathcal{F}_{\mathbf{z}_1,\cdots,\mathbf{z}_N}|$$

Vapnik-Chervonenkis

Theorem 5 (Vapnik-Chervonenkis). *For any $\delta > 0$, with probability at least $1-\delta$,*

$$\forall g \in \mathcal{G}, R(g) \leq R_n(g) + 2\sqrt{2\frac{\log S_{\mathcal{G}}(2n) + \log\frac{2}{\delta}}{n}}$$

Definition 2 (VC Dimension). *The VC dimension of a class $\mathcal{G}$ is the largest n such that*

$$S_{\mathcal{G}}(n) = 2^n.$$

- The VC dimension of linear classifiers in $\mathbb{R}^d$ is $d+1$.
- Given the VC dimension h of a family $\mathcal{G}$, we can prove

$$\forall g \in \mathcal{G}, R(g) \leq R_n(g) + 2\sqrt{2\frac{h\log\frac{2en}{h} + \log\frac{2}{\delta}}{n}}$$