CSCI B609:
"Foundations of Data Science"

Lecture 1 & 2: Intro

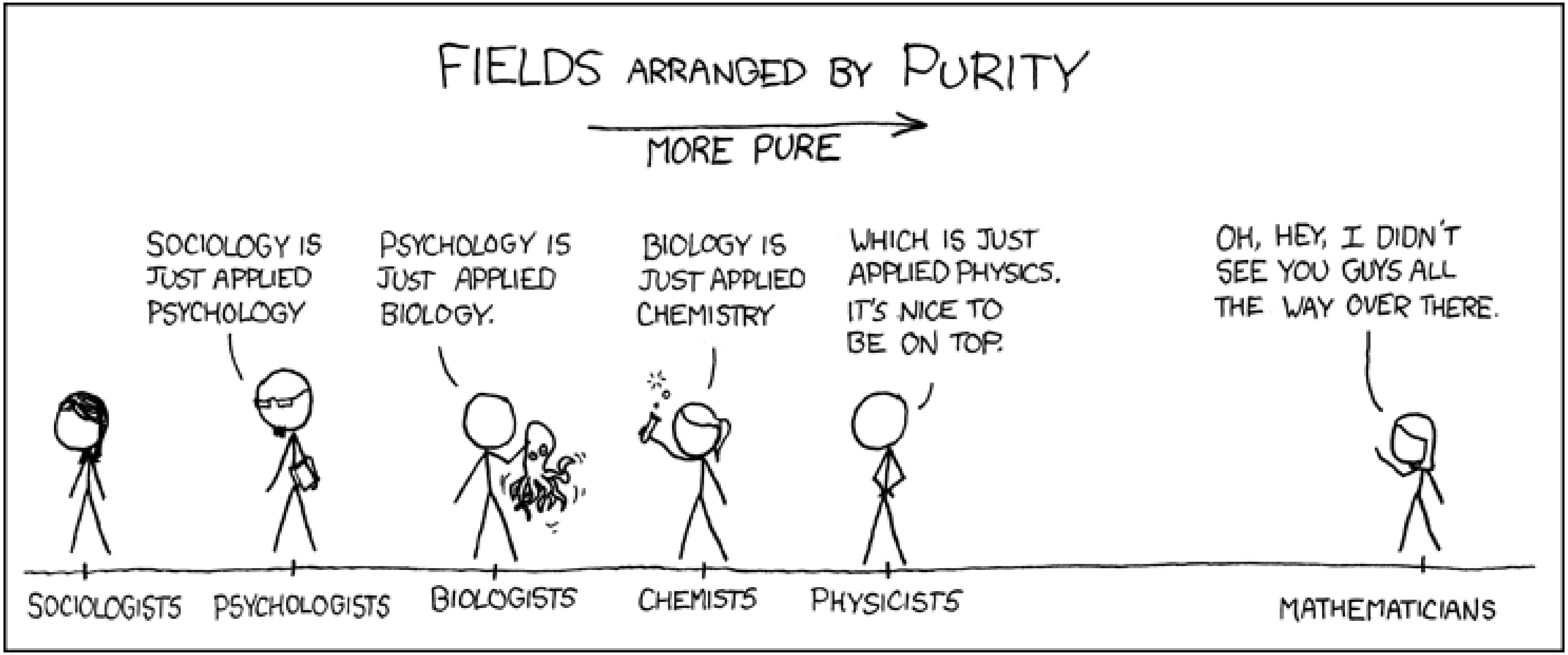
Slides at http://grigory.us/data-science-class.html

