CSE 312

Foundations of Computing II

Lecture 20: Tail Bounds -- Markov , Chebyshev, and Chernoff Bounds

Review: Joint PMFs and Joint Range

Definition. Let X and Y be discrete random variables. The **Joint PMF** of X and Y is

$$p_{X,Y}(a,b) = P(X=a, Y=b) \quad = P(X=a)\,P(\cancel{X}Y=b)$$

Definition. Let X and Y be discrete random variables and $p_{X,Y}(a,b)$ their joint PMF. The **marginal PMF** of X

$$p_X(a) = \sum_{b \in \Omega_Y} p_{X,Y}(a,b) \quad @\ Y=b.$$

$$= P(X=a \wedge Y=b)$$

Y=b

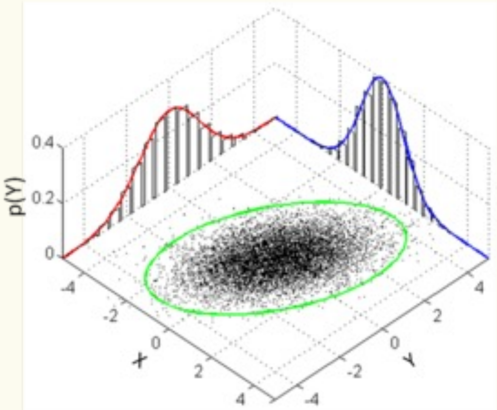
Review: Continuous distributions on $\mathbb{R}\times\mathbb{R}$

Definition. The **joint probability density function (PDF)** of continuous random variables X and Y is a function $f_{X,Y}$ defined on $\mathbb{R}\times\mathbb{R}$ such that

- $f_{X,Y}(x,y) \geq 0$ for all $x, y \in \mathbb{R}$
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}x\,\mathrm{d}y = 1$

The **(marginal) PDFs** f_X and f_Y are given by

- $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y$
- $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x$



Independence and joint distributions

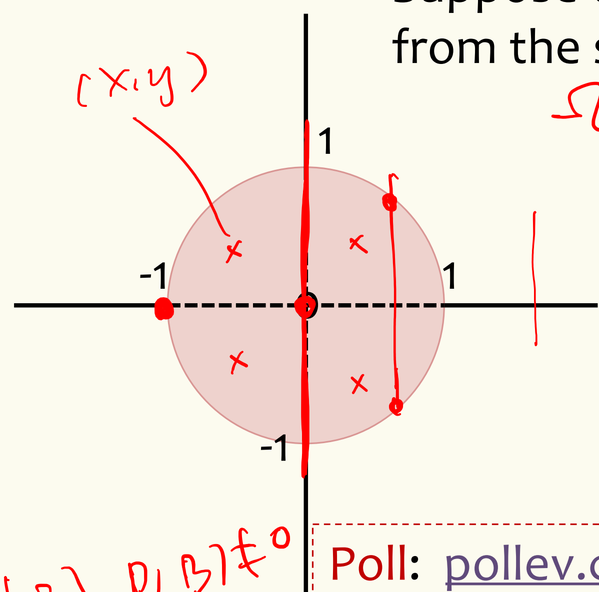
Definition. Discrete random variables X and Y are **independent** iff

- $p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y)$ for all $x \in \Omega_X, y \in \Omega_Y$

Definition. Continuous random variables X and Y are **independent** iff

- $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$ for all $x, y \in \mathbb{R}$

Example – Uniform distribution on a unit disk



Suppose that a pair of random variables (X, Y) is chosen uniformly from the set of real points (x, y) such that $x^2 + y^2 \leq 1$

$\Omega_{XY} = \{(x,y) \mid x^2+y^2 \leq 1\}$

This is a disk of radius 1 which has area π

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{\pi} & \text{if } x^2 + y^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$\neq f_X(x) \cdot f_Y(y)$

LOTP

$$f_X(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi} \mathrm{d}y$$

$$= 2\sqrt{1-x^2}/\pi$$

$P(A|B)\ P(B) \neq 0$

$= P(A)$

Poll: pollev.com/rachel312

Are X and Y independent?

