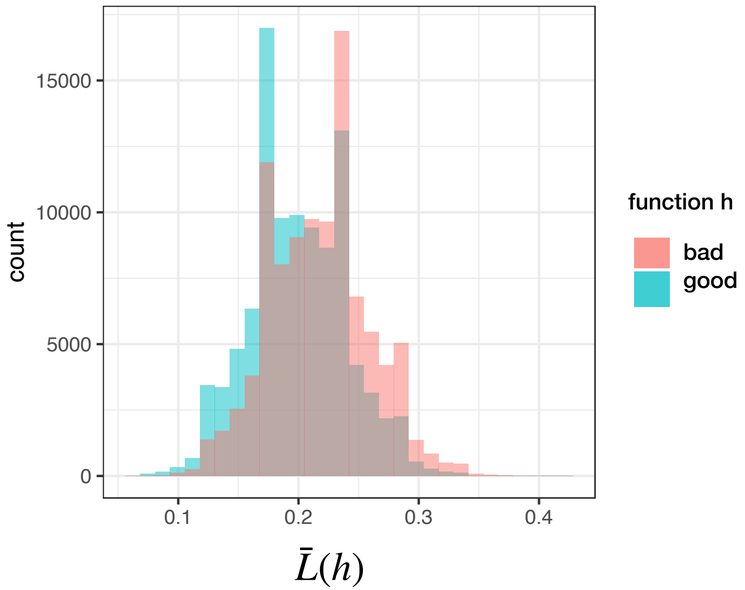
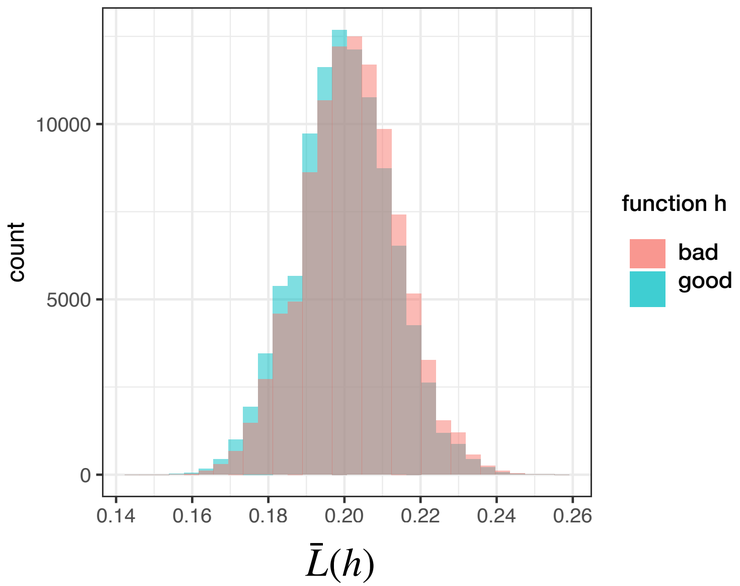
An experiment

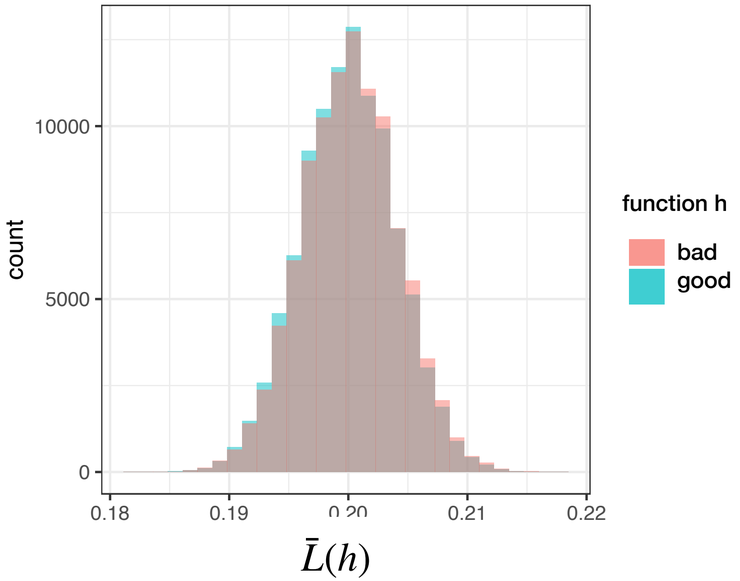
- Two models: $|\mathscr{H}| = 2$
- Good model: accuracy 80%
- Bad model: accuracy 80% - 4/n
- Does ERM find the good model with probability $90\,\% = 1 - \delta$?