Grigory Yaroslavtsev

http://grigory.us

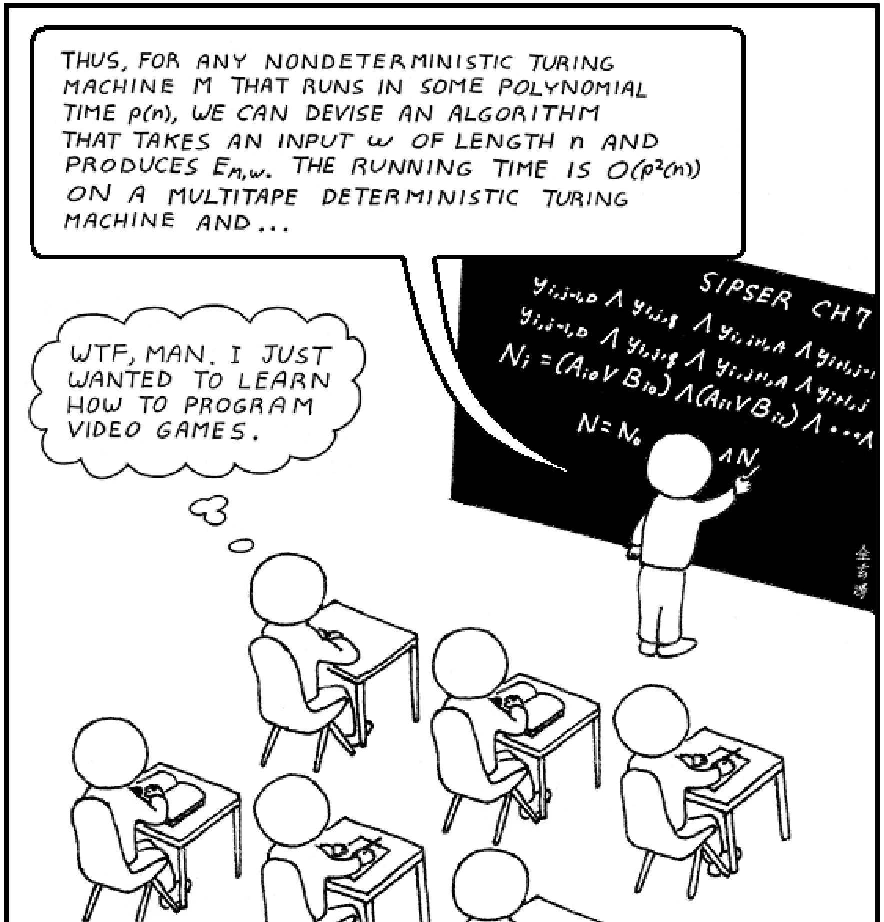
Disclaimers

- A lot of Math!



Disclaimers

- No programming!



Class info

- **Advanced graduate class, not an intro-level class**
- Primary audience: Ph.D. students
- MW 16:00 – 17:15, Ballantine 310
- Grading:
  - Class attendance/participation (20%)
  - Homework assignments (40%)
    - **Only accepted via e-mail in LaTeX-generated PDF format**
    - No handwritten homework accepted
  - Project (40%)
- Text: Blum-Hopcroft-Kannan, “Foundations of Data Science”
  - http://grigory.us/files/bhk-book.pdf
  - 06/09/16 version
- Office hours announced later
- Slides will be posted

Plan for today

- Lecture: first 45 minutes:
  - Basic probability
  - Inequalities for random variables
  - Concentration bounds
- Quiz: last 20 minutes:
  - Tests background knowledge
  - Graded but doesn’t count towards final grade
  - Quiz too hard => take intro-level classes first

Expectation

- $\boldsymbol{X}$ = random variable with values $x_1, \dots, x_n, \dots$
- If $\boldsymbol{X}$ is continuous then **all sums replaced with integrals**
- Expectation $\mathbb{E}[\boldsymbol{X}]$

$$\mathbb{E}[\boldsymbol{X}] = \sum_{i=1}^{\infty} \mathrm{x_i} \cdot \Pr[\boldsymbol{X} = x_i]$$

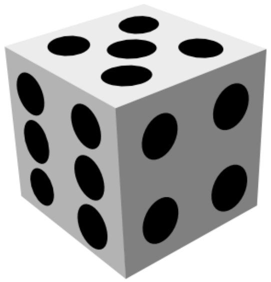
- Properties (linearity):

$$\mathbb{E}[c\boldsymbol{X}] = c\mathbb{E}[\boldsymbol{X}]$$
$$\mathbb{E}[\boldsymbol{X} + \boldsymbol{Y}] = \mathbb{E}[\boldsymbol{X}] + \mathbb{E}[\boldsymbol{Y}]$$

