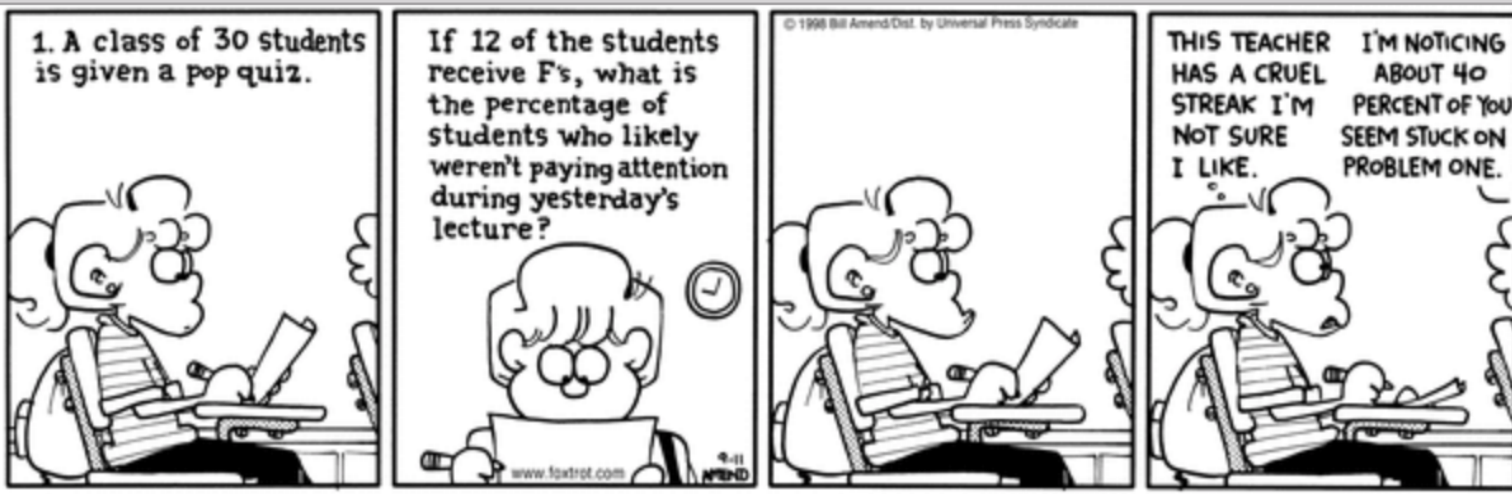
Stat 88: Probability & Mathematical Statistics in Data Science



Lecture 22: 3/12/2021

Sections 6.3, 6.4, 7.1

Markov and Chebyshev's Inequalities problems, Sums of RVs

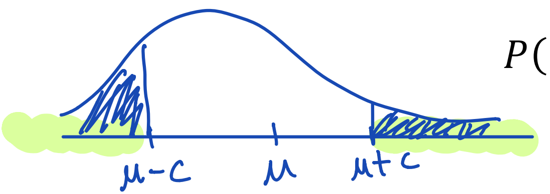
Bounding the tail probabilities

- **Markov's Inequality**: For a nonnegative rv X, and constant $c > 0$

$$P(X \geq c) \leq \frac{E(X)}{c}$$

If it is possible to try to use both Chebyshev & Markov's ineq. try both. Use the ineq that gives you a better bound.
Better = larger # for a lower bound & smaller # for an upperbound.

- **Chebyshev's inequality**: For a random variable X, with mean μ and standard deviation σ, for any positive constant $c > 0$, we have:



$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2} = \frac{Var(X)}{c^2} = \frac{\sigma^2}{c^2}$$

If $a, b \geq 0$
$a+b \leq c$ then $a \leq c$
$b \leq c$

- Ex: Is it possible that half of US flights have delay times at least 3 times the national average?