12

a. Yes

b. No

Joint Expectation

Definition. Let X and Y be discrete random variables and $p_{X,Y}(a,b)$ their joint PMF. The **expectation** of some function $g(x,y)$ with inputs X and Y

$$\mathbb{E}[\boxed{g(X,Y)}] = \sum_{a \in \Omega_X} \sum_{b \in \Omega_Y} g(a,b) \cdot p_{X,Y}(a,b)$$

$\mathbb{E}[(X+Y)^2]$

Agenda

- Joint Distributions
  - Cartesian Products
  - Joint PMFs and Joint Range
  - Marginal Distribution
- Conditional Expectation and Law of Total Expectation
- Conditional expectation and LTE for continuous RVs

Conditional Expectation

Definition. Let X be a discrete random variable then the **conditional expectation** of X given event A is

$$\mathbb{E}[X \mid A] = \sum_{x \in \Omega_X} x \cdot \boxed{P(X = x \mid A)}$$

Notes:

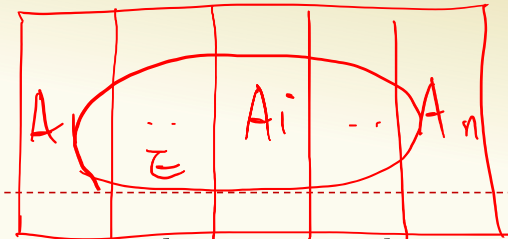
- Can be phrased as a "random variable version"

$$\mathbb{E}[X| Y = y]$$

- Linearity of expectation still applies here

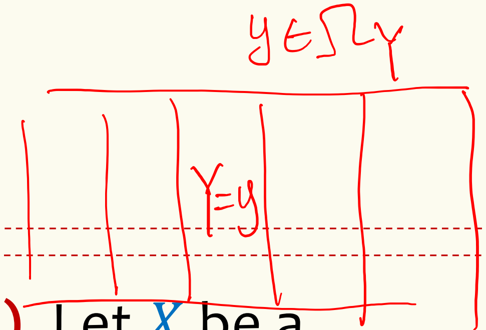
$$\mathbb{E}[aX + bY + c \mid A] = a\,\mathbb{E}[X \mid A] + b\,\mathbb{E}[Y \mid A] + c$$

Law of Total Expectation



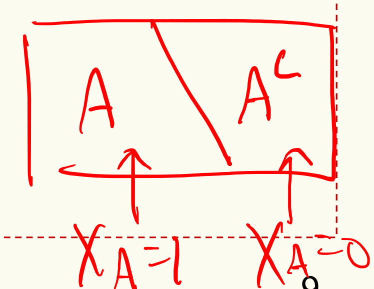
Law of Total Expectation (event version). Let X be a random variable and let events $A_1, \dots, A_n$ partition the sample space. Then,

$$\boxed{\mathbb{E}[X]} = \sum_{i=1}^{n} \mathbb{E}[X \mid A_i] \cdot P(A_i)$$



Law of Total Expectation (random variable version). Let X be a random variable and Y be a discrete random variable. Then,

$$\mathbb{E}[X] = \sum_{y \in \Omega_Y} \mathbb{E}[X \mid Y = y] \cdot P(Y = y)$$



Proof of Law of Total Expectation (not covered)

Follows from Law of Total Probability and manipulating sums

$$\mathbb{E}[X] = \sum_{x \in \Omega_X} x \cdot P(X = x)$$

$$= \sum_{x \in \Omega_X} x \cdot \sum_{i=1}^{n} P(X = x \mid A_i) \cdot P(A_i)$$ (by LTP)

$$= \sum_{i=1}^{n} P(A_i) \sum_{x \in \Omega_X} x \cdot P(X = x \mid A_i)$$ (change order of sums)

$$= \sum_{i=1}^{n} P(A_i) \cdot \mathbb{E}[X \mid A_i]$$ (def of cond. expect.)

Example – Flipping a Random Number of Coins

i

Suppose someone gave us $Y \sim \text{Poi}(5)$ fair coins and we wanted to compute the expected number of heads X from flipping those coins.