- Useful fact: if all $x_i \geq 0$ and integer then

$$\mathbb{E}[\boldsymbol{X}] = \sum_{i=1}^{\infty} \Pr[\boldsymbol{X} \geq i]$$

Expectation



- Example: dice has values $1, 2, \ldots, 6$ with probability 1/6

$$\mathbb{E}[\text{Value}] = \sum_{i=1}^{6} i \cdot \Pr[Value = i]$$

$$= \frac{1}{6} \sum_{i=1}^{6} i = \frac{21}{6} = 3.5$$

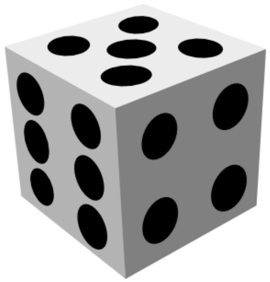
Variance

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

$$Var[\boldsymbol{X}] = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2] =$$
$$= \mathbb{E}[\boldsymbol{X}^2 \ - 2\,\mathbf{X} \cdot \mathbb{E}[\mathbf{X}] + \mathbb{E}[\mathbf{X}]^2]$$
$$= \mathbb{E}[\boldsymbol{X}^2] \ - 2\mathbb{E}[\mathbf{X} \cdot \mathbb{E}[\mathbf{X}]] + \mathbb{E}[\mathbb{E}[\mathbf{X}]^2]$$

- $\mathbb{E}[\mathrm{X}]$ is some fixed value (a constant)
- $2\ \mathbb{E}[\mathbf{X} \cdot \mathbb{E}[\mathbf{X}]] = 2\ \mathbb{E}[\mathbf{X}] \cdot \mathbb{E}[\mathbf{X}] = 2\ \mathbb{E}^2[\boldsymbol{X}]$
- $\mathbb{E}[\mathbb{E}[\mathbf{X}]^2] = \ \mathbb{E}^2[\mathbf{X}]$
- $Var[\boldsymbol{X}] = \mathbb{E}[\boldsymbol{X}^2] - 2\ \mathbb{E}^2[\boldsymbol{X}] + \ \mathbb{E}^2[\mathbf{X}] = \mathbb{E}[\boldsymbol{X}^2] \ - \ \mathbb{E}^2[\mathbf{X}]$
- Corollary: $Var[c\boldsymbol{X}] \ = \ c^2 Var[\boldsymbol{X}]$

Variance



- Example (Variance of a fair dice):

$$\mathbb{E}[Value] = 3.5$$

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

$$=\mathbb{E}[(Value\ - 3.5)^2]$$

$$= \sum_{i=1}^{6}(i\ - 3.5)^2 \cdot Pr\,[Value = i]$$

$$= \frac{1}{6}\sum_{i=1}^{6}(i\ - 3.5)^2$$

$$= \frac{1}{6}\,[\left(1 - 3.5\right)^2 + \left(2 - 3.5\right)^2 + \left(3 - 3.5\right)^2$$

$$+(4 - 3.5)^2 + (5 - 3.5)^2 + (6 - 3.5)^2]$$

$$= \frac{1}{6}[6.25\ + 2.25\ + 0.25\ + 0.25\ + 2.25 + 6.25\ ]$$

$$= \frac{8.75}{3} \approx 2.917$$

Independence

- Two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are **independent** if and only if (iff) for every x, y:

$$\Pr[\boldsymbol{X} = x, \boldsymbol{Y} = y] = \Pr[\boldsymbol{X} = x] \cdot \Pr[\boldsymbol{Y} = y]$$

- Variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are **mutually independent** iff

$$\Pr[\boldsymbol{X}_\mathbf{1} = x_1, \ldots, \boldsymbol{X}_n = x_n] = \prod_{i=1}^{n} \Pr[\boldsymbol{X}_{\boldsymbol{i}} = x_i]$$

- Variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are **pairwise independent** iff for all pairs i,j

$$\Pr\left[\boldsymbol{X}_{\boldsymbol{i}} = x_i, \boldsymbol{X}_j = x_j\right] = \Pr[\boldsymbol{X}_{\boldsymbol{i}} = x_i] \Pr\left[\boldsymbol{X}_{\boldsymbol{j}} = x_j\right]$$

Independence: Example

- Ratings of mortgage securities
  - AAA = 1% probability of default (over X years)
  - AA = 2% probability of default
  - A = 5% probability of default
  - B = 10% probability of default
  - C = 50% probability of default
  - D = 100% probability of default
- You are a portfolio holder with 1000 AAA securities?
  - Are they all independent?
  - Is probability of all defaulting $(0.01)^{1000} = 10^{-2000}$?

