Introduction to Machine Learning
CMU-10701

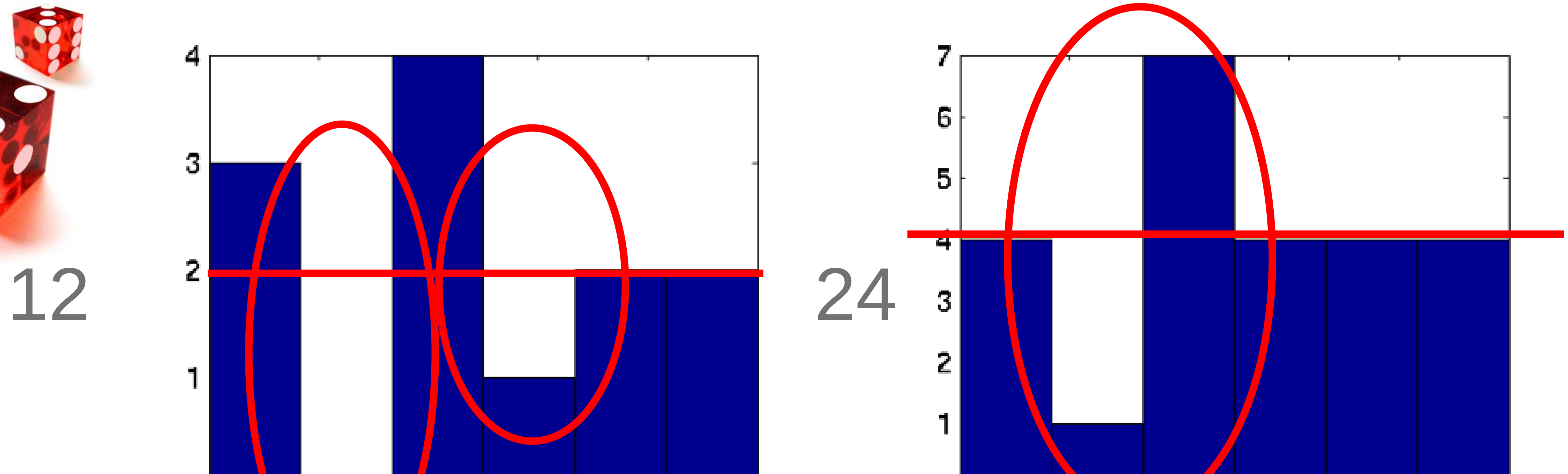
Stochastic Convergence and Tail Bounds

Barnabás Póczos

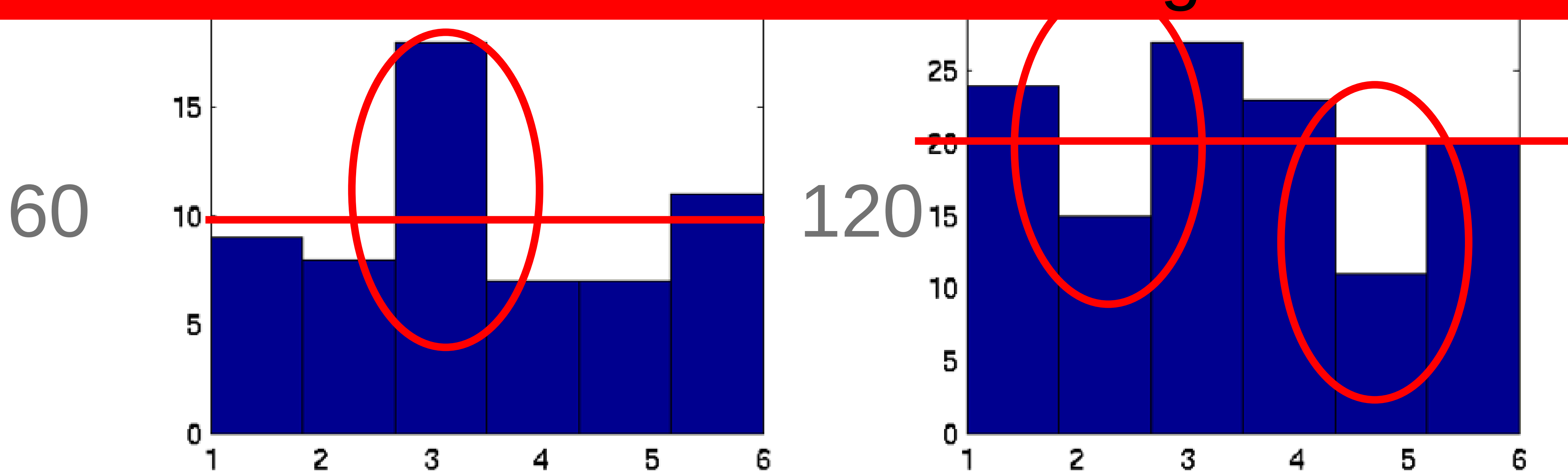
MACHINE LEARNING DEPARTMENT
Carnegie Mellon.
School of Computer Science

Basic Estimation Theory

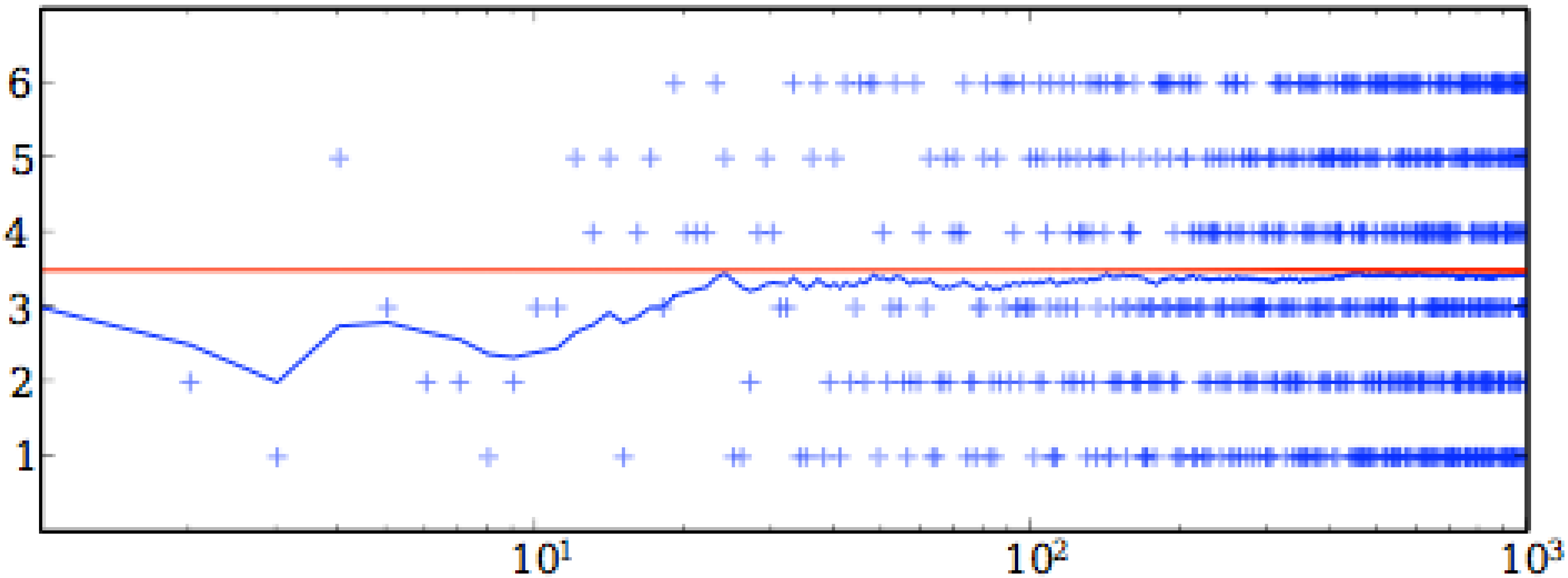
Rolling a Dice,
Estimation of parameters $\theta_1, \theta_2, \ldots, \theta_6$



Does the MLE estimation converge to the right value?
How fast does it converge?

