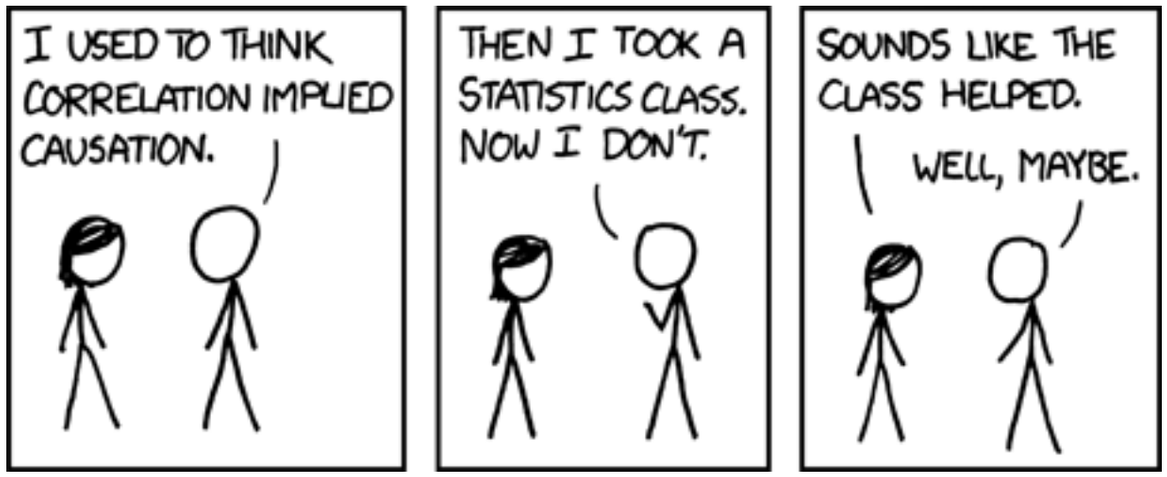


MACHINE LEARNING
DEPARTMENT

10701 Machine Learning

Recitation 7 - Tail bounds and Averages

Ahmed Hefny
Slides mostly by Alex Smola

Carnegie Mellon University

Why this stuff ?

- Can machine learning work ?

Why this stuff ?

- Can machine learning work ?

- Yes, otherwise:
  - No Google
  - No spam-filters
  - No face detectors
  - No 701 midterm
  - I’d be living my life

Why this stuff ?

- Will machine learning *always* work ?

Why this stuff ?

- Will machine learning *always* work ?

- No,

Why this stuff ?

- We need some theory to analyze machine learning algorithms.

- We will go through basic tools used to build theory.

- How well can we estimate stuff from data ?
- What is the convergence behavior of empirical averages ?

Outline

- Estimation Example
- Convergence of Averages
  - Law of Large Numbers
  - Central Limit Theorem
- Inequalities and Tail Bounds
  - Markov Inequality
  - Chebychev's Inequality
  - Heoffding's and McDiarmid's Inequalities
- Proof Tools
  - Union Bound
  - Fourier Analysis
  - Characteristic Functions

Estimating Probabilities

Discrete Distribution

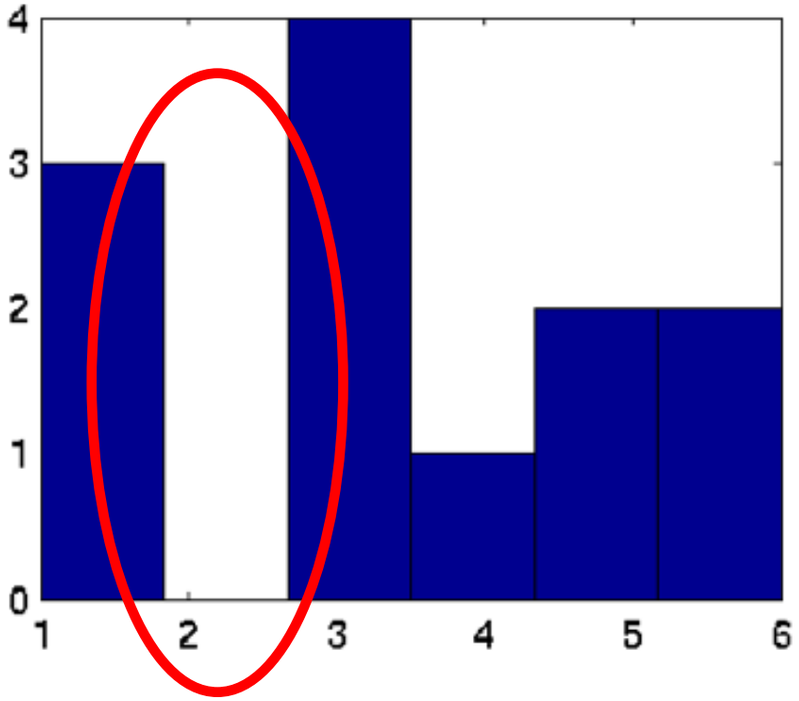
- n outcomes (e.g. faces of a dice)
- Data likelihood $p(X;\pi) = \prod_i \pi_i^{n_i}$

- Maximum Likelihood Estimation
- Constrained optimization problem ... or ...
- Incorporate constraint via $p(x;\theta) = \frac{\exp \theta_x}{\sum_{x'} \exp \theta_{x'}}$
- Taking derivatives yields

$$\theta_i = \log \frac{n_i}{\sum_j n_j} \Longleftrightarrow p(x=i) = \frac{n_i}{\sum_j n_j}$$