$\frac{i}{2}$

By the Law of Total Expectation

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} \mathbb{E}[X \mid \boxed{Y = i}] \cdot P(Y = i) = \sum_{i=0}^{\infty} \frac{i}{2} \cdot P(Y = i)$$

$$= \frac{1}{2} \cdot \sum_{i=0}^{\infty} i \cdot P(Y = i)$$

$$= \frac{1}{2} \cdot \mathbb{E}[Y] = \frac{1}{2} \cdot 5 = 2.5$$

Agenda

- Joint Distributions
  - Cartesian Products
  - Joint PMFs and Joint Range
  - Marginal Distribution
- Conditional Expectation and Law of Total Expectation
- Conditional expectation and LTE for continuous RVs

Conditional Expectation again…

Definition. Let X be a discrete random variable; then the **conditional expectation** of X given event A is

$$\mathbb{E}[X \mid A] = \sum_{x \in \Omega_X} x \cdot P(X = x \mid A)$$

Therefore for X and Y discrete random variables, the conditional expectation of X given $Y = y$ is

$$\mathbb{E}[X \mid Y = y] = \sum_{x \in \Omega_X} x \cdot P(X = x \mid Y = y) = \sum_{x \in \Omega_X} x \cdot p_{X|Y}(x|y)$$

where we **define** $p_{X|Y}(x|y) = P(X = x \mid Y = y) = \dfrac{p_{X,Y}(x, y)}{p_Y(y)}$

Conditional Expectation – Discrete & Continuous

Discrete: Conditional PMF: $p_{X|Y}(x|y) = \dfrac{p_{X,Y}(x,y)}{p_Y(y)}$

Conditional Expectation: $\mathbb{E}[X \mid Y = y] = \sum_{x \in \Omega_X} x \cdot p_{X|Y}(x|y)$

Continuous: Conditional PDF: $f_{X|Y}(x|y) = \dfrac{f_{X,Y}(x,y)}{f_Y(y)}$

Conditional Expectation: $\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y)\, dx$

Law of Total Expectation - continuous

Law of Total Expectation (event version). Let X be a random variable and let events $A_1, \dots, A_n$ partition the sample space. Then,

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X \mid A_i] \cdot P(A_i)$$

Law of Total Expectation (random variable version). Let X and Y be continuous random variables. Then,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} \mathbb{E}[X \mid Y = y] \cdot f_Y(y)\, dy$$

Using LTE for Continuous RVs

PDF for $Exp(\lambda)$ is $\begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{o.w.} \end{cases}$

Expectation is $1/\lambda$

$Y=1$ / $Y=2$

Suppose that we first choose $Y \sim Exp(1/2)$ and then choose $X \sim Exp(Y)$. What is $\mathbb{E}[X]$?

$X \sim Exp(1)$ $\sim Exp(2)$

$$f_{X|Y}(x|y) = \boxed{y\, e^{-xy}}$$

$\boxed{y \text{ is fixed here}}$

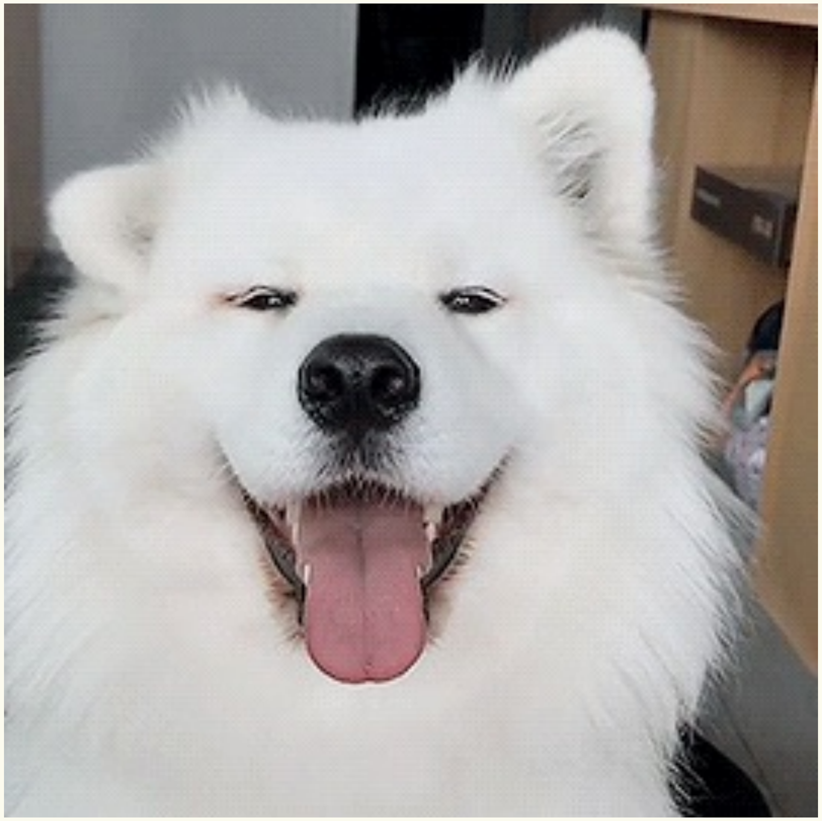
$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y)\, dx = \int_{-\infty}^{\infty} x \cdot y\, e^{-xy}\, dx = \frac{1}{y}$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} \mathbb{E}[X \mid Y = y]\, f_Y(y)\, dy = \int_{-\infty}^{\infty} \frac{1}{y} \cdot 2\, e^{-y/2} dx = 2$$