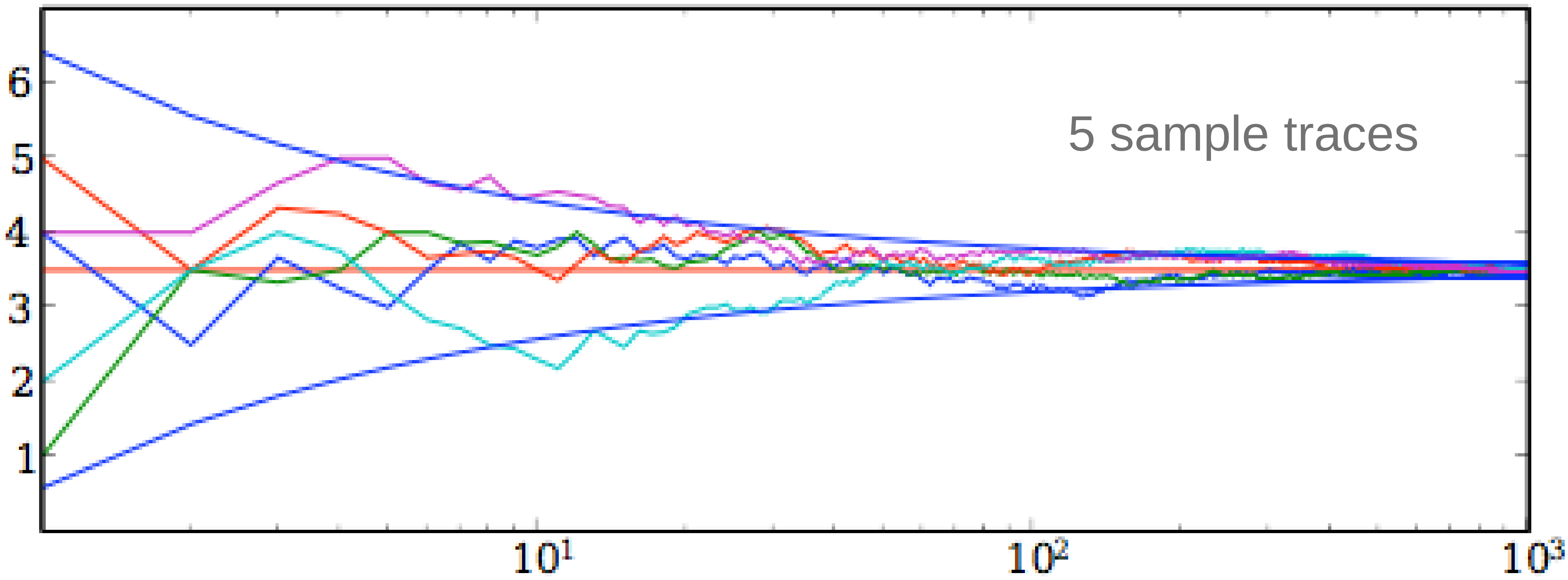


Rolling a Dice
Calculating the Empirical Average



Does the empirical average converge to the true mean?
How fast does it converge?

Rolling a Dice, Calculating the Empirical Average



How fast do they converge? $\mu \pm \sqrt{\mathrm{Var}(x)/n}$

Key Questions

- Do empirical averages converge?
- Does the MLE converge in the dice rolling problem?
- What do we mean on convergence?
- What is the rate of convergence?

I want to know the coin parameter $\theta \in [0,1]$ within $\varepsilon = 0.1$ error, with probability at least $1-\delta = 0.95$.

How many flips do I need?

Outline

Theory:

- Stochastic Convergences:
  - Weak convergence = Convergence in distribution
  - Convergence in probability
  - Strong (almost surely)
  - Convergence in L_p norm

- Limit theorems:
  - Law of large numbers
  - Central limit theorem

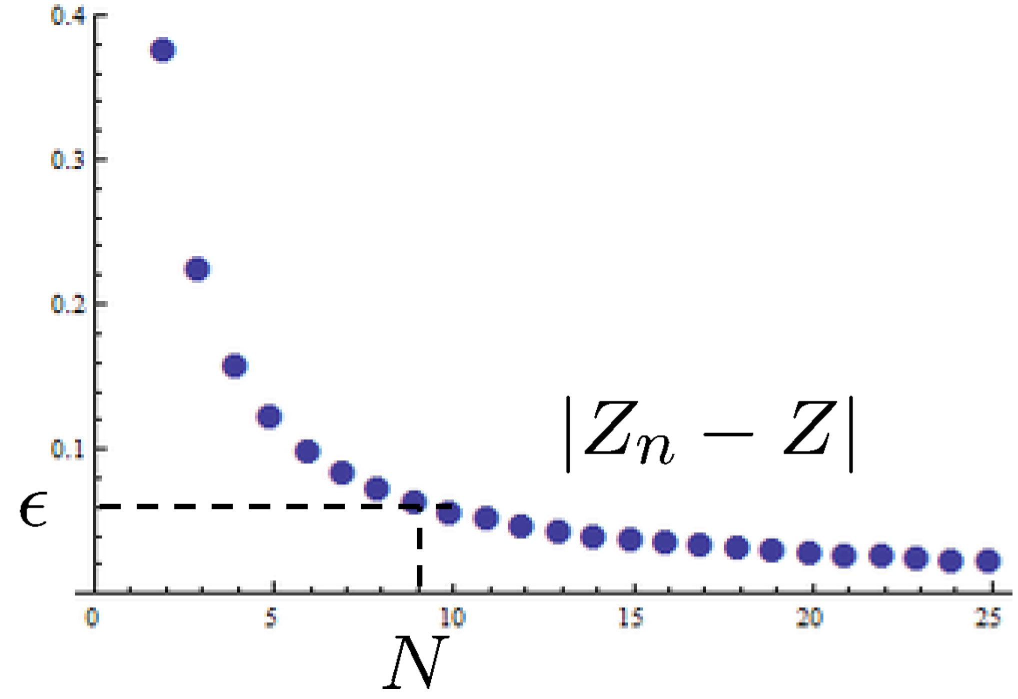
- Tail bounds:
  - Markov, Chebyshev

Stochastic convergence definitions and properties

Convergence of vectors

In $\mathbb{R}^n$ the $Z_n \to Z$ convergence definition is easy:

For each $\epsilon > 0$, there exists a $N > 0$ treshold number such that, for every $n > N$, we have $|Z_n - Z| < \epsilon$.



What do we mean on the convergence of random variables $Z_n \to Z$?

Convergence in Distribution = Convergence Weakly = Convergence in Law

Let $\{Z, Z_1, Z_2, \dots\}$ be a sequence of random variables.

F_n and F are the cumulative distribution functions of Z_n and Z.

Notation:

$$Z_n \xrightarrow{d} Z, \quad Z_n \xrightarrow{\mathcal{D}} Z, \quad Z_n \xrightarrow{\mathcal{L}} Z, \quad Z_n \xrightarrow{d} \mathcal{L}_Z,$$

$$Z_n \rightsquigarrow Z, \quad Z_n \Rightarrow Z, \quad \mathcal{L}(Z_n) \to \mathcal{L}(Z), \quad F_n \xrightarrow{w} F$$

Definition:

$$\lim_{n\to\infty} F_n(z) = F(z), \ \forall z \in \mathbb{R} \text{ at which } F \text{ is continuous}$$

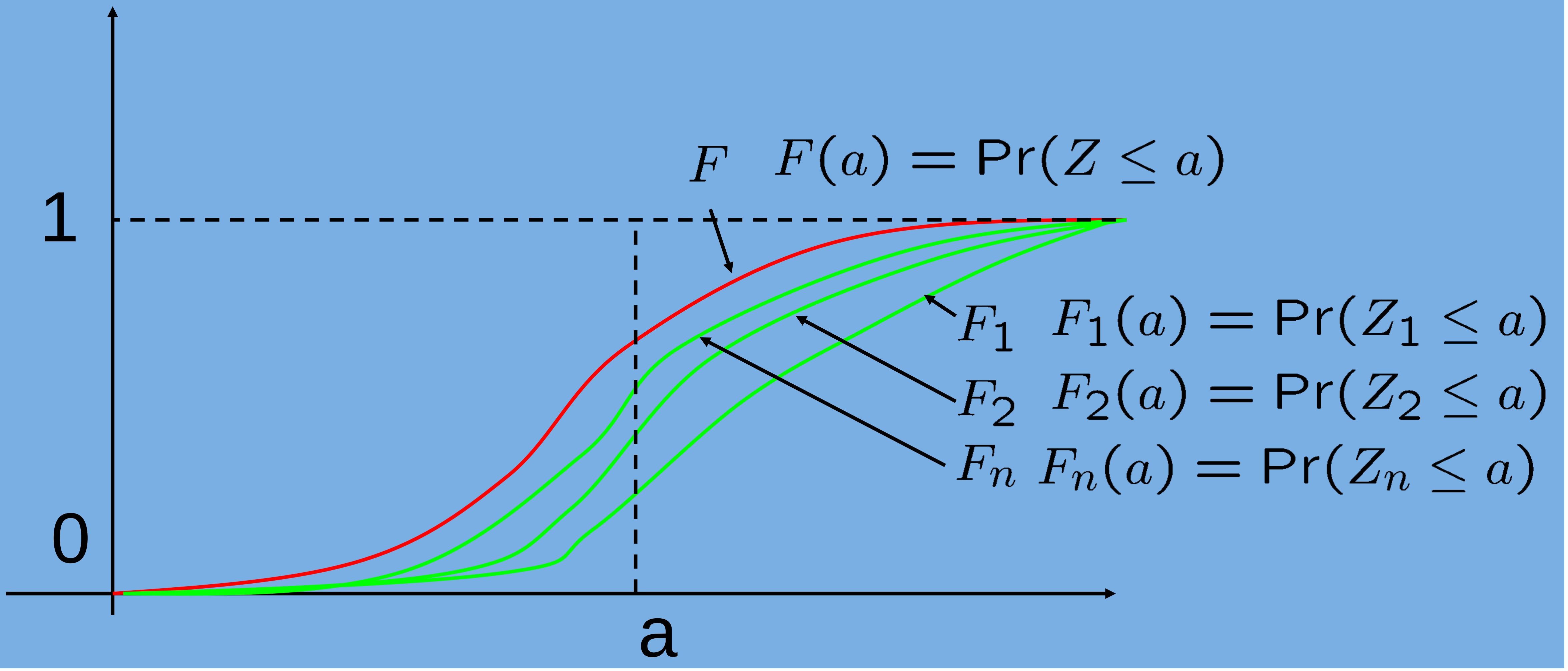
This is the “weakest” convergence.