Non neg r.v, no SD given → Markov

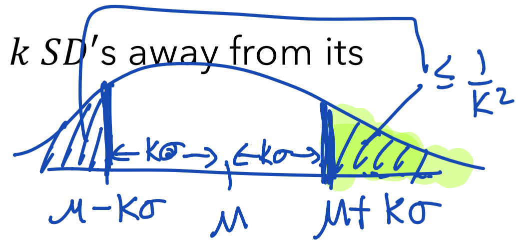
$$P(X \geq c) \leq \frac{\mu_x}{c} \longrightarrow P(X \geq 3\mu_x) \leq \frac{\cancel{\mu_x}}{3\cancel{\mu_x}} = \frac{1}{3}, \mu_x > 0$$

No! Not possible. $\left(\frac{1}{3} < \frac{1}{2}\right)$

Chebyshev's inequality interpreted as distances

- Say that $E(X)$ is the origin, and we are measuring distances in terms of $SD(X)$.
- We want to know the chance that the rv X is at least k SD's away from its mean:

$$P(|X-\mu| \geq k \cdot \sigma) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$



- What if we are only interested in one tail? A certain type of light bulb has an average lifetime of 10,000 hours. The SD of bulb lifetimes is 550 hours. What decimal fraction of bulbs could last more than 11,980 hours?

$\mu = 10{,}000$, $\sigma = 550$ $\quad$ X = lifetime of a randomly selected bulb

$$P(X \geq 11{,}980) = P(X - \mu \geq 11980 - \underbrace{10000}_{\mu})$$

$$= P(X-\mu \geq 1980) \leq \frac{\sigma^2}{c^2} = \frac{550^2}{1980^2} \approx 0.07$$

Using $k\sigma$, figure out k, $1980 = k \cdot 550$, so $k = 3.6$

$$P(X-\mu \geq 1980) \leq \frac{1}{(3.6)^2}, \quad P(X-\mu \geq 1980) \leq 0.07$$

Chebyshev or Markov?

$$P(X \geq c) \leq \frac{E(X)}{c}$$

- Suppose X is a non-negative random variable with expectation 60 and SD 5.

(a) What can we say about $P(X \geq 70)$?

$$P(X \geq 70) = P(X - \mu \geq 70 - 60)$$

$$P(X - \mu \geq 10) \leftarrow \text{Chebyshev}$$

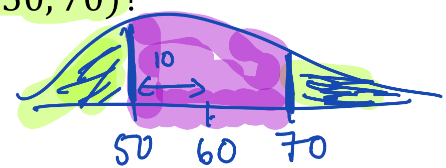
$\uparrow$ Markov $\quad P(X \geq 70) \leq \frac{60}{70} = \frac{6}{7}$

$$P(X - \mu \geq 10) \leq \frac{\sigma^2}{c^2} = \frac{25}{100} = \frac{1}{4}$$

(b) What is the chance that X is outside the interval $(50, 70)$?

$$P(|X - \mu| \geq \underbrace{10}_{k\sigma}) = P(|X - \mu| \geq 2 * 5) \leq \frac{1}{2^2} = \frac{1}{4}$$

$k = 2$



(c) What about $P\left(X \in (50, 70)\right)$?

$$P(X \in (50,70)) = 1 - P(X \geq 70 \text{ or } X \leq 50)$$

$$= 1 - P(|X - \mu| \geq 10) \geq 1 - \frac{1}{4} = \frac{3}{4}$$

$$P(X \in (50,70)) \geq 3/4.$$

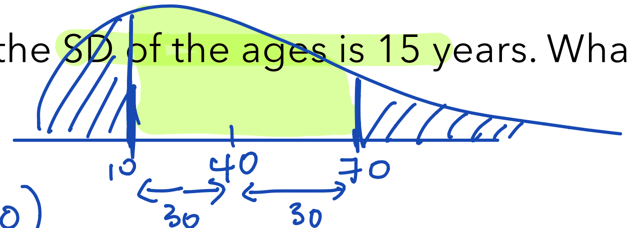
Exercise 6.5.6

Ages in a population have a mean of 40 years. Let X be the age of a person picked at random from the population.