Reference Sheet (with continuous RVs)

	Discrete	Continuous
Joint PMF/PDF	$p_{X,Y}(x,y) = P(X=x, Y=y)$	$f_{X,Y}(x,y) \neq P(X=x, Y=y)$
Joint CDF	$F_{X,Y}(x,y) = \sum_{t \leq x} \sum_{s \leq y} p_{X,Y}(t,s)$	$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(t,s)\,ds\,dt$
Normalization	$\sum_{x} \sum_{y} p_{X,Y}(x,y) = 1$	$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1$
Marginal PMF/PDF	$p_X(x) = \sum_{y} p_{X,Y}(x,y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$
Expectation	$E[g(X,Y)] = \sum_{x} \sum_{y} g(x,y) p_{X,Y}(x,y)$	$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy$
Conditional PMF/PDF	$p_{X\mid Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$	$f_{X\mid Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
Conditional Expectation	$E[X \mid Y=y] = \sum_{x} x p_{X\mid Y}(x \mid y)$	$E[X \mid Y=y] = \int_{-\infty}^{\infty} x f_{X\mid Y}(x \mid y)\,dx$
Independence	$\forall x, y, p_{X,Y}(x,y) = p_X(x) p_Y(y)$	$\forall x, y, f_{X,Y}(x,y) = f_X(x) f_Y(y)$

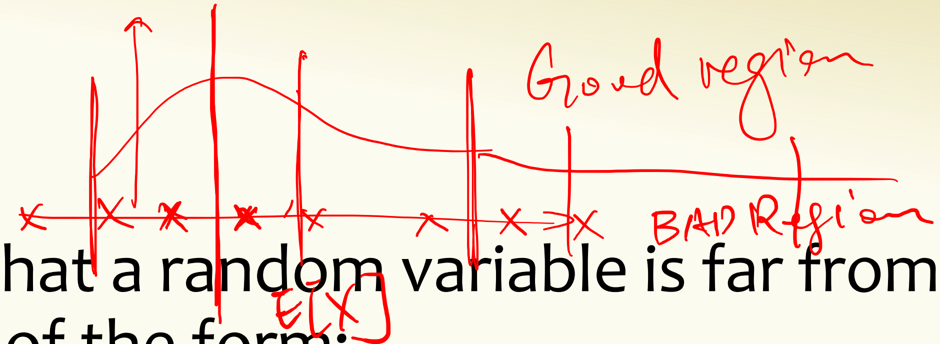
Brain Break



Agenda

- Markov's Inequality
- Chebyshev's Inequality
- Chernoff-Hoeffding Bound

Tail Bounds (Idea)



Bounding the probability that a random variable is far from its mean. Usually statements of the form:

$$P(X \geq a) \leq b$$
$$P(|X - \mathbb{E}[X]| \geq a) \leq b$$

Useful tool when

- An approximation that is easy to compute is sufficient
- The process is too complex to analyze exactly