Convergence in Distribution = Convergence Weakly = Convergence in Law

Only the distribution functions converge!
(NOT the values of the random variables)

$Z_n(\omega)$ can be very different of $Z(\omega)$

Random variable Z_n can be independent of random variable Z.

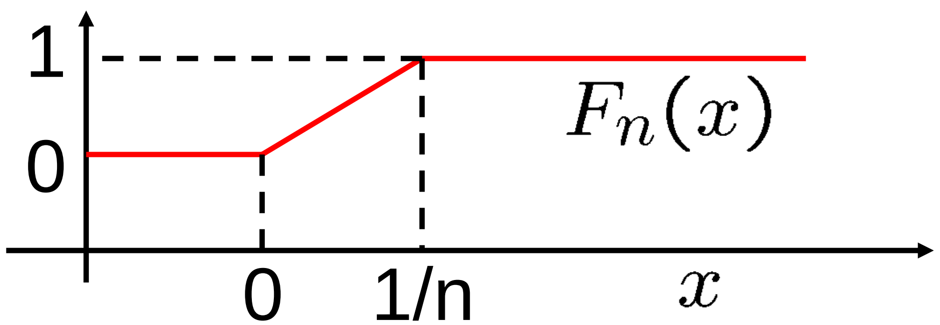


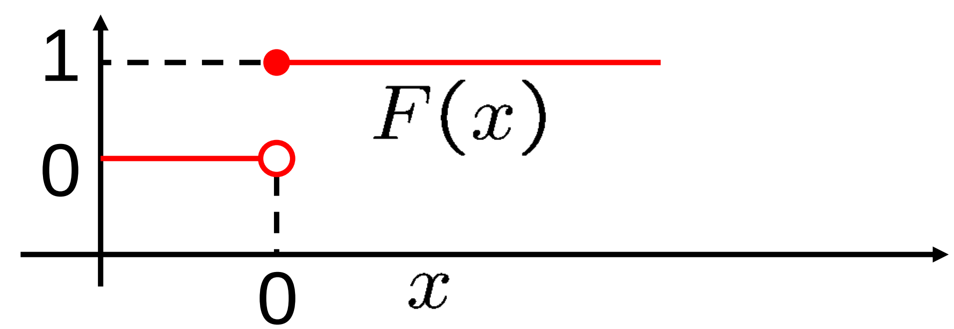
Convergence in Distribution = Convergence Weakly = Convergence in Law

Continuity is important!

Example: Let $Z_n \sim U[0, \frac{1}{n}]$. Then $Z_n \xrightarrow{d} 0$ degenerate variable.

Proof: $F_n(x) = 0$ when $x \leq 0$, and $F_n(x) = 1$ when $x \geq \frac{1}{n}$





The limit random variable is constant 0:

$F(0) = 1$, even though $F_n(0) = 0$ for all n.
$\Rightarrow$ the convergence of cdfs fails at $x = 0$ where F is discontinuous.

In this example the limit Z is discrete, not random (constant 0), although Z_n is a continuous random variable.

Convergence in Distribution = Convergence Weakly = Convergence in Law

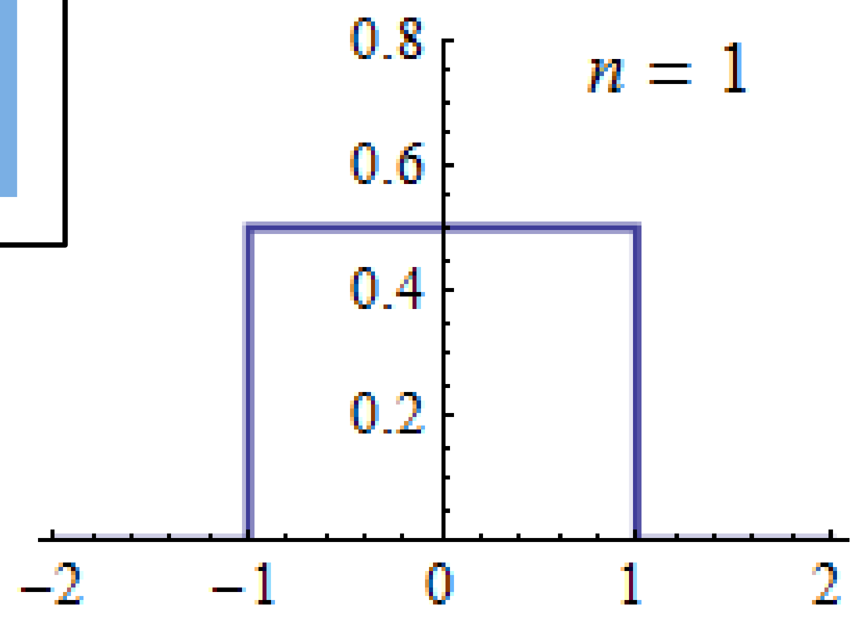
Properties

- For large n, $\Pr(Z_n \le a) \approx \Pr(Z \le a)$, $\forall a$ continuity point of F
  Z_n and Z can still be independent even if their distributions are the same!
- $\mathbb{E}[f(Z_n)] \to \mathbb{E}[f(Z)]$, if f is bounded continuous function.
- *Scheffe's theorem:*
  convergence of the probability density functions $\Rightarrow$ convergence in distribution

$$p_{Z_n}(a) \xrightarrow{n\to\infty} p_Z(a), \text{ for all } a \Rightarrow Z_n \xrightarrow{d} Z.$$
$$p_{Z_n}(a) \xrightarrow{n\to\infty} p_Z(a), \text{ for all } a \not\Leftarrow Z_n \xrightarrow{d} Z.$$

Example:
(Central Limit Theorem)

$$X_n \sim U[-1,1].$$
$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i.$$
$$Z_n \xrightarrow{d} Z \sim \mathcal{N}(0,1)$$