Tossing a Die

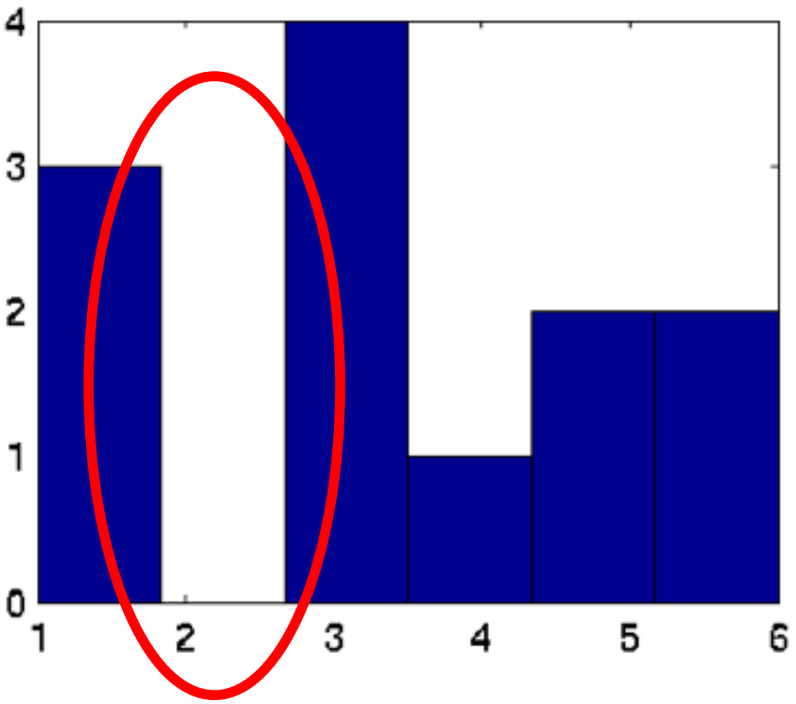
12



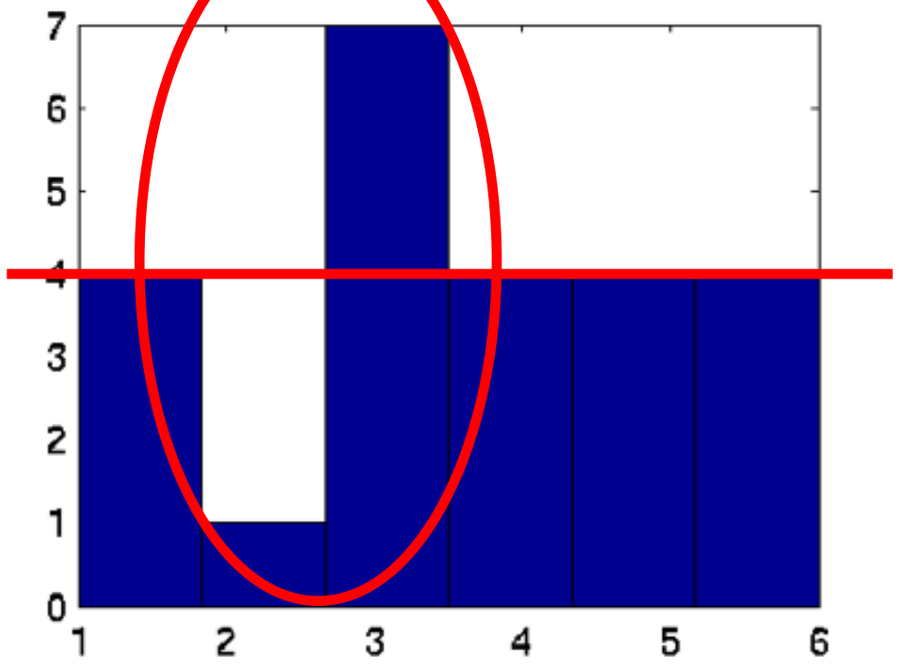
60

Tossing a Die

12



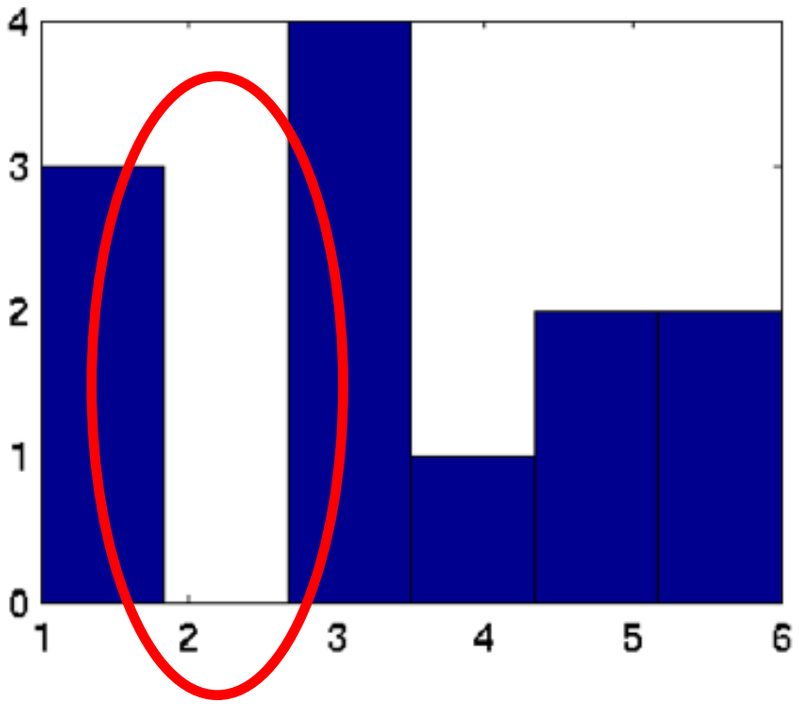
24



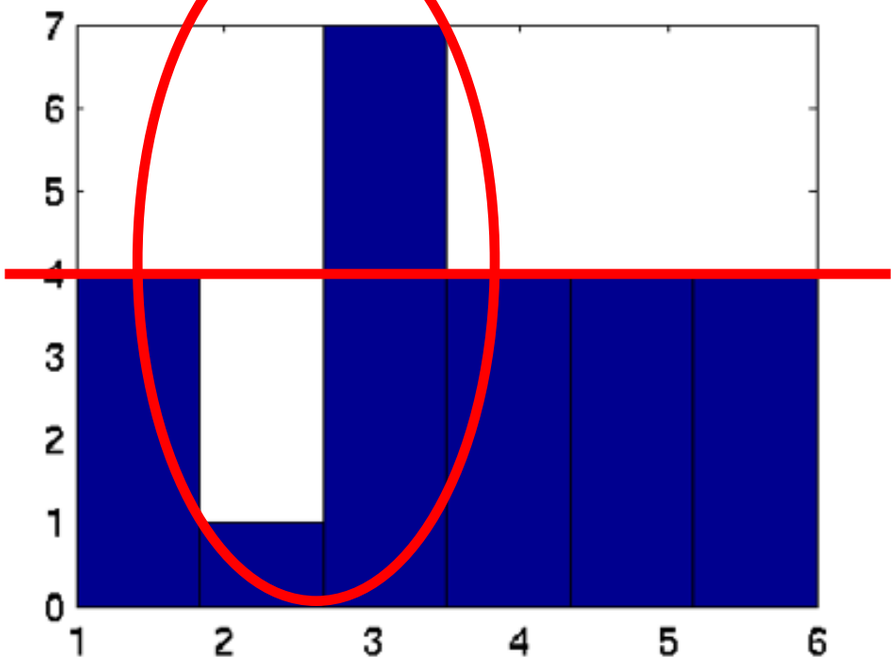
60

Tossing a Die

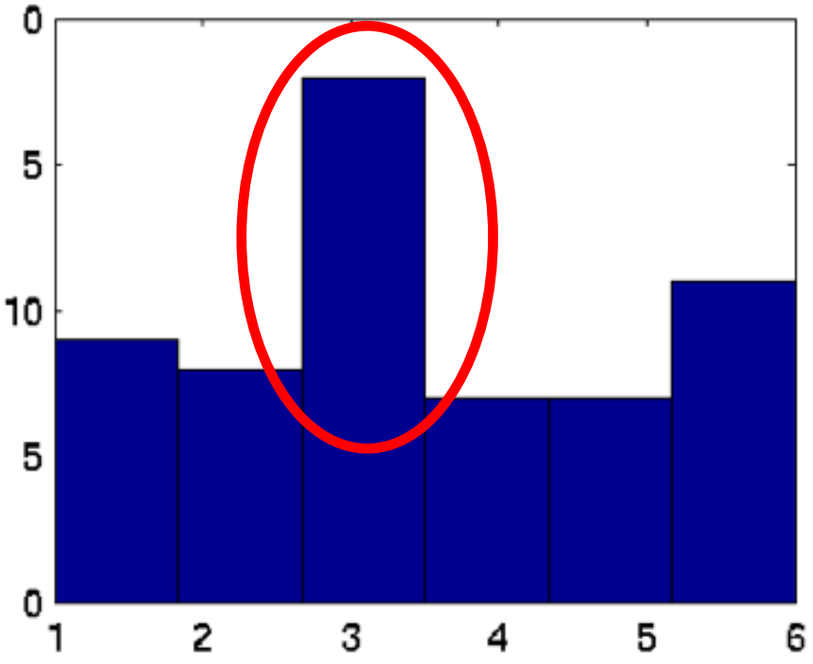
12



24



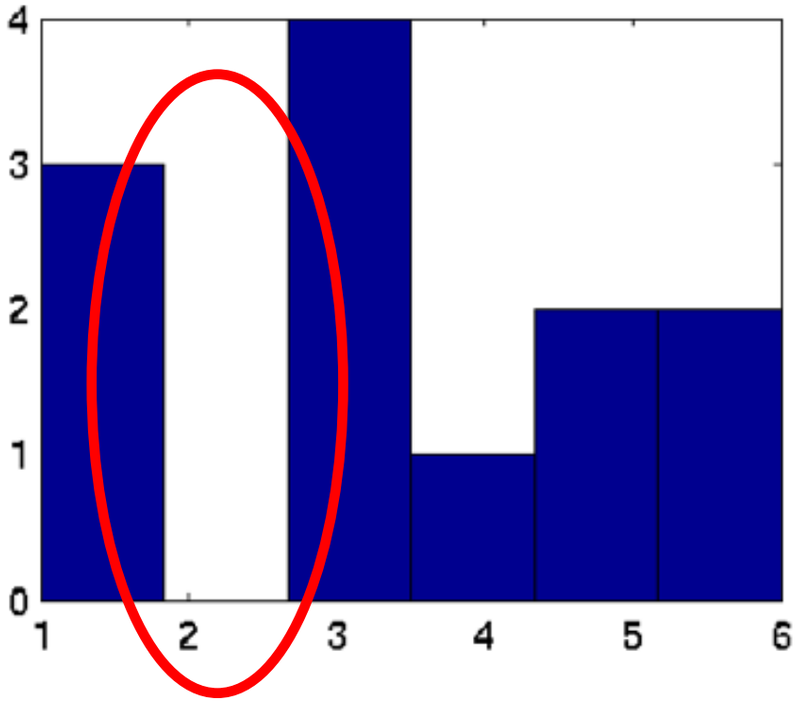
60



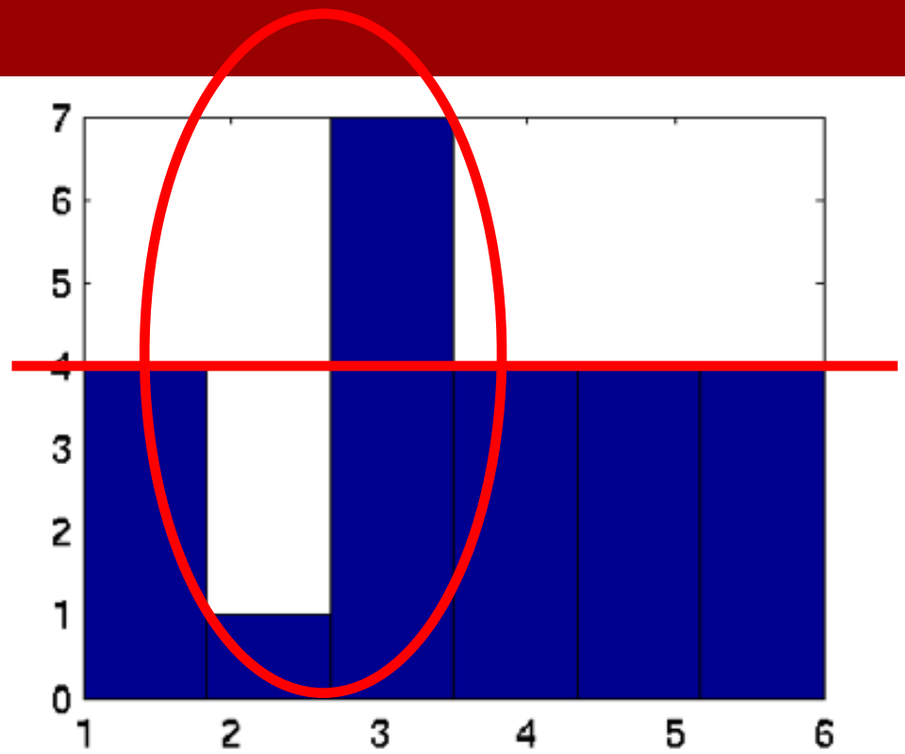
ersity

Tossing a Die