Conditional Probabilities

- For two events E_1 and E_2:

$$\Pr[E_2|E_1] = \frac{\Pr[E_1 \ and \ E_2]}{\Pr[E_1]}$$

- If two random variables (r.vs) are independent

$$\Pr[X_2 = x_2|X_1 = x_1]$$

$$= \frac{\Pr[X_1{=}x_1 \ and \ X_2{=}x_2]}{\Pr[X_1{=}x_1]} \text{ (by definition)}$$

$$= \frac{\Pr[X_1{=}x_1]Pr[X_2{=}x_2]}{\Pr[X_1{=}x_1]} \text{ (by independence)}$$

$$= \Pr[X_2 = x_2]$$

Union Bound

For any events $E_1, \ldots, E_k$:

$$\Pr[E_1 or\ E_2\ or\ \ldots or\ E_k]$$
$$\leq \Pr[E_1] + \Pr[E_2] + \ldots + \Pr[E_k]$$

- **Pro**: Works even for dependent variables!
- **Con**: Sometimes very loose, especially for **mutually** independent events
  $\Pr[E_1\ or\ E_2\ or\ \ldots or\ E_k] = 1\ - \prod_{i=1}^{k}(1\ - \Pr[E_i])$

Independence and Linearity of Expectation/Variance

- Linearity of expectation (even for dependent variables!):

$$\mathbb{E}\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} \mathbb{E}[X_i]$$

- Linearity of variance (only for **pairwise independent** variables!)

$$Var\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} Var[X_i]$$

Part 2: Inequalities

- Markov inequality
- Chebyshev inequality
- Chernoff bound

Markov's Inequality

- If $\boldsymbol{X}$ is a non-negative r.v. then for every $c > 0$:

$$\Pr\left[\boldsymbol{X} \geq c\ \mathbb{E}[\boldsymbol{X}]\right] \leq \frac{1}{c}$$

- **Proof**

$$\mathbb{E}[\boldsymbol{X}] = \sum_i i \cdot \Pr[\boldsymbol{X} = i] \qquad \text{(by definition)}$$

$$\geq \sum_{i=c\mathbb{E}[\boldsymbol{X}]}^{\infty} i \cdot \Pr[\boldsymbol{X} = i] \qquad \text{(pick only some i's)}$$

$$\geq \sum_{i=c\mathbb{E}[\boldsymbol{X}]}^{\infty} c\mathbb{E}[\boldsymbol{X}] \cdot \Pr[\boldsymbol{X} = i] \qquad (i \geq c\mathbb{E}[\boldsymbol{X}])$$

$$= c\mathbb{E}[\boldsymbol{X}] \sum_{i=c\mathbb{E}[\boldsymbol{X}]}^{\infty} \Pr[\boldsymbol{X} = i] \qquad \text{(by linearity)}$$

$$= c\mathbb{E}[\boldsymbol{X}] \Pr\left[\boldsymbol{X} \geq c\ \mathbb{E}[\boldsymbol{X}]\right] \qquad \text{(same as above)}$$

$$\Rightarrow \Pr\left[\boldsymbol{X} \geq c\ \mathbb{E}[\boldsymbol{X}]\right] \leq \frac{1}{c}$$

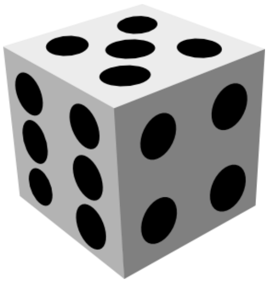
Markov's Inequality

- For every $c > 0$: $\Pr\left[\boldsymbol{X} \geq c\ \mathbb{E}[\boldsymbol{X}]\right] \leq \frac{1}{c}$
- **Corollary** $(\mathrm{c}' = c\ \mathbb{E}[\boldsymbol{X}])$ **:**

For every $c' > 0$: $\Pr[\boldsymbol{X} \geq c'\ ] \leq \frac{\mathbb{E}[\boldsymbol{X}]}{c'}$

- **Pro**: always works!
- **Cons**:
  - Not very precise
  - Doesn't work for the lower tail: $\Pr\left[\boldsymbol{X} \leq c\ \mathbb{E}[\boldsymbol{X}]\right]$

Markov Inequality: Example



Markov 1: For every $c > 0$:

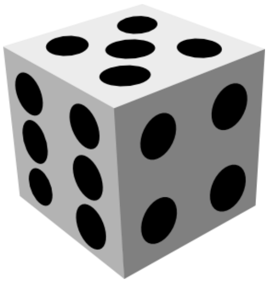
$$\Pr\left[\boldsymbol{X} \geq c\ \mathbb{E}[\boldsymbol{X}]\right] \leq \frac{1}{c}$$