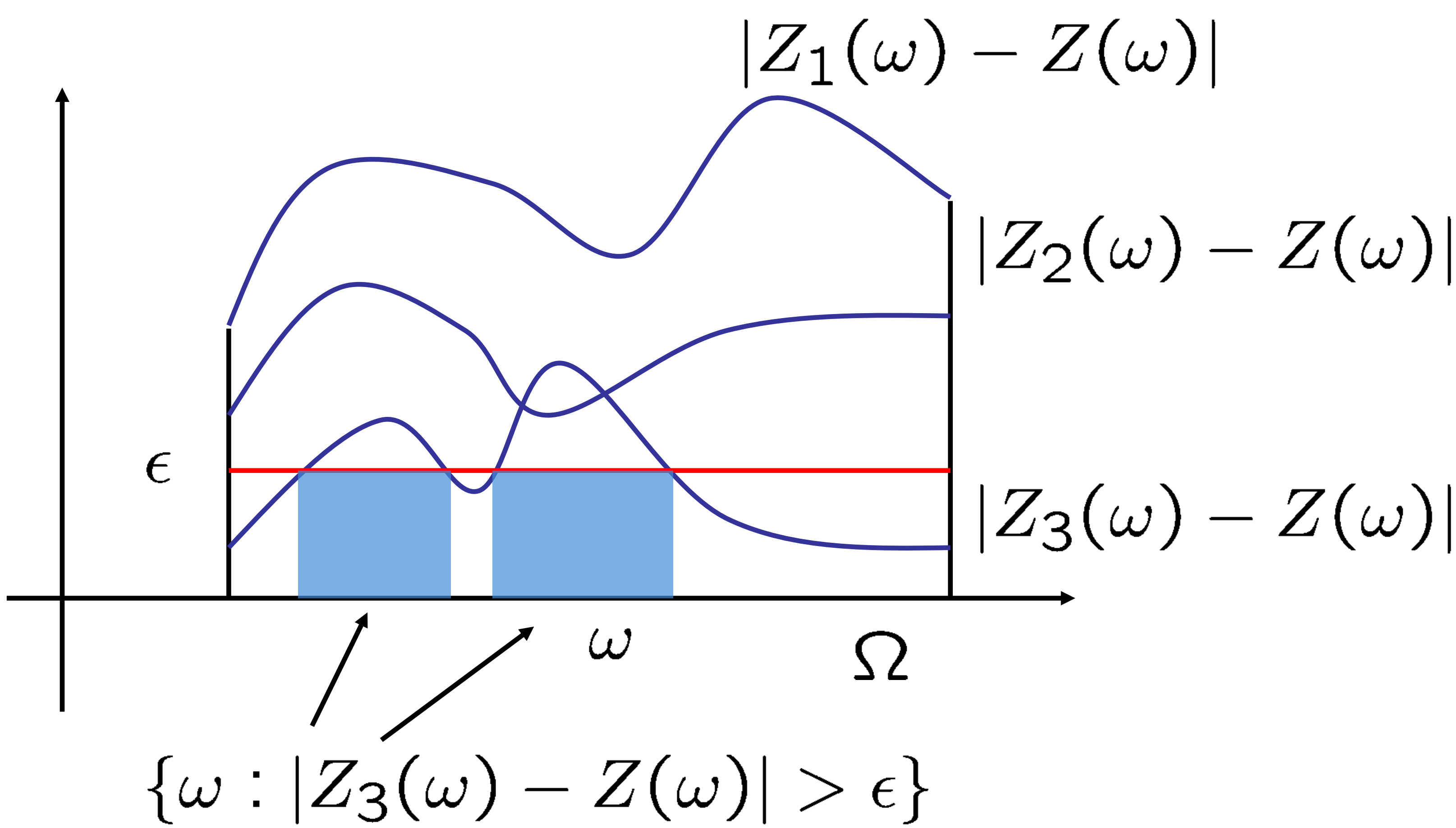


Convergence in Probability

Notation: $Z_n \xrightarrow{p} Z$

Definition: $\forall \varepsilon > 0 \; \lim_{n \to \infty} \Pr\left(|Z_n - Z| \geq \varepsilon\right) = 0.$

$\forall \varepsilon > 0 \; \lim_{n \to \infty} \Pr\left(|Z_n - Z| < \varepsilon\right) = 1.$

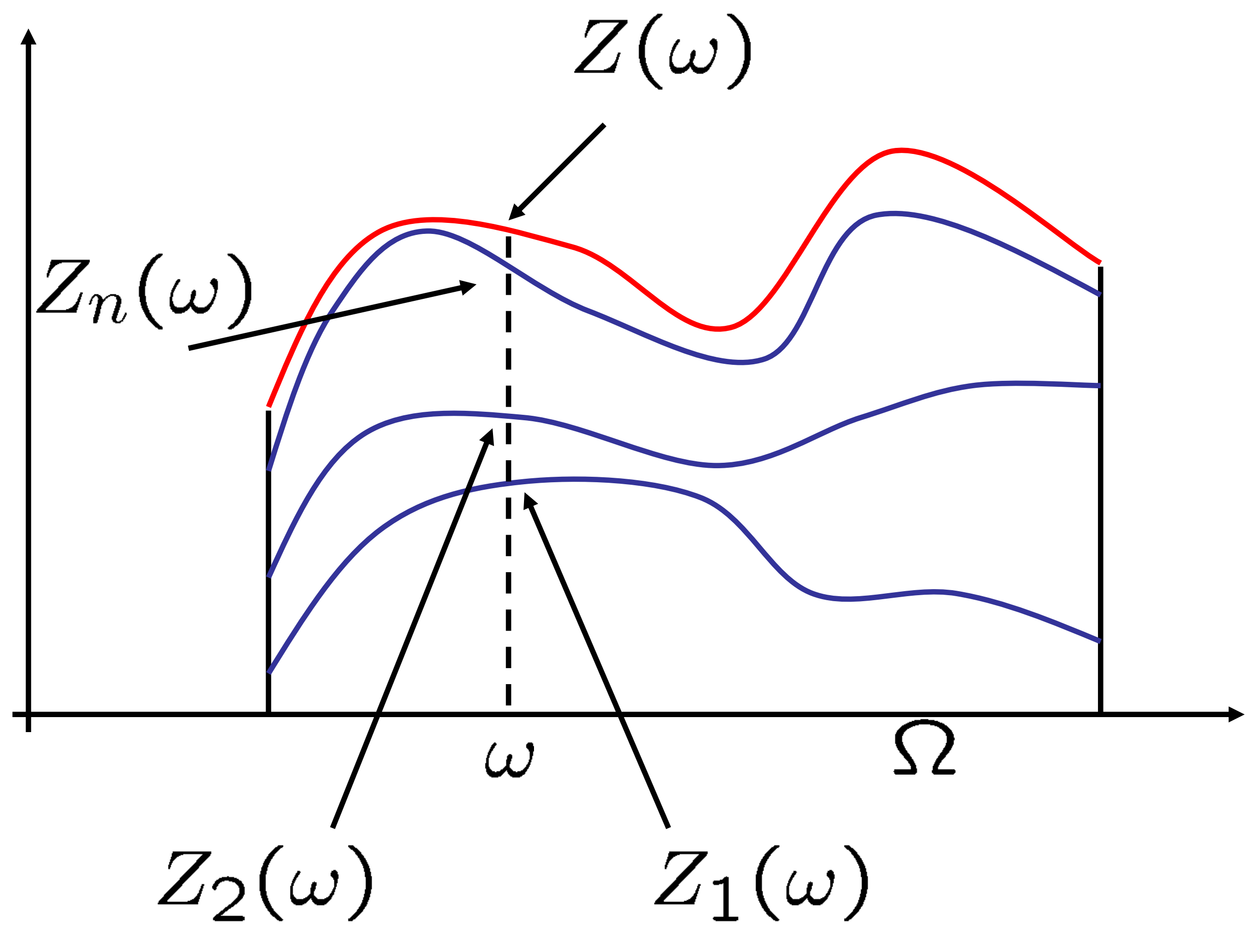


This indeed measures how far the values of $Z_n(\omega)$ and $Z(\omega)$ are from each other.

Almost Surely Convergence

Notation: $Z_n \xrightarrow{a.s.} Z \quad Z_n \to Z$ (w.p. 1)

Definition: $\Pr\left(\omega \in \Omega : \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\right) = 1.$



Convergence in p-th mean, L_p norm

Notation: $Z_n \xrightarrow{L_p} Z$

Definition: $\lim_{n \to \infty} \mathbb{E}\left[|Z_n - Z|^p\right] = 0$

Properties:

$$Z_n \xrightarrow{a.s.} Z$$

$$\Downarrow$$

$$Z_n \xrightarrow{p} Z \Rightarrow Z_n \xrightarrow{d} Z$$

$$\Uparrow$$

$$Z_n \xrightarrow{L_p} Z$$

Counter Examples

$$Z_n \xrightarrow{d} Z \nRightarrow Z_n \xrightarrow{p} Z$$

$$Z_n \xrightarrow{p} Z \nRightarrow Z_n \xrightarrow{a.s.} Z$$

$$Z_n \xrightarrow{p} Z \nRightarrow Z_n \xrightarrow{L_p} Z$$

$$Z_n \xrightarrow{a.s.} Z \nRightarrow Z_n \xrightarrow{L_p} Z$$

$$Z_n \xrightarrow{L_p} Z \nRightarrow Z_n \xrightarrow{a.s.} Z$$

$$Z_n \xrightarrow{a.s.} Z$$

$$Z_n \xrightarrow{p} Z \Rightarrow Z_n \xrightarrow{d} Z$$

$$Z_n \xrightarrow{L_p} Z$$

$$Z_n \xrightarrow{d} Z \Rightarrow \mathbb{E}[f(Z_n)] \to \mathbb{E}[f(Z)]\text{, if } f \text{ is bounded continuous function.}$$

$$Z_n \xrightarrow{d} Z \nRightarrow \mathbb{E}[f(Z_n)] \to \mathbb{E}[f(Z)]\text{, if } f \text{ is general function.}$$

Further Readings on Stochastic convergence

- [**http://en.wikipedia.org/wiki/Convergence_of_random_variables**](http://en.wikipedia.org/wiki/Convergence_of_random_variables)

- **Patrick Billingsley**: Probability and Measure

- **Patrick Billingsley**: Convergence of Probability Measures

Finite sample tail bounds

Useful tools!