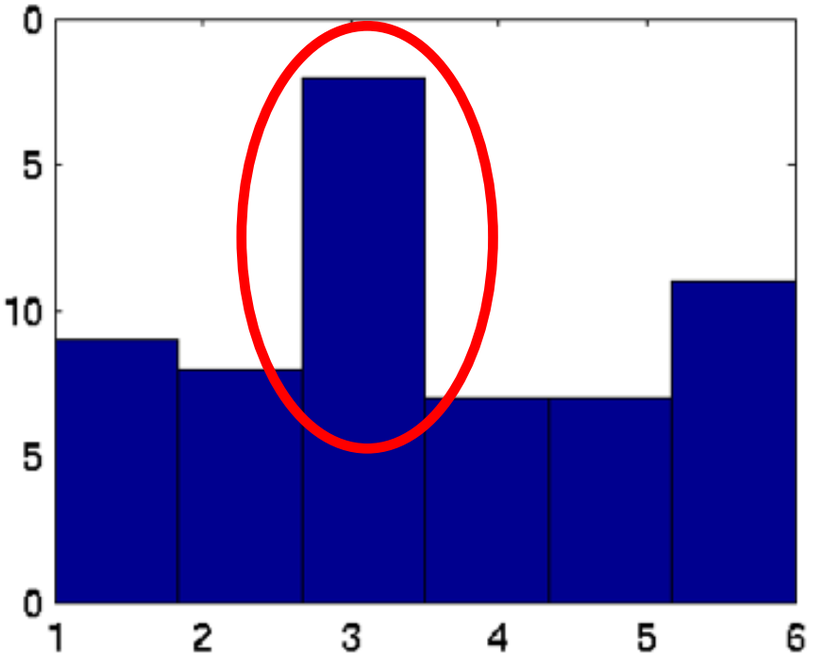
12



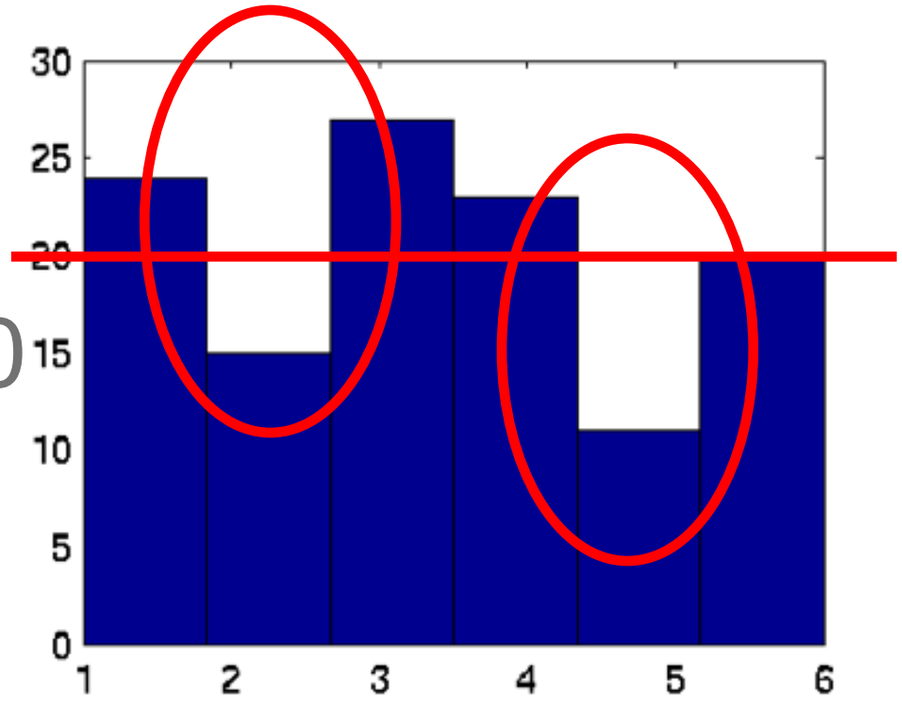
24



60

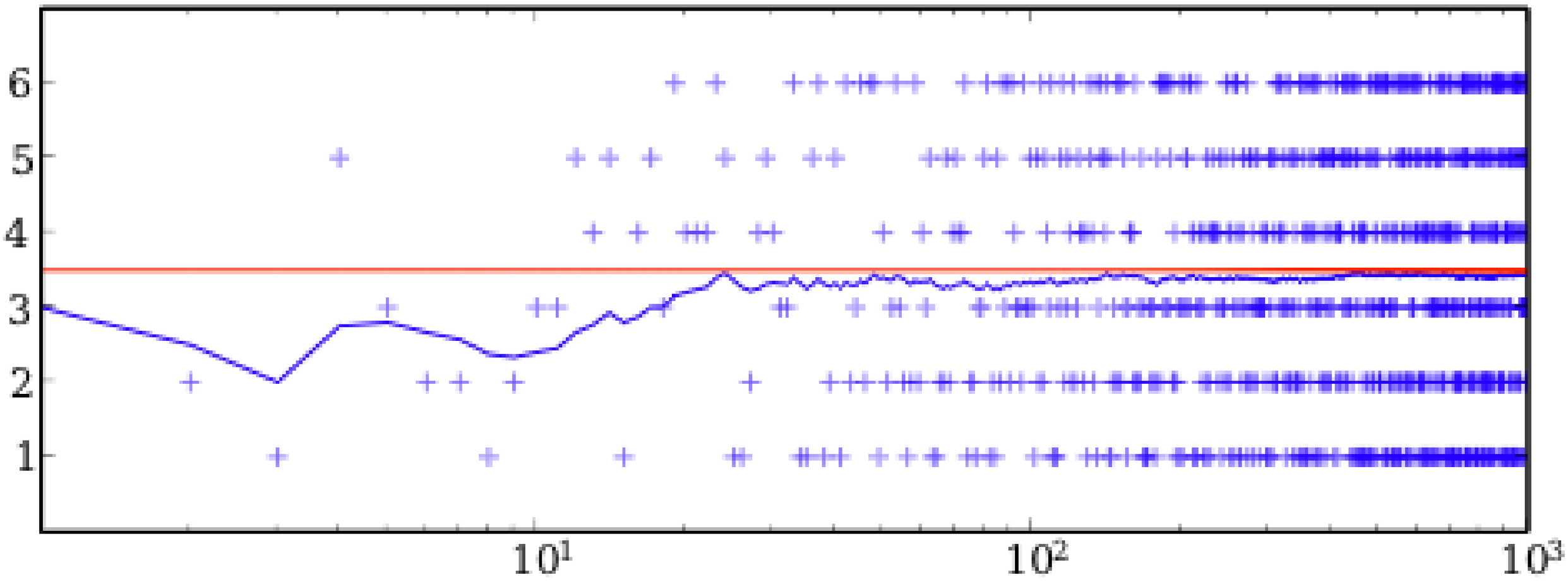


120

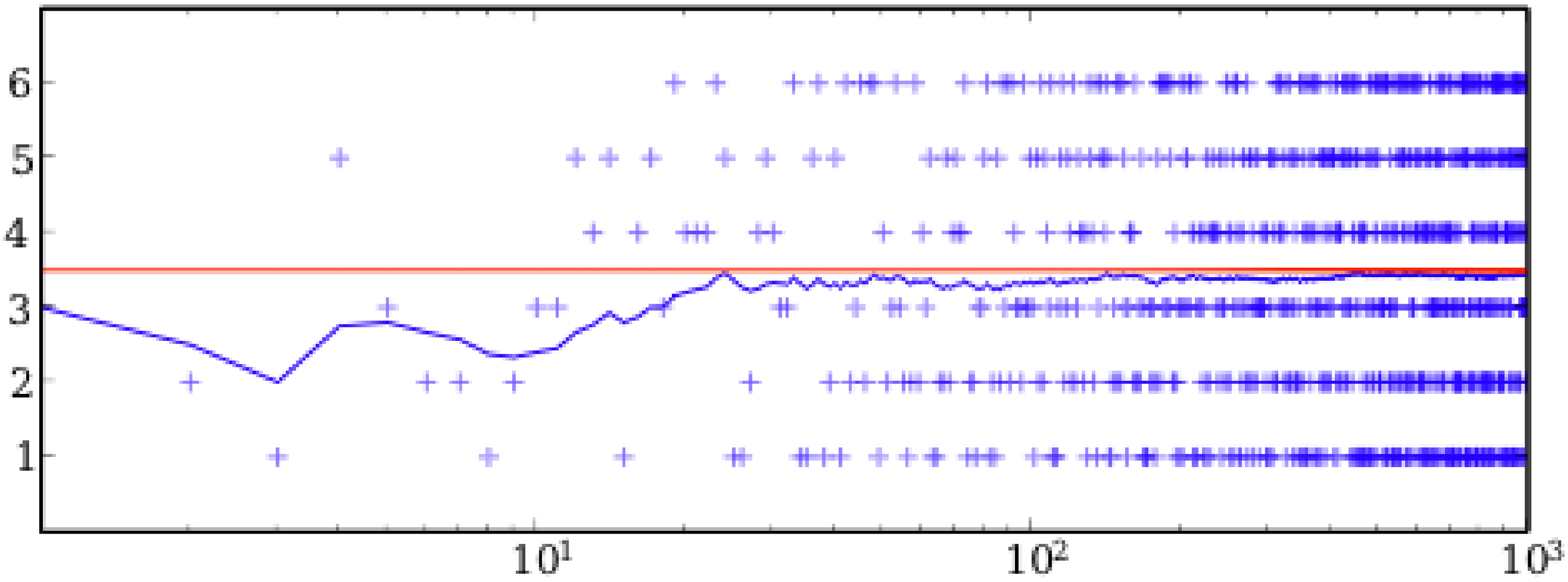


ersity

Empirical average for a die



Empirical average for a die



is it guaranteed to converge ?
how quickly does it converge?

MACHINE LEARNING
DEPARTMENT

Convergence of Empirical Averages

Carnegie Mellon University

Expectations

- Random variable x with probability measure
- Expected value of f(x)

$$\mathbf{E}[f(x)] = \int f(x)dp(x)$$

- Special case - discrete probability mass

$$\Pr\{x = c\} = \mathbf{E}[\{x = c\}] = \int \{x = c\}\, dp(x)$$

(same trick works for intervals)

- Draw x_i identically and independently from p
- Empirical average

$$\mathbf{E}_{\text{emp}}[f(x)] = \frac{1}{n}\sum_{i=1}^{n} f(x_i) \text{ and } \Pr_{\text{emp}}\{x = c\} = \frac{1}{n}\sum_{i=1}^{n} \{x_i = c\}$$

Law of Large Numbers

- Random variables x_i with mean $\mu = \mathbf{E}[x_i]$
- Empirical average $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} x_i$

- Weak Law of Large Numbers

$$\lim_{n\to\infty} \Pr\left(|\hat{\mu}_n - \mu| \leq \epsilon\right) = 1 \text{ for any } \epsilon > 0$$