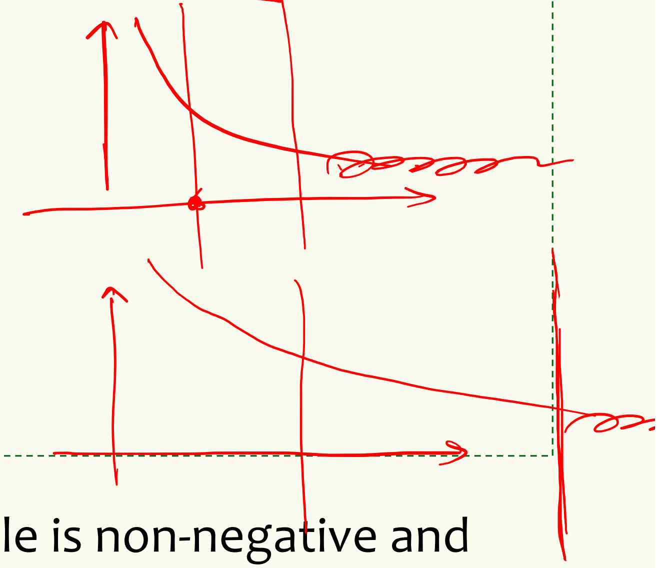
Markov's Inequality

Theorem. Let X be a random variable taking only *non-negative* values. Then, for any $t > 0$,

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

(Alternative form) For any $k \geq 1$,

$$P(X \geq k \cdot \mathbb{E}[X]) \leq \frac{1}{k}$$



Incredibly simplistic – only requires that the random variable is non-negative and only needs you to know <u>expectation</u>. You don't need to know **anything else** about the distribution of X.

Markov's Inequality – Proof I

Theorem. Let X be a (discrete) random variable taking only non-negative values. Then, for any $t > 0$,

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

$$\mathbb{E}[X] = \sum_x x \cdot P(X = x)$$

$$= \sum_{x \geq t} x \cdot P(X = x) + \sum_{x < t} x \cdot P(X = x)$$

≥ 0 because $x \geq 0$ whenever $P(X = x) \geq 0$ (X takes only non-negative values)

$$\geq \sum_{x \geq t} x \cdot P(X = x)$$

$$\geq \sum_{x \geq t} t \cdot P(X = x) = t \cdot P(X \geq t)$$

Follows by re-arranging terms

…

Markov's Inequality – Proof II

Theorem. Let X be a (continuous) random variable taking only non-negative values. Then, for any $t > 0$,

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

$$\mathbb{E}[X] = \int_0^\infty x \cdot f_X(x)\, \mathrm{d}x$$

$$= \underline{\int_t^\infty x \cdot f_X(x)\, \mathrm{d}x} + \underline{\int_0^t x \cdot f_X(x)\, \mathrm{d}x}$$

$$\geq \int_t^\infty x \cdot f_X(x)\, \mathrm{d}x$$

$$\geq \int_t^\infty t \cdot f_X(x)\, \mathrm{d}x = t \cdot \int_t^\infty f_X(x)\, \mathrm{d}x = t \cdot P(X \geq t)$$

so $P(X \geq t) \leq \mathbb{E}[X]/t$ as before

Example – Geometric Random Variable

Let X be ~~geometric~~ RV with parameter p

$P(X = i) = (1-p)^{i-1}p$

$\mathbb{E}[X] = \frac{1}{p}$

"X is the number of times Alice needs to flip a biased coin until she sees heads, if heads occurs with probability p?

What is the probability that $X \geq \underline{2\mathbb{E}[X] = 2/p}$?

Markov's inequality: $\underline{P(X \geq 2\mathbb{E}[X]) \leq \frac{1}{2}}$

Example

$P(X \geq k \cdot \mathbb{E}[X]) \leq \frac{1}{k}$

$E[X] = 25$

Suppose that the average number of ads you will see on a website is 25. Give an upper bound p on the probability of seeing a website with 75 or more ads.

$P[X \geq 75] \leq \frac{E[X]}{75}$

$= \frac{25}{75}$

$= \frac{1}{3}$

Poll: pollev.com/rachel312

a. $0 \leq p < 0.25$

b. $0.25 \leq p < 0.5$

c. $0.5 \leq p < 0.75$

d. $0.75 \leq p$

e. Unable to compute

$p = \frac{25}{75}$

Example

$$P(X \geq k \cdot \mathbb{E}[X]) \leq \frac{1}{k}$$

Suppose that the average number of ads you will see on a website is 25. Give an upper bound on the probability of seeing a website with 20 or more ads.