Gauss Markov inequality

If X is a nonnegative random variable and $a > 0$, then

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Proof: Decompose the expectation

$$\begin{aligned}\Pr(X \geq a) &= \int_a^\infty p(x)dx \\ &\leq \int_a^\infty \frac{x}{a}p(x)dx = \frac{1}{a}\int_a^\infty xp(x)dx \\ &\leq \frac{1}{a}\int_0^\infty xp(x)dx = \frac{\mathbb{E}[X]}{a}\end{aligned}$$

Corollary: Chebyshev's inequality

Chebyshev inequality

If X is any nonnegative random variable and $a > 0$, then

$$\text{Pr}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\text{Var}(X)}{a^2}$$

Here Var(X) is the variance of X, defined as:

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Proof:

Gauss Markov: $\text{Pr}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$

Apply Gauss-Markov to $(X - \mathbb{E}[X])^2$ with a^2:

$$\text{Pr}((X - \mathbb{E}[X])^2 \geq a^2) \leq \frac{\text{Var}(X)}{a^2}$$

Generalizations of Chebyshev's inequality

Chebyshev: $\Pr(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$

where σ^2 is the variance and $\mu = \mathbb{E}[X]$ is the mean.

This is equivalent to this: $\Pr(-a \leq X - \mu \leq a) \geq 1 - \frac{\sigma^2}{a^2}$

Symmetric two-sided case (X is symmetric distribution**)**

$$\Pr(k_1 < X < k_2) \geq 1 - \frac{4\sigma^2}{(k_2 - k_1)^2}$$

Asymmetric two-sided case (X is asymmetric distribution**)**

$$\Pr(k_1 < X < k_2) \geq \frac{4[(\mu - k_1)(k_2 - \mu) - \sigma^2]}{(k_2 - k_1)^2}$$

There are lots of other generalizations, for example multivariate *X*.

Higher moments?

Markov: $\Pr(|X-\mu| \geq a) \leq \frac{\mathbb{E}[|X-\mu|]}{a}$

Chebyshev: $\Pr(|X-\mu| \geq a) \leq \frac{\mathbb{E}[|X-\mu|^2]}{a^2}$

Higher moments: $\Pr(|X-\mu| \geq a) \leq \frac{\mathbb{E}(|X-\mu|^n)}{a^n}$

where n ≥ 1

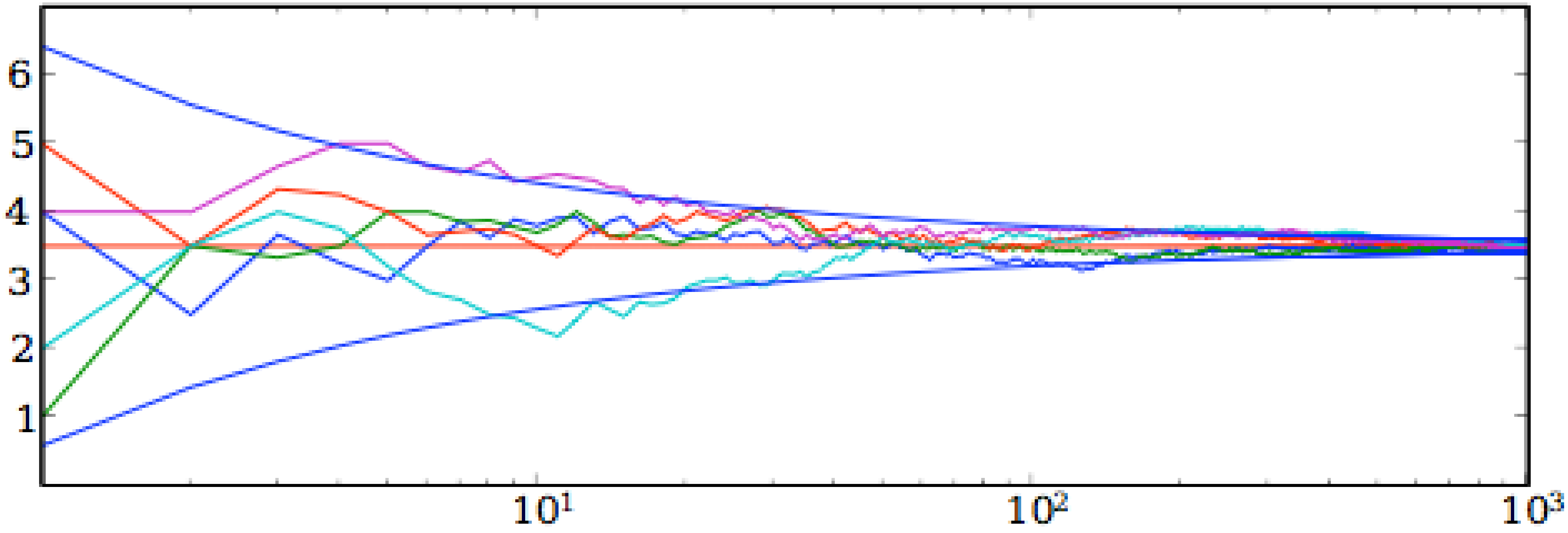
Other functions instead of polynomials?

Exp function: $\Pr(X \geq a) \leq e^{-ta}\mathbb{E}(e^{tX})$ where $a, t, X \geq 0$

Proof: $\Pr(X \geq a) = \Pr(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}}$ (Markov ineq.)

Law of Large Numbers

Do empirical averages converge?



Chebyshev's inequality is good enough to study the question:
Do the empirical averages converge to the true mean?

Answer: Yes, they do. (Law of large numbers)

Law of Large Numbers

$X_1, \ldots, X_n$ i.i.d. random variables with mean $\mu = \mathbb{E}[X_i]$

Empiricial average: $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

Weak Law of Large Numbers: $\widehat{\mu}_n \xrightarrow{p} \mu$

$$\forall \varepsilon > 0 \ \lim_{n\to\infty} \Pr\left(|\widehat{\mu}_n - \mu| \geq \varepsilon\right) = 0.$$

Strong Law of Large Numbers: $\widehat{\mu}_n \xrightarrow{a.s.} \mu$

$$\Pr\left(\omega \in \Omega : \lim_{n\to\infty} \widehat{\mu}_n(\omega) = \mu\right) = 1.$$

Weak Law of Large Numbers

Proof I:

$X_1, \ldots, X_n$ i.i.d., $\mu = \mathbb{E}[X_i]$ $\quad \widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

Assume finite variance. (Not very important) $\text{Var}(X_i) = \sigma^2$, (for all i)

$$\text{Var}(\widehat{\mu}_n) = \text{Var}(\tfrac{1}{n}(X_1 + \cdots + X_n)) = \frac{1}{n^2}\text{Var}(X_1 + \cdots + X_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

$\mathbb{E}[\widehat{\mu}_n] = \mu.$

Using Chebyshev's inequality on $\widehat{\mu}_n$ results in $\Pr(|\widehat{\mu}_n - \mu| \geq \varepsilon) \leq \dfrac{\sigma^2}{n\varepsilon^2}.$

Therefore,

$$\Pr(|\widehat{\mu}_n - \mu| < \varepsilon) = 1 - \Pr(|\widehat{\mu}_n - \mu| \geq \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}.$$

As *n* approaches infinity, this expression approaches 1.

$$\Rightarrow \widehat{\mu}_n \xrightarrow{P} \mu \quad \text{for} \quad n \to \infty.$$

Fourier Transform
and Characteristic Function

Fourier Transform

Fourier transform

unitary transf.

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

Inverse Fourier transform

$$f(x) = \mathcal{F}^{-1}[\hat{f}](x) = \int_{\mathbb{R}^d} \hat{f}(\omega) \exp(2\pi i \langle \omega, x \rangle) d\omega$$

Other conventions: Where to put 2π?

Not preferred: not unitary transf.
Doesn't preserve inner product

$$\hat{f}(\omega) = \int_{\mathbb{R}^n} f(x) \exp(-i \langle \omega, x \rangle)\, dx.$$

$$f(x) = \mathcal{F}^{-1}[\hat{f}](x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{f}(\omega) \exp(i \langle \omega, x \rangle)\, d\omega$$

unitary transf.

$$\hat{f}(\omega) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(x) \exp(-i \langle \omega, x \rangle)\, dx$$

$$f(x) = \mathcal{F}^{-1}[\hat{f}](x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \hat{f}(\omega) \exp(i \langle \omega, x \rangle)\, d\omega$$

Fourier Transform

Fourier transform

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

Inverse Fourier transform

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

Properties:

Inverse is really inverse: $F \circ F^{-1}[g] = g \;\; F^{-1} \circ F[f] = f$

and lots of other important ones…

Fourier transformation will be used to define the characteristic function, and represent the distributions in an alternative way.