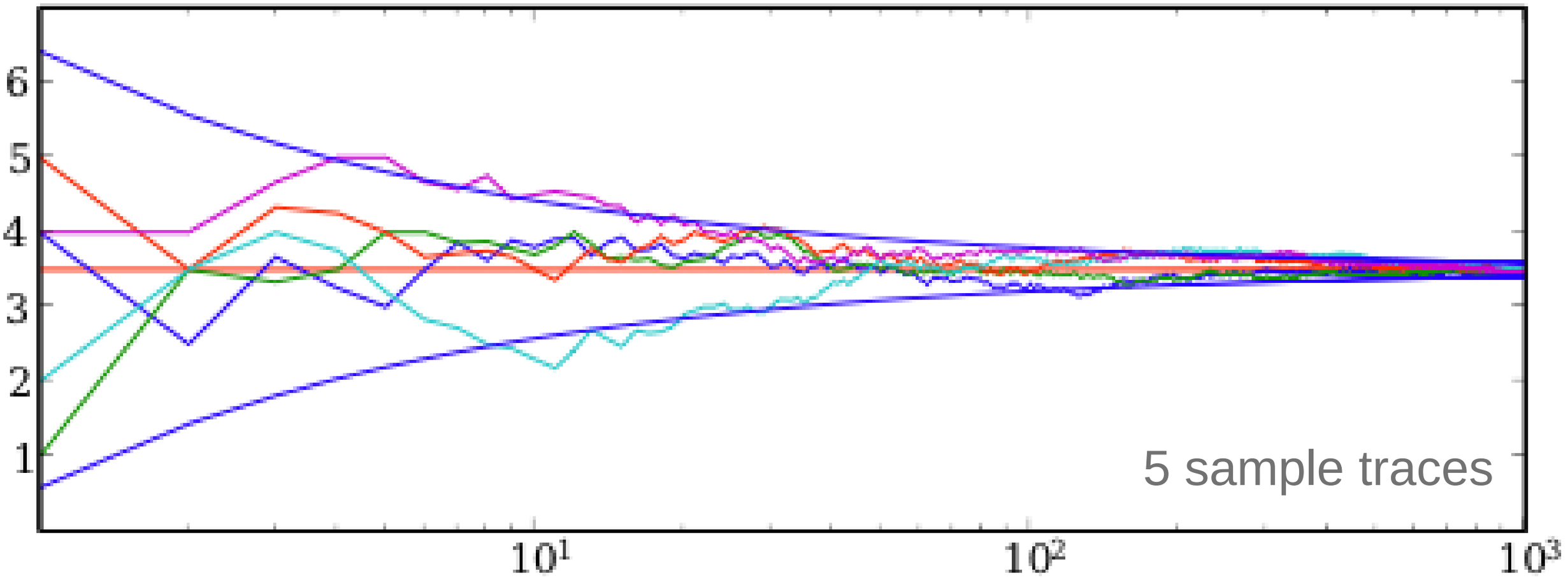
convergence in probability

- Strong Law of Large Numbers

$$\Pr\left(\lim_{n\to\infty} \hat{\mu}_n = \mu\right) = 1$$

Almost sure convergence

Empirical average for a dice



Central Limit Theorem

- Independent random variables x_i with mean μ_i and standard deviation σ_i
- The random variable

$$z_n := \left[\sum_{i=1}^{n} \sigma_i^2\right]^{-\frac{1}{2}} \left[\sum_{i=1}^{n} x_i - \mu_i\right]$$

converges to a Normal Distribution $\mathcal{N}(0,1)$

- Special case - IID random variables & average

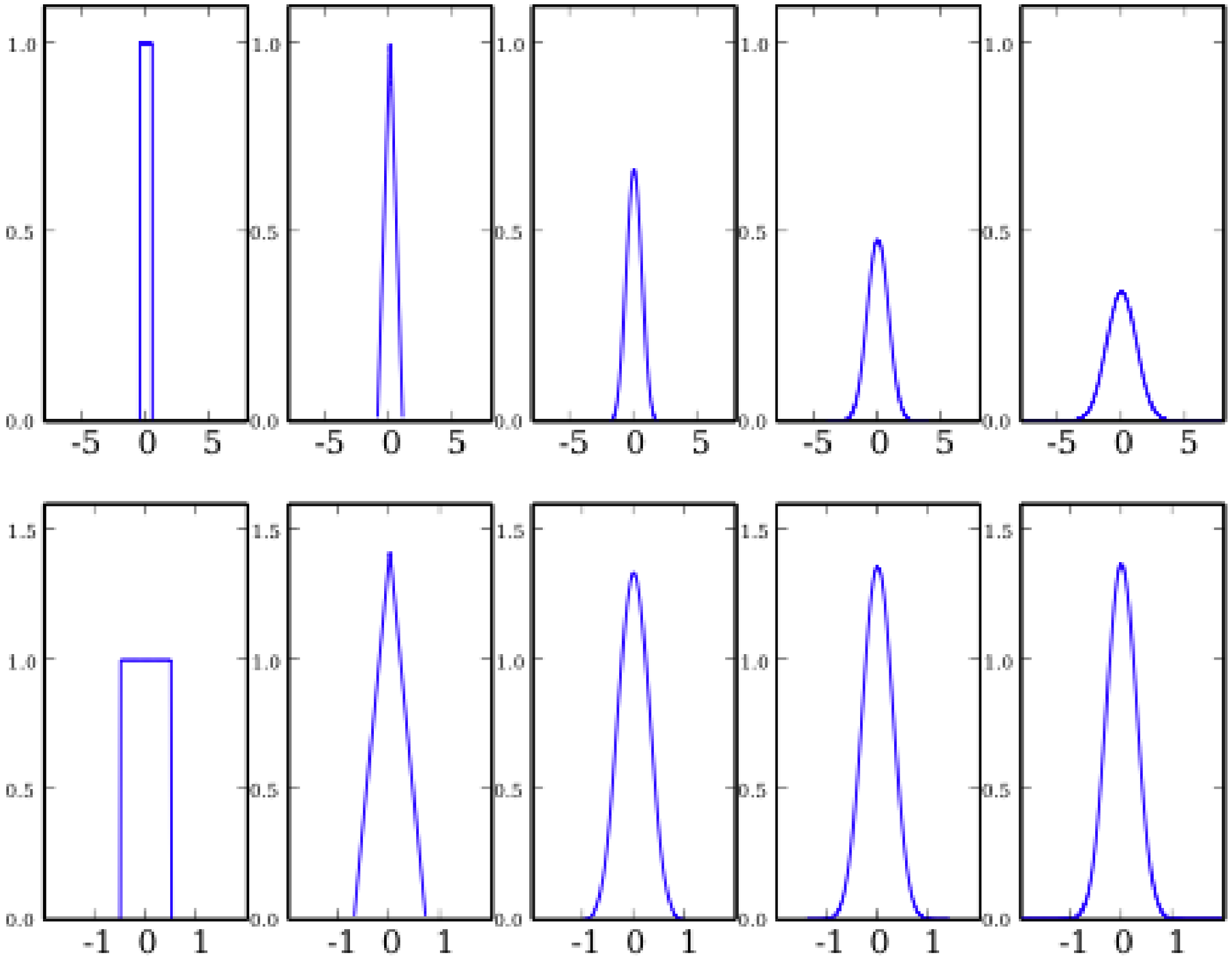
$$\frac{\sqrt{n}}{\sigma}\left[\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right] \to \mathcal{N}(0,1)$$