$n = 100$



$n = 1000$



$n = 10000$

red and blue high overlap → ERM wrong ~half the time → need more samples!

Concentration inequalities

Setting: I.I.D. random variables $Z_i \in [0,1]$, $\mathbb{E}[Z_i] = \mu$, variance$(Z_i) = \sigma^2$

Law of large numbers: $\frac{1}{n}\sum_{i=1}^{n} Z_i \xrightarrow{\mathsf{P}} \mu$ as $n \to \infty$

Central limit theorem: $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right) \xrightarrow{\mathsf{d}} \mathcal{N}(0,1)$ as $n \to \infty$

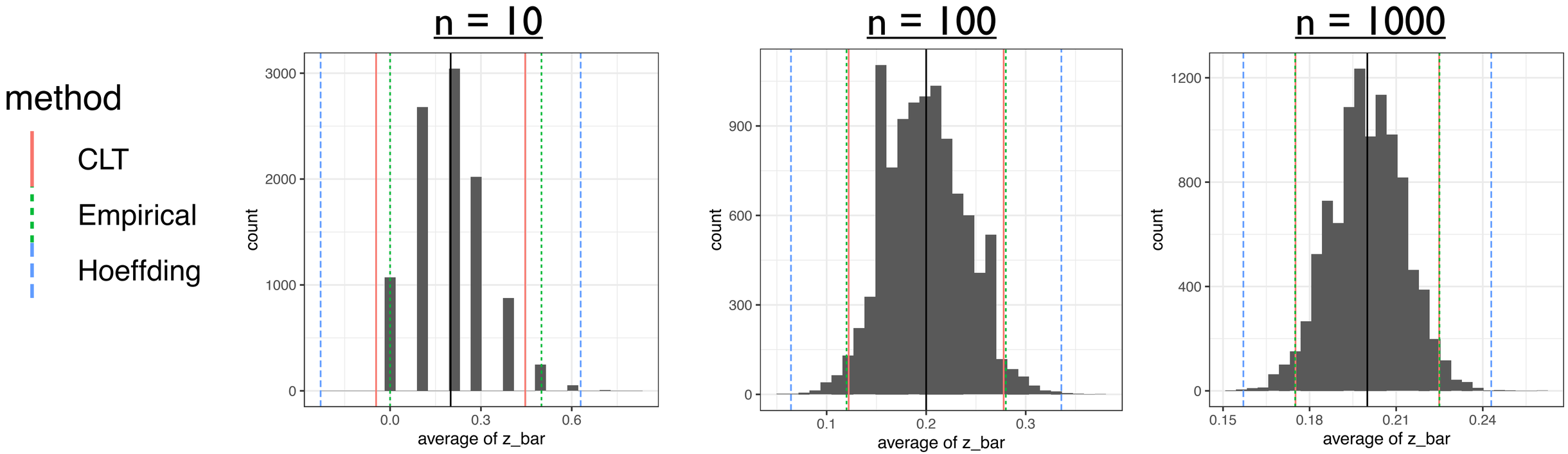
Hoeffding's inequality: $P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| \geq t\right) \leq 2e^{-2nt^2}$ for all n

a "concentration inequality"

Concentration, example

Setting: I.I.D. random variables $Z_i \sim \text{Bern}(.2)$, $\mathbb{E}[Z_i] = .2$ $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$

Look at 95% bounds on $\bar{Z}$ based on Hoeffding and CLT.



- Hoeffding is valid in all three panels
- CLT is a better approximation with large n, not exact with n = 10
- Hoeffding is not tight for large n (it does not approach the true quantiles)