Characteristic function

How can we describe a random variable?

- cumulative distribution function (cdf)

$$F_X(x) = \Pr(X \leq x) = \mathbb{E}\left[\mathbf{1}_{\{X \leq x\}}\right]$$

- probability density function (pdf)

The Characteristic function provides an alternative way for describing a random variable

Definition:

$$\varphi_X(t) = \mathbb{E}\left[e^{i\langle t,x\rangle}\right] = \int_{\mathbb{R}^d} e^{i\langle t,x\rangle}\, dF_X(x) = \int_{\mathbb{R}^d} e^{i\langle t,x\rangle} f_X(x)\, dx$$

The Fourier transform of the density

Characteristic function

$$\varphi_X(t) = \mathbb{E}\left[e^{i\langle t,x\rangle}\right] = \int_{\mathbb{R}^d} e^{i\langle t,x\rangle}\, dF_X(x) = \int_{\mathbb{R}^d} e^{i\langle t,x\rangle} f_X(x)\, dx$$

Properties

- $\varphi_X(t)$ of a real-valued random variable X always exists.

  For example, Cauchy doesn't have mean but still has characteristic function.

- Continuous on the entire space, even if *X* is not continuous.
- Bounded, even if *X* is not bounded $|\varphi_X(t)| \leq 1, \ \forall t \in \mathbb{R}^d$.
- Bijection between cdf and characteristic functions: For any two random variables $X_1,\ X_2,\ F_{X_1} = F_{X_2} \ \Leftrightarrow \ \varphi_{X_1} = \varphi_{X_2}$
- $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$ if $X \perp\!\!\!\perp Y$.
- $\varphi_{\frac{1}{n}X}(t) = \varphi_X(\frac{t}{n})$
- Characteristic function of constant *a*: $\varphi_{\delta_a}(t) = \exp(i\langle t, a\rangle)$
- Levi's: continuity theorem $\varphi_{X_n}(t) \to \varphi_X(t) \quad \forall t \in \mathbb{R} \Rightarrow X_n \xrightarrow{\mathcal{D}} X$

Weak Law of Large Numbers

Proof II: Goal: $\widehat{\mu}_n \xrightarrow{D} \mu$.

Taylor's theorem for complex functions

$$\exp(itx) = 1 + itx + o(t), \quad t \to 0$$

The Characteristic function

$$\varphi_X(t) = \mathbb{E}[\exp(itX)] = 1 + it\mu + o(t)$$

Properties of characteristic functions :

$$\varphi_{\frac{1}{n}X}(t) = \varphi_X(\tfrac{t}{n}) \quad \text{and} \quad \varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t) \quad \text{if } X \perp\!\!\!\perp Y.$$

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$$

$$\Rightarrow \varphi_{\widehat{\mu}_n}(t) = \left[\varphi_X\left(\frac{t}{n}\right)\right]^n = \left[1 + i\mu\frac{t}{n} + o\left(\frac{t}{n}\right)\right]^n \xrightarrow{n\to\infty} e^{it\mu} = 1 + t\mu + \ldots$$

mean

Levi's continuity theorem $\Rightarrow$ Limit is a constant distribution with mean μ

"Convergence rate" for LLN

Gauss-Markov:

$$\Pr(|\widehat{\mu}_n - \mu| < \varepsilon) \geq 1 - \frac{\mathbb{E}[|\widehat{\mu}_n - \mu|]}{\varepsilon} = 1 - \delta$$

Doesn't give rate

Chebyshev:

$$P(|\widehat{\mu}_n - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2} = 1 - \delta. \Rightarrow |\widehat{\mu}_n - \mu| < \varepsilon = \frac{\sigma}{\sqrt{n\delta}}$$

with probability 1-δ

Can we get smaller, logarithmic error in δ???

$$\sqrt{\log \tfrac{1}{\delta}} \ll \frac{1}{\sqrt{\delta}} \text{ if } 0 < \delta < 1$$

Further Readings on LLN, Characteristic Functions, etc

- http://en.wikipedia.org/wiki/Levy_continuity_theorem
- http://en.wikipedia.org/wiki/Law_of_large_numbers
- http://en.wikipedia.org/wiki/Characteristic_function_(probability_theory)
- http://en.wikipedia.org/wiki/Fourier_transform

More tail bounds

More useful tools!

Hoeffding's inequality (1963)

$$\left.\begin{array}{r} X_1, \ldots, X_n \text{ independent} \\ X_i \in [a_i, b_i] \\ \varepsilon > 0 \end{array}\right\} \Rightarrow$$

$$\Rightarrow \begin{cases} \mathbb{P}(|\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i)| > \varepsilon) \leq 2\exp\left(\frac{-2n\varepsilon^2}{\frac{1}{n}\sum_{i=1}^n (b_i - a_i)^2}\right) \\ \text{two-sided} \\ \\ \mathbb{P}(\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i) > \varepsilon) \leq \exp\left(\frac{-2n\varepsilon^2}{\frac{1}{n}\sum_{i=1}^n (b_i - a_i)^2}\right) \\ \text{one-sided} \end{cases}$$

It only contains the range of the variables, but not the variances.

"Convergence rate" for LLN from Hoeffding

Hoeffding Let $c^2 = \frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2$

$$\Rightarrow \Pr(|\widehat{\mu}_n - \mu| > \varepsilon) \le 2\exp\left(\frac{-2n\varepsilon^2}{c^2}\right)$$

$$\delta = 2\exp\left(\frac{-2n\varepsilon^2}{c^2}\right)$$

$$\log\frac{\delta}{2} = \frac{-2n\varepsilon^2}{c^2}$$

$$\frac{c^2}{2n}\log\frac{2}{\delta} = \varepsilon^2$$

$$\varepsilon = c\sqrt{\frac{\log 2 - \log\delta}{2n}}$$

$$\Rightarrow |\widehat{\mu}_n - \mu| < \varepsilon = c\sqrt{\frac{1}{2n}\log\frac{2}{\delta}} \qquad \ll \frac{\sigma}{\sqrt{n\delta}}$$

Proof of Hoeffding's Inequality

A few minutes of calculations.

Bernstein's inequality (1946)

$$\left.\begin{array}{r} X_1, \ldots, X_n \text{ indep.} \\ X_i \in [a,b] \\ \sigma^2 = \frac{1}{n}\sum_{i=1}^n Var(X_i) \\ \varepsilon > 0 \end{array}\right\} \Rightarrow$$

$$\Rightarrow \mathbb{P}(|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X_i| > \varepsilon) \le 2\exp\left(\frac{-n\varepsilon^2}{2\sigma^2 + \frac{2}{3}\varepsilon(b-a)}\right)$$

It contains the variances, too, and can give tighter bounds than Hoeffding.

Benett's inequality (1962)

$$\left.\begin{array}{r} X_1, \ldots, X_n \text{ indep.} \\ \mathbb{E}X_i = 0 \\ |X_i| \leq a \\ \sigma^2 = \frac{1}{n}\sum_{i=1}^n Var(X_i) \\ h(u) \doteq (1+u)\log(1+u) - u, \quad u \geq 0 \end{array}\right\} \Rightarrow$$

$$\Rightarrow \mathbb{P}\left(\sum_{i=1}^n X_i > t\right) \leq \exp\left(-\frac{n\sigma^2}{a^2} h\left(\frac{at}{n\sigma^2}\right)\right)$$

Benett's inequality $\Rightarrow$ Bernstein's inequality.

Proof:

$$h(u) \geq \frac{u^2}{2+2u/3} \qquad t = n\varepsilon \qquad n\sigma^2 h\left(\frac{n\varepsilon}{n\sigma^2}\right) \geq \ldots \geq \frac{n\varepsilon^2}{2\sigma^2 + \frac{2}{3}\varepsilon}$$

McDiarmid's Bounded Difference Inequality

Suppose $X_1, X_2, \ldots, X_n$ are independent and assume that

$$\sup_{x_1, x_2, \ldots, x_n, \widehat{x}_i} |f(x_1, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_{i-1}, \widehat{x}_i, x_{i+1}, \ldots, x_n)| \le c_i$$

$$\text{for } 1 \le i \le n$$

(In other words, replacing the i-th coordinate x_i by some other value changes the value of f by at most c_i.)