$O\left(n^{-\frac{1}{2}}\right)$ convergence

Central Limit Theorem in Practice

unscaled

scaled



Tail Bounds

Simple tail bounds

- Gauss Markov inequality
  Non-negative Random variable X with mean μ

$$\Pr(X \geq \epsilon) \leq \mu/\epsilon$$

- Proof - decompose expectation

$$\Pr(X \geq \epsilon) = \int_{\epsilon}^{\infty} dp(x) \leq \int_{\epsilon}^{\infty} \frac{x}{\epsilon} dp(x) \leq \epsilon^{-1} \int_{0}^{\infty} x dp(x) = \frac{\mu}{\epsilon}.$$

Simple tail bounds

- Chebyshev inequality
  Random variable X with mean μ and variance σ^2

$$\Pr(\|\hat{\mu}_m - \mu\| > \epsilon) \leq \sigma^2 m^{-1} \epsilon^{-2} \text{ or equivalently } \epsilon \leq \sigma / \sqrt{m\delta}$$

δ

- Proof - applying Gauss-Markov to $Y = (X - \mu)^2$ with confidence ε^2 yields the result.

Simple tail bounds

- Chebyshev inequality
  Random variable X with mean μ and variance σ^2

$$\Pr(\|\hat{\mu}_m - \mu\| > \epsilon) \leq \sigma^2 m^{-1} \epsilon^{-2} \text{ or equivalently } \epsilon \leq \sigma / \sqrt{m\delta}$$

δ

- Proof - applying Gauss-Markov to $Y = (X - \mu)^2$ with confidence ε^2 yields the result.

Correct ?

Simple tail bounds

- Chebyshev inequality
  Random variable X with mean μ and variance σ^2

$$\Pr(\|\hat{\mu}_m - \mu\| > \epsilon) \leq \sigma^2 m^{-1} \epsilon^{-2} \text{ or equivalently } \epsilon \leq \sigma / \sqrt{m\delta}$$

δ

- Proof - applying Gauss-Markov to $Y = (X - \mu)^2$ with confidence ε^2 yields the result.

Approximately Correct ?

Simple tail bounds

- Chebyshev inequality
  Random variable X with mean μ and variance σ^2

$$\Pr(|\hat{\mu}_m - \mu\| > \epsilon) \leq \sigma^2 m^{-1} \epsilon^{-2} \text{ or equivalently } \epsilon \leq \sigma / \sqrt{m\delta}$$

δ

- Proof - applying Gauss-Markov to $Y = (X - \mu)^2$ with confidence ε^2 yields the result.

Probably Approximately Correct !

Scaling behavior

- Gauss-Markov $\epsilon \leq \frac{\mu}{\delta}$

  Scales properly in μ but expensive in δ

- Chebyshev $\epsilon \leq \frac{\sigma}{\sqrt{m\delta}}$

Proper scaling in σ but still bad in δ

Can we get logarithmic scaling in δ?

Chernoff bound

- For Bernoulli Random Variable with P(x=1)=p
- *Ex:* n independent tosses from biased coin with p probability of getting head

$$\Pr(\hat{\mu}_n - p \geq \epsilon) \leq \exp(-2n\epsilon^2)$$

Chernoff bound

- Proof: We show that

$$\Pr\left\{\sum_i x_i \geq nq\right\} \leq \exp\left(-nK(q,p)\right) \leq \exp\left(-2n(p-q)^2\right)$$

- Where

Pinsker's inequality

$$K(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$$

$$\text{w.l.o.g.} q > p \text{ and set } k \geq qn$$

$$\frac{\Pr\{\sum_i x_i = k|q\}}{\Pr\{\sum_i x_i = k|p\}} = \frac{q^k(1-q)^{n-k}}{p^k(1-p)^{n-k}} \geq \frac{q^{qn}(1-q)^{n-qn}}{p^{qn}(1-p)^{n-qn}} = \exp\left(nK(q,p)\right)$$

$$\sum_{k\geq nq}\Pr\left\{\sum_i x_i = k|p\right\} \leq \sum_{k\geq nq}\Pr\left\{\sum_i x_i = k|q\right\}\exp(-nK(q,p)) \leq \exp(-nK(q,p))$$

Heoffding's Inequality

- If X_i have bounded range c

$$\Pr\left(|\hat{\mu}_m - \mu| > \epsilon\right) \leq 2\exp\left(-\frac{2m\epsilon^2}{c^2}\right).$$

Heoffding's Inequality

- Scaling Behavior

$$\delta := \Pr\left(|\hat{\mu}_m - \mu| > \epsilon\right) \leq 2\exp\left(-\frac{2m\epsilon^2}{c^2}\right)$$

$$\implies \log \delta/2 \leq -\frac{2m\epsilon^2}{c^2}$$

$$\implies \epsilon \leq c\sqrt{\frac{\log 2 - \log \delta}{2m}}$$

- This helps when we need to combine several tail bounds since we only pay logarithmically in terms of their combination.