Poll: pollev.com/rachel312

a. $0 \leq p < 0.25$
b. $0.25 \leq p < 0.5$
c. $0.5 \leq p < 0.75$
d. $0.75 \leq p$
e. Unable to compute

Example – Geometric Random Variable

Let X be geometric RV with parameter p

$$P(X = i) = (1-p)^{i-1}p \qquad \mathbb{E}[X] = \frac{1}{p}$$

"X is **Next, we will see that we can get better tail bounds using variance** *e sees heads, if ... probability p?*

What is the probability that $X \geq 2\mathbb{E}[X] = 2/p$?

Markov's inequality: $P(X \geq 2\mathbb{E}[X]) \leq \frac{1}{2}$

Agenda

- Markov's Inequality
- Chebyshev's Inequality
- Chernoff-Hoeffding Bound

Chebyshev's Inequality

Theorem. Let X be a random variable. Then, for any $t > 0$,

$$P(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

Proof: Define $Z = X - \mathbb{E}[X]$. Then $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[Z^2]$.

$$P(|Z| \geq t) = P(Z^2 \geq t^2) \leq \frac{\mathbb{E}[Z^2]}{t^2} = \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{t^2} = \frac{\mathrm{Var}(X)}{t^2}$$

$|Z| \geq t$ iff $Z^2 \geq t^2$

Markov's inequality ($Z^2 \geq 0$)

Example – Geometric Random Variable

Let X be geometric RV *with* parameter p

$$P(X = i) = (1 - p)^{i-1} p \qquad \mathbb{E}[X] = \frac{1}{p} \qquad \mathrm{Var}(X) = \frac{1-p}{p^2}$$

What is the probability that $X \geq 2\mathbb{E}(X) = 2/p$*?*

Markov: $P(X \geq 2\mathbb{E}[X]) \leq \frac{1}{2}$

Chebyshev: $P(X \geq 2\mathbb{E}[X]) \leq P(|X - \mathbb{E}[X]| \geq \mathbb{E}[X]) \leq \frac{\mathrm{Var}(X)}{\mathbb{E}[X]^2} = 1 - p$

Better if $p > 1/2$ ☺

Example

$P(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}$.

Suppose that the average number of ads you will see on a website is 25 and the standard deviation of the number of ads is 4. Give an upper bound on the probability of seeing a website with 30 or more ads.

Poll: Where does that upper bound p lie?
pollev.com/rachel312

a. $0 \leq p < 0.25$
b. $0.25 \leq p < 0.5$
c. $0.5 \leq p < 0.75$
d. $0.75 \leq p$
e. Unable to compute

Chebyshev's Inequality – Repeated Experiments

"How many times does Alice need to flip a biased coin until she sees heads n times, if heads occurs with probability p?

X = # of flips until n times "heads"

X_i = # of flips between $(i-1)$-st and i-th "heads"

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

Note: $X_1, \dots, X_n$ are independent and geometric with parameter p

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{n}{p}$$

$$\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{n(1-p)}{p^2}$$

Chebyshev's Inequality – Coin Flips

"How many times does Alice need to flip a biased coin <u>until she sees heads n</u> times, if heads occurs with probability p?

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{n}{p} \qquad \mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{n(1-p)}{p^2}$$

What is the probability that $X \geq 2\mathbb{E}[X] = 2n/p$?

<u>Markov:</u> $P(X \geq 2\mathbb{E}[X]) \leq \frac{1}{2}$

<u>Chebyshev:</u> $P(X \geq 2\mathbb{E}[X]) \leq P(|X - \mathbb{E}[X]| \geq \mathbb{E}[X]) \leq \frac{\mathrm{Var}(X)}{\mathbb{E}[X]^2} = \frac{1-p}{n}$

Goes to zero as $n \to \infty$ ☺

Tail Bounds

Useful for approximations of complex systems. How good the approximation is depends on the actual distribution and the context you are using it in.

- Very often loose upper-bounds are okay when designing for the worst case

Generally (but not always) making more assumptions about your random variable leads to a more accurate upper-bound.