- Example:

$$\Pr\left[Value \geq 1.5 \cdot\ \mathbb{E}[Value]\right] = \Pr[Value \geq 1.5 \cdot 3.5] = \Pr[Value \geq 5.25] \leq \frac{1}{1.5} = \frac{2}{3}$$

$$\Pr\left[Value \geq 2 \cdot\ \mathbb{E}[Value]\right] = \Pr[Value \geq 2 \cdot 3.5]$$
$$= \Pr[Value \geq 7] \leq \frac{1}{2}$$

Markov Inequality: Example



Markov 2: For every $c > 0$:

$$\Pr[\boldsymbol{X} \geq c\ ] \leq \frac{\mathbb{E}[\boldsymbol{X}]}{c}$$

- Example:

$$\Pr[Value \geq 4] \leq \frac{\mathbb{E}[Value]}{4} = \frac{3.5}{4} = 0.875 \ (= \ \mathbf{0.5})$$

$$\Pr[Value \geq 5] \leq \frac{\mathbb{E}[Value]}{5} = \frac{3.5}{5} = 0.7 \quad (\approx \mathbf{0.33})$$

$$\Pr[Value \geq 6] \leq \frac{\mathbb{E}[Value]}{6} = \frac{3.5}{6} \approx 0.58 \quad (\approx \mathbf{0.17})$$

$$\Pr[Value \geq 3] \leq \frac{\mathbb{E}[Value]}{3} = \frac{3.5}{3} \approx 1.17 \quad (\approx \mathbf{0.66})$$

Quiz analysis: P1, part 1

x, y are independent variables with uniform distribution over [0,1]

- $\mathbb{E}[\mathrm{x}] = ½$
- $\mathbb{E}[x^2] = \int_0^1 x^2 dx = \frac{1}{3}$
- $\mathbb{E}[x - y] = \mathbb{E}[x] - \mathbb{E}[y] = 1/2 - 1/2 = 0$
- $\mathbb{E}[xy] = \mathbb{E}[x]\ \mathbb{E}[y] = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
- $\mathbb{E}[(x - y)^2] = \mathbb{E}[x^2] - 2\mathbb{E}[xy] + \mathbb{E}[y^2]$

$$= \frac{1}{3} - 2 \times \frac{1}{4} + \frac{1}{3} = 1/6$$

Quiz analysis: P1, part 2

- What is the expected squared distance between two points generated uniformly at random inside a d-dimensional hypercube $[0,1]^d$?
- $\mathbb{E}\left[\sum_{i=1}^{d}(x_i - y_i)^2\right] = d \times \mathbb{E}[(x_i - y_i)^2] = \frac{d}{6}$

Quiz analysis: P2

- For fixed $a \geq 1$ show an example when Markov's inequality is tight, i.e.
$$\Pr[\mathrm{X} \geq a] = \frac{\mathbb{E}[X]}{a}$$
- Example: $X = a$ (with probability 1)
- $\mathbb{E}[X] = a, \Pr[\mathrm{X} \geq a] = \frac{\mathbb{E}[X]}{a} = 1$

Quiz analysis: P3

- What is the variance of the first coordinate x_1 of a vector x drawn from a uniform distribution over a unit **d**-dimensional sphere (set of points such that $||x||_2 = 1$)?
- Var$[x_1]$ = $\mathbb{E}[x_1^2] - \mathbb{E}^2[x_1]$
- $\mathbb{E}[x_1] = 0$ (by symmetry)
- $\mathbb{E}[x_1^2] = \frac{1}{d}\mathbb{E}[\sum_{i=1}^{d} x_i^2] = \frac{1}{d}$

Quiz analysis: P4

- Sort a sequence of integers in $O(n^2)$ time
  - Expected solution: Bubblesort, Insertionsort, etc.
- Sort a sequence of integers in $O(n \log n)$ time
  - Expected solution: Quicksort (in expectation), Mergesort (worst-case)

Core Classes to Take

- **B503 (Algorithms), MW + TR**
- **B551 (Elements of Artificial Intelligence), TR**
- **B555 (Machine Learning), MW, this time**
- **B561 (Databases), MW + TR**
- **B565 (Data Mining), TR**

Chebyshev's Inequality

- For every $c > 0$:

$$\Pr\left[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c\sqrt{Var[\boldsymbol{X}]}\right] \leq \frac{1}{c^2}$$