It follows that

$$\Pr\{f(X_1, X_2, \ldots, X_n) - E[f(X_1, X_2, \ldots, X_n)] \ge \varepsilon\} \le \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

$$\Pr\{E[f(X_1, X_2, \ldots, X_n)] - f(X_1, X_2, \ldots, X_n) \ge \varepsilon\} \le \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

$$\Pr\{|E[f(X_1, X_2, \ldots, X_n)] - f(X_1, X_2, \ldots, X_n)| \ge \varepsilon\} \le 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Further Readings on Tail bounds

http://en.wikipedia.org/wiki/Hoeffding's_inequality

http://en.wikipedia.org/wiki/Doob_martingale (McDiarmid)

http://en.wikipedia.org/wiki/Bennett%27s_inequality

http://en.wikipedia.org/wiki/Markov%27s_inequality

http://en.wikipedia.org/wiki/Chebyshev%27s_inequality

http://en.wikipedia.org/wiki/Bernstein_inequalities_(probability_theory)

Limit Distribution?

Central Limit Theorem

Let $X_1, \ldots, X_n$ be i.i.d $E[X_i] = \mu$ and $Var[X_i] = \sigma^2$.
LLN: $\frac{X_1 + \ldots + X_n}{n} - \mu \xrightarrow{a.s.} 0$

Lindeberg-Lévi CLT: $X_1, \ldots, X_n$ i.i.d, $E[X_i] = \mu$, and $Var[X_i] = \sigma^2$.

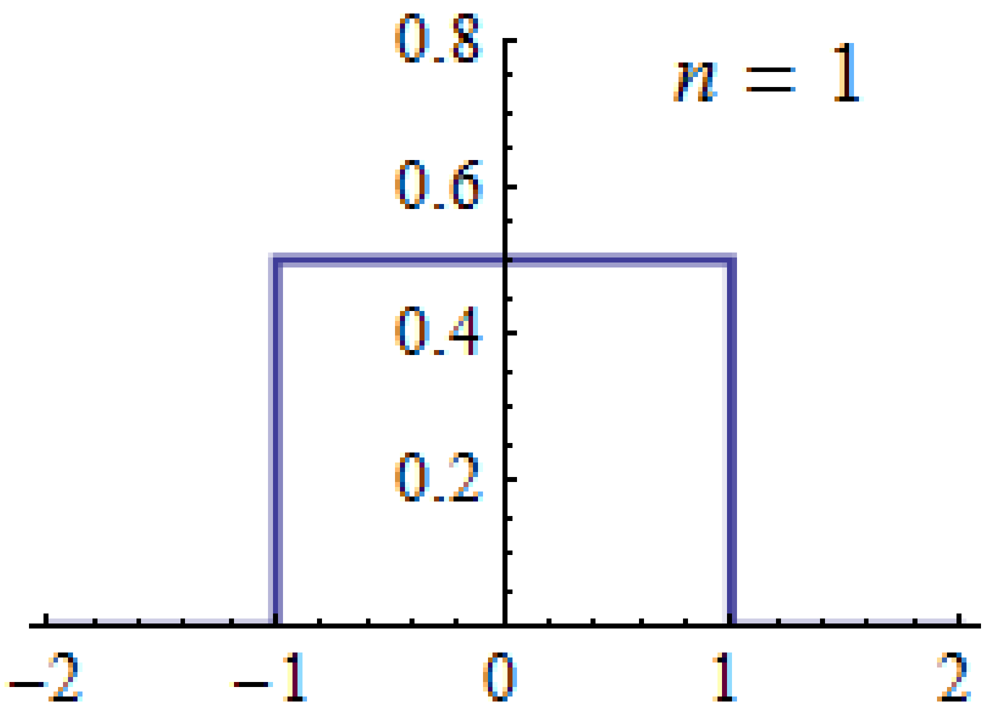
$$\Rightarrow \sqrt{n}\left(\frac{X_1 + \ldots + X_n}{n} - \mu\right) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

Lyapunov CLT:

$E[X_i] = \mu_i$, $Var[X_i] = \sigma_i^2$, $s_n^2 = \sum_{i=1}^{n} \sigma_i^2$.

+ some other conditions

$$\Rightarrow \frac{1}{s_n}\left(\sum_{i=1}^{n} X_i - \mu_i\right) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$



Generalizations: multi dim, time processes

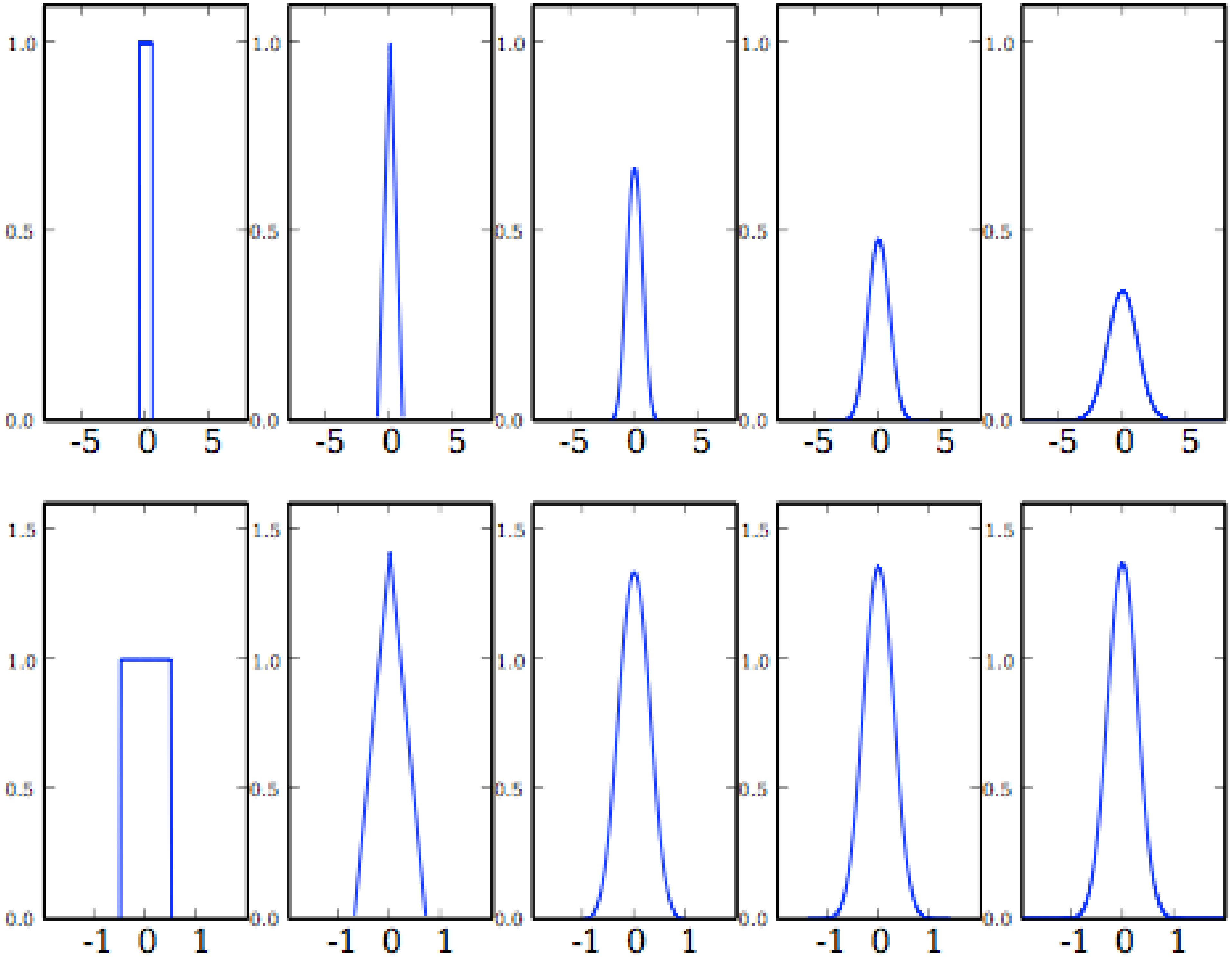
Central Limit Theorem in Practice

unscaled

$$\sum_{i=1}^{n} X_i$$

scaled

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$$



Proof of CLT

Let $\mathbb{E}[Y] = 0$, and $Var(Y) = 1$. From Taylor series around 0:

$$\exp(ity) = 1 + ity + \frac{i^2}{2}t^2y^2 + o(|t|^2)$$

$$\Rightarrow \varphi_Y(t) = \mathbb{E}[\exp(itY)] = 1 - \frac{t^2}{2} + o(t^2), \quad t \to 0$$