McDiarmid Inequality

- Generalization of Heoffding's Inequality
- Independent random variables X_i
- Function $f : \mathcal{X}^m \to \mathbb{R}$
- Deviation from expected value

$$\Pr\left(|f(x_1, \ldots, x_m) - \mathbf{E}_{X_1,\ldots,X_m}[f(x_1, \ldots, x_m)]| > \epsilon\right) \leq 2\exp\left(-2\epsilon^2 C^{-2}\right).$$

- Here C is given by $C^2 = \sum_{i=1}^{m} c_i^2$ where

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leq c_i$$

- f is average and X_i have bounded range c → Heoffding's Inequality

More tail bounds

- Higher order moments
  - Bernstein inequality (needs variance bound)

$$\Pr(|\hat{\mu}_m - \mu| \geq \epsilon) \leq 2\exp\left(-\frac{(n\epsilon)^2/2}{\sum_i E[X_i^2] + Cn\epsilon/3}\right)$$

- Absolute / relative error bounds
- Bounds for (weakly) dependent random variables

Summary

- Markov [X is non-negative]
- Chebychev [Finite variance]
- Heoffding [Bound on range]
- Brenestein [Bounded on range + Bound on second moment]

Tighter Bounds
More Assumptions

Tools for the proof

Union Bound

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B)$$

In general

$$P(\bigcup_i A_i) \leq \sum_i P(A_i)$$

Fourier Transform

- Fourier transform relations

$$F[f](\omega) := (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^n} f(x) \exp(-i \langle \omega, x \rangle) dx$$

$$F^{-1}[g](x) := (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^n} g(\omega) \exp(i \langle \omega, x \rangle) d\omega.$$

- Useful identities
  - Identity $F^{-1} \circ F = F \circ F^{-1} = \mathrm{Id}$
  - Derivative $F[\partial_x f] = -i\omega F[f]$
  - Convolution (also holds for inverse transform)

$$F[f \circ g] = (2\pi)^{\frac{d}{2}} F[f] \cdot F[g]$$

The Characteristic Function Method

- Characteristic function

$$\phi_X(\omega) := F^{-1}[p(x)] = \int \exp(i \langle \omega, x \rangle) dp(x)$$

- For X and Y independent we have
  - Joint distribution is convolution

$$p_{X+Y}(z) = \int p_X(z-y) p_Y(y) dy = p_X \circ p_Y$$

  - Characteristic function is product

$$\phi_{X+Y}(\omega) = \phi_X(\omega) \cdot \phi_Y(\omega)$$

  - Proof - plug in definition of Fourier transform
- Characteristic function is unique

Proof - Weak law of large numbers

- Require that expectation exists
- Taylor expansion of exponential

$$\exp(iwx) = 1 + i\langle w, x\rangle + o(|w|)$$
$$\text{and hence } \phi_X(\omega) = 1 + iw\mathbf{E}_X[x] + o(|w|).$$

  (need to assume that we can bound the tail)

- Average of random variables

$$\phi_{\hat{\mu}_m}(\omega) = \left(1 + \frac{i}{m} w\mu + o(m^{-1}|w|)\right)^m$$

convolution

- Limit is constant distribution

vanishing higher order terms

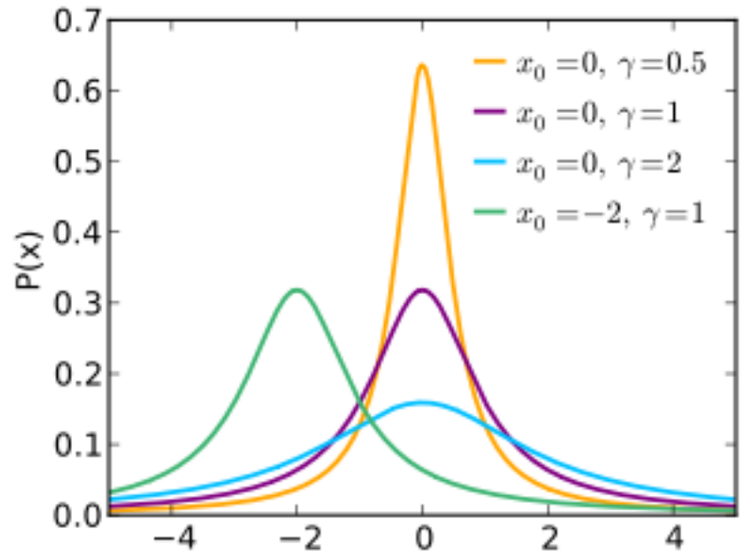
$$\phi_{\hat{\mu}_m}(\omega) \to \exp i\omega\mu = 1 + i\omega\mu + \dots$$

mean

Warning

- Moments may not always exist
  - Cauchy distribution

$$p(x) = \frac{1}{\pi}\frac{1}{1+x^2}$$



- For the mean to exist the following integral would have to converge

$$\int |x| dp(x) \geq \frac{2}{\pi}\int_1^\infty \frac{x}{1+x^2}dx \geq \frac{1}{\pi}\int_1^\infty \frac{1}{x}dx = \infty$$

Proof - Central limit theorem

- Require that second order moment exists (we assume they're all identical WLOG)
- Characteristic function

$$\exp(iwx) = 1 + iwx - \frac{1}{2}w^2x^2 + o(|w|^2)$$

$$\text{and hence } \phi_X(\omega) = 1 + iw\mathbf{E}_X[x] - \frac{1}{2}w^2\text{var}_X[x] + o(|w|^2)$$

- Subtract out mean (centering)

$$z_n := \left[\sum_{i=1}^{n} \sigma_i^2\right]^{-\frac{1}{2}} \left[\sum_{i=1}^{n} x_i - \mu_i\right]$$

$$\phi_{Z_m}(\omega) = \left(1 - \frac{1}{2m}w^2 + o(m^{-1}\,|w|^2)\right)^m \rightarrow \exp\left(-\frac{1}{2}w^2\right) \text{ for } m \rightarrow \infty$$

- This is the FT of a Normal Distribution

Conclusion & what's next ?

We looked at basic building blocks of learning theory

- Convergence of empirical averages
- Tail bounds
- Union bound

Conclusion & what's next ?

Evaluate classifier C on N data points and estimate accuracy. Can we upper-bound estimation error ?

Conclusion & what's next ?

Evaluate classifier C on N data points and estimate accuracy. Can we upper-bound estimation error ?

Yes, Chernoff bound / Heoffding's inequality

Conclusion & what's next ?

Evaluate a set classifiers on N data points and pick the one with best accuracy. Can we upper-bound estimation error ?

Conclusion & what's next ?

Evaluate a set classifiers on N data points and pick the one with best accuracy. Can we upper-bound estimation error ?

Yes, Chernoff bound
/ Heoffding's inequality
+ union bound

Conclusion & what's next ?

What if the set of classifiers is *infinite* ??