- Proof:

$$\Pr\left[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c\sqrt{Var[\boldsymbol{X}]}\right]$$

$= \Pr[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]|^2 \geq c^2 Var[\boldsymbol{X}]]$ (by squaring)

$= \Pr[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]|^2 \geq c^2 \mathbb{E}[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]|^2]]$ (def. of Var)

$\leq \frac{1}{c^2}$ (by Markov's inequality)

Chebyshev's Inequality

- For every $c > 0$:

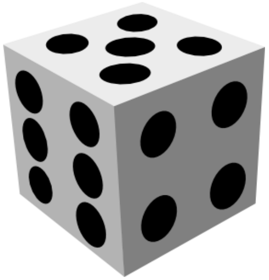
$$\Pr\left[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c\sqrt{Var[\boldsymbol{X}]}\right] \leq \frac{1}{c^2}$$

- **Corollary** ($c' = c\sqrt{Var[\boldsymbol{X}]}$):

For every $c' > 0$:

$$\Pr[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c'] \leq \frac{Var[\boldsymbol{X}]}{c'^2}$$

Chebyshev: Example



- For every $c' > 0$: $\Pr[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c'] \leq \frac{Var[\boldsymbol{X}]}{c'^2}$

$$\mathbb{E}[Value] = 3.5; \;\; Var\,[Value] \approx 2.91$$

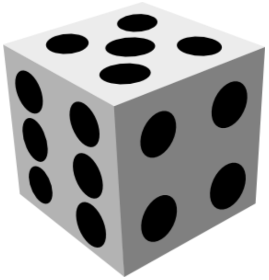
$$\Pr[Value \geq 4 \; or \; Value \leq 3] =$$

$$Pr[|Value - 3.5| \geq 0.5] \leq \frac{2.91}{0.5^2} \approx 11.64 \; (= \mathbf{1})$$

$$\Pr[Value \geq 5 \; or \; Value \leq 2] \leq \frac{2.91}{1.5^2} \approx 1.29 \quad (\approx \mathbf{0.66})$$

$$\Pr[Value \geq 6 \; or \; Value \leq 1] \leq \frac{2.91}{2.5^2} \approx 0.47 \quad (\approx \mathbf{0.33})$$

Chebyshev: Example



- Roll a dice 10 times:

$Value_{10}$ = Average value over 10 rolls

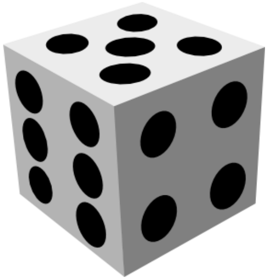
$$\Pr[Value_{10} \geq 4 \; or \; Value_{10} \leq 3] = ?$$

- X_i = value of the i-th roll, $\boldsymbol{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$
- Variance ($=$ by linearity for **independent** r.vs):

$$Var[\boldsymbol{X}] = Var\left[\frac{1}{10}\sum_{i=1}^{10} X_i\right] = \frac{1}{100} Var\left[\sum_{i=1}^{10} X_i\right]$$

$$= \frac{1}{100}\sum_{i=1}^{10} Var[X_i] \approx \frac{1}{100} \cdot 10 \cdot 2.91 = 0.291$$

Chebyshev: Example



- Roll a dice 10 times:

$$Value_{10} = \text{Average value over 10 rolls}$$
$$\Pr[Value_{10} \geq 4 \; or \; Value_{10} \leq 3] = ?$$

- $Var[Value_{10}] = 0.291$ (if n rolls then 2.91 / n)
- $\Pr[Value_{10} \geq 4 \; or \; Value_{10} \leq 3] \leq \frac{0.291}{0.5^2} \approx 1.16$
- $\Pr[Value_{n} \geq 4 \; or \; Value_{n} \leq 3] \leq \frac{2.91}{n \cdot 0.5^2} \approx \frac{11.6}{n}$

Chernoff bound

- Let $X_1 \ldots X_{\boldsymbol{t}}$ be independent and identically distributed r.vs with range [0,1] and expectation μ.
- Then if $X = \frac{1}{t}\sum_i X_i$ and $1 > \delta > 0$,

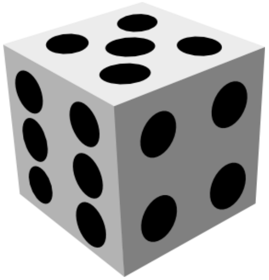
$$\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\mu \boldsymbol{t} \delta^2}{3}\right)$$