Let $Y_i = \frac{X_i - \mu}{\sigma}$ and let $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu_i}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$ $\quad \mathbb{E}[Y_i] = 0$, $Var(Y_i) = 1$

Properties of characteristic functions :

$\varphi_{\frac{1}{\sqrt{n}}Z}(t) = \varphi_Z(\frac{t}{\sqrt{n}})$ and $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$ if $X \perp\!\!\!\perp Y$.

$$\Rightarrow \varphi_{Z_n}(t) = \prod_{i=1}^{n} \varphi_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n \to e^{-t^2/2}, \quad n \to \infty$$

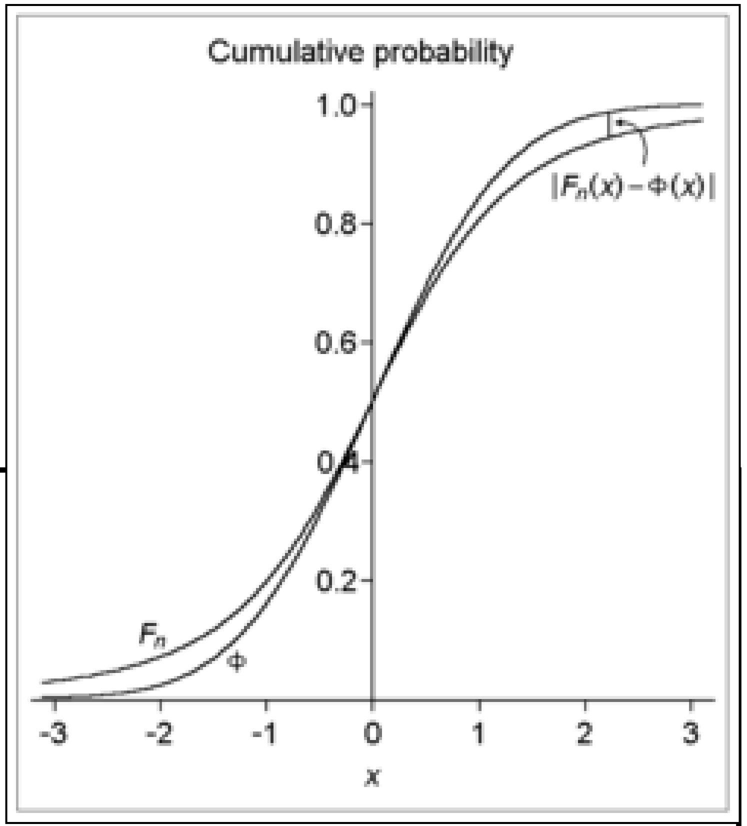
characteristic function of Gauss distribution

Levi's continuity theorem + uniqueness $\Rightarrow$ CLT

How fast do we converge to Gauss distribution?

CLT: $\sqrt{n}\left(\frac{X_1+\ldots+X_n}{n} - \mu\right) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$

It doesn't tell us anything about the convergence rate.



Berry-Esseen Theorem

Let $X_1, \ldots, X_n$ be i.i.d.

$\mathbb{E}[X_1] = \mu,\ \mathbb{E}[X_1^2] = \sigma^2\ \mathbb{E}[|X_1|^3] = \rho < \infty$

Let $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu_i}{\sigma}$

F_n is the cdf of Z_n $\qquad \Phi(x)$ is the cdf of $\mathcal{N}(0,1)$.

Then $\exists C > 0$ such that for all x and n, $|F_n(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3 \sqrt{n}}$.

Independently discovered by A. C. Berry (in 1941) and C.-G. Esseen (1942)

Did we answer the questions we asked?

- Do empirical averages converge?
- What do we mean on convergence?
- What is the rate of convergence?
- What is the limit distrib. of “standardized” averages?

Next time we will continue with these questions:

- ❑ How good are the ML algorithms on unknown test sets?
- ❑ How many training samples do we need to achieve small error?
- ❑ What is the smallest possible error we can achieve?

Further Readings on CLT

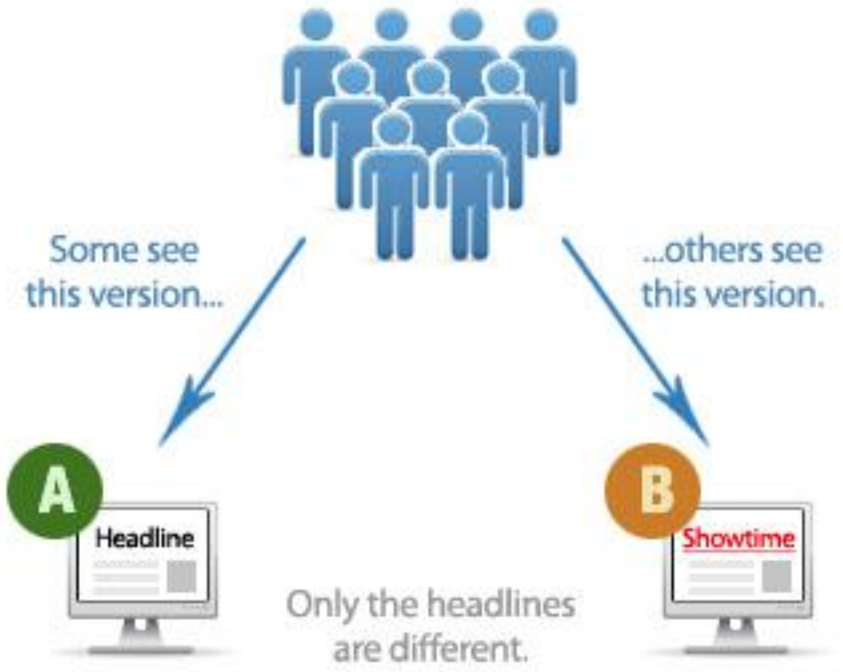
- http://en.wikipedia.org/wiki/Central_limit_theorem
- http://en.wikipedia.org/wiki/Law_of_the_iterated_logarithm

Tail bounds in practice

A/B testing

- Two possible webpage layouts
- Which layout is better?

Experiment
- Some users see A
- The others see design B



How many trials do we need to decide which page attracts more clicks?

A/B testing

Let us simplify this question a bit:

Assume that in group A

p(click|A) = 0.10 click and p(noclick|A) = 0.90

Assume that in group B

p(click|B) = 0.11 click and p(noclick|B) = 0.89

Assume also that we *know* these probabilities in group A, but we *don't know* yet them in group B.

We want to estimate p(click|B) with less than 0.01 error

Chebyshev Inequality

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad X_i = \begin{cases} 1 & \text{click} \\ 0 & \text{no click} \end{cases}$$

Chebyshev: $$\Pr(|\hat{\mu}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

- In group B the click probability is $\mu = 0.11$ (we don't know this yet)
- Want failure probability of δ=5%

• If we have no prior knowledge, we can only bound the variance by σ^2 = 0.25 (Uniform distribution hast the largest variance 0.25)

$$\Pr(|\hat{\mu}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} < \delta \Rightarrow \frac{\sigma^2}{\delta\varepsilon^2} < n \Rightarrow \frac{0.25}{0.05 \cdot 0.01^2} = 50{,}000 < n$$

• If we know that the click probability is < 0.15, then we can bound σ^2 at 0.15 * 0.85 = 0.1275. This requires at least 25,500 users.

Hoeffding's bound

- Hoeffding Let $c^2 = \frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2$

$$\Rightarrow \Pr(|\hat{\mu}_n - \mu| > \varepsilon) \le 2\exp\left(\frac{-2n\varepsilon^2}{c^2}\right)$$

- Random variable has bounded range [0, 1] (click or no click), hence c=1
- Solve Hoeffding's inequality for *n:*

$$2\exp\left(\frac{-2n\varepsilon^2}{c^2}\right) \le \delta \quad \Rightarrow \left(\frac{-2n\varepsilon^2}{c^2}\right) \le \log(\delta/2) \Rightarrow -2n\varepsilon^2 \le c^2\log(\delta/2)$$

$$\Rightarrow n > \frac{c^2\log(2/\delta)}{2\varepsilon^2} = 1 \cdot \frac{\log(2/0.05)}{2 \cdot 0.01^2} = 18,445$$

This is better than Chebyshev.

What we have learned today

Theory:

- Stochastic Convergences:
  - Weak convergence = Convergence in distribution
  - Convergence in probability
  - Strong (almost surely)
  - Convergence in L_p norm

- Limit theorems:
  - Law of large numbers
  - Central limit theorem

- Tail bounds:
  - Markov, Chebyshev
  - Hoeffding, Benett, McDiarmid

Thanks for your attention ☺