Chernoff bound (corollary)

- Let $X_1 \ldots X_t$ be independent and identically distributed r.vs with range [0, **c**] and expectation μ.
- Then if $X = \frac{1}{t}\sum_i X_i$ and $1 > \delta > 0$,

$$\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\mu t \delta^2}{3\boldsymbol{c}}\right)$$

Chernoff: Example

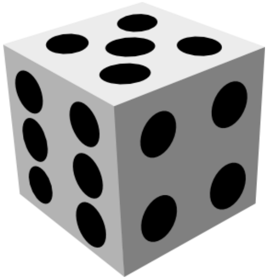


- $\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\mu t \delta^2}{3c}\right)$
- Roll a dice 10 times:

$$Value_{10} = \text{Average value over 10 rolls}$$
$$\Pr[Value_{10} \geq 4 \; or \; Value_{10} \leq 3] = ?$$

  - $X = Value_{10}, \; t = 10, \; \mathrm{c} = 6$
  - $\mu = \mathbb{E}[X_i] = 3.5$
  - $\delta = \frac{0.5}{3.5} = \frac{1}{7}$
- $\Pr[Value_{10} \geq 4 \; or \; Value_{10} \leq 3] \leq 2\exp\left(-\frac{3.5 \cdot 10}{3 \cdot 6 \cdot 49}\right) =$
  $2\exp\left(-\frac{35}{882}\right) \approx 2 \cdot 0.96 = 1.92$

Chernoff: Example



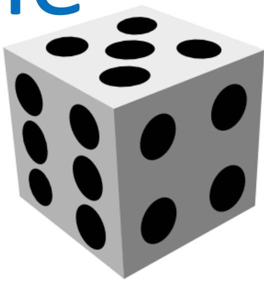
- $\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\mu t \delta^2}{3c}\right)$
- Roll a dice 1000 times:

  $Value_{1000}$ = Average value over 1000 rolls

  $\Pr[Value_{1000} \geq 4 \; or \; Value_{1000} \leq 3] = ?$

  - $X = Value_{1000}, \; t = 1000, \; \mathrm{c} = 6$
  - $\mu = \mathbb{E}[X_i] = 3.5$
  - $\delta = \frac{0.5}{3.5} = \frac{1}{7}$
- $\Pr[Value_{10} \geq 4 \; or \; Value_{10} \leq 3] \leq$
  $2\exp\left(-\frac{3.5 \cdot 1000}{3 \cdot 6 \cdot 49}\right) = 2\exp\left(-\frac{3500}{882}\right) \approx$
  $2 \cdot \exp(-3.96) \approx 2 \cdot 0.02 = 0.04$

Chernoff v.s Chebyshev: Example



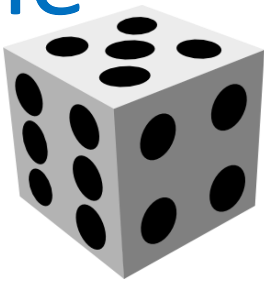
Let $\sigma = Var[X_i]$:

- Chebyshev: $\Pr[|\boldsymbol{X} - \mu| \geq c'] \leq \frac{Var[\boldsymbol{X}]}{c'^2} = \frac{\sigma}{\boldsymbol{t}\, \boldsymbol{c}'^2}$
- Chernoff: $\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\mu \boldsymbol{t} \delta^2}{3c}\right)$

If $\boldsymbol{t}$ is very big:

- Values $\mu, \sigma, \delta, c, c'$ are all constants!
  - Chebyshev: $\Pr[|\boldsymbol{X} - \mu| \geq z] = O\left(\frac{1}{\boldsymbol{t}}\right)$
  - Chernoff: $\Pr[|\boldsymbol{X} - \mu| \geq z] = e^{-\Omega(\boldsymbol{t})}$

Chernoff v.s Chebyshev: Example



Large values of t is exactly what we need!

- Chebyshev: $\Pr[|\boldsymbol{X} - \mu| \geq z] = O\left(\frac{1}{t}\right)$
- Chernoff: $\Pr[|\boldsymbol{X} - \mu| \geq z] = e^{-\Omega(t)}$

So is Chernoff always better for us?

- Yes, if we have i.i.d. variables.
- No, if we have dependent or only pairwise independent random varaibles.
- If the variables are not identical – Chernoff-type bounds exist.