a) If possible, find $P(X \geq 80)$. If it's not possible, explain why, and find the best upper bound you can based on the information given.

$$P(X \geq 80) \leq \frac{40}{80} = \frac{1}{2}$$

b) Suppose you are told in addition that the SD of the ages is 15 years. What can you say about $P(10 < X < 70)$?



$$\underbrace{P(10 < X < 70)}_{\geq 1/4} = P(|X-\mu| < 30)$$

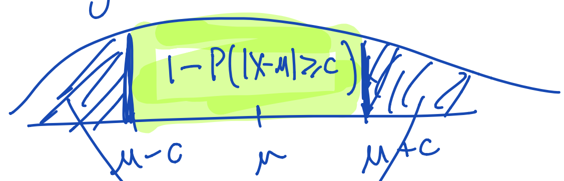
$$= P(|X-\mu| < \underbrace{2 * 15}_{2\sigma}) \geq 1 - \frac{1}{k^2} = 1 - \frac{1}{4} = 3/4$$

c) With the information as in Part b, what can you say about $P(10 \leq X \leq 70)$?

$$\underbrace{P(10 \leq X \leq 70)}_{\text{added } 10,\,70 \text{ as possible values}} \geq P(10 < X < 70) \geq \frac{3}{4}$$

Chebyshev's ineq. $P(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2}$

$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$



bounded above by $\frac{\sigma^2}{c^2}$

Examples

Suppose that each year, Berkeley admits 15,000 students on average, with an SD of 5,000 students. Assuming that the application pool is roughly the same across years, find the smallest upper bound you can on the probability that Berkeley will admit at least 22,500 students in 2021.

$\mu = 15000 \quad , \sigma = 5000$

Let X = # of students admitted

$$P(X \geq 22{,}500) \leq \frac{\overset{10}{\cancel{15000}}}{\underset{15}{\cancel{22500}}} = \frac{2}{3} \quad \text{(Markov)}$$

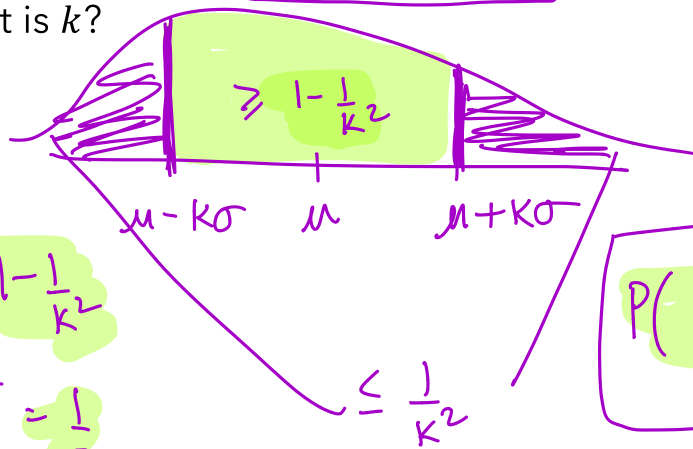
Chebyshev

$$P(X - \mu \geq 22500 - 15000)$$

$$= P(X - \mu \geq 7500) \leq \frac{(5000)^2}{(7500)^2} = \frac{4}{9}$$

Example

Suppose a list of numbers $x = \{x_1, \ldots, x_n\}$ has mean μ and standard deviation σ . Let k be the smallest number of standard deviations away from μ we must go to ensure the range $(\mu - k\sigma, \mu + k\sigma)$ contains at least 50% of the data in x. What is k?



$$\boxed{P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}}$$

$$P(X \in (\mu - k\sigma, \mu + k\sigma)) \geq 1 - \frac{1}{k^2}$$

$$1 - \frac{1}{k^2} = 0.5 = \frac{1}{2}$$

$$\frac{1}{k^2} = \frac{1}{2}, \quad \boxed{k = \sqrt{2}}$$

Example

$\mu = 1, \quad \sigma = 2$

A list of non negative numbers has an average of 1 and an SD of 2. Let p be the proportion of numbers over 4. To get an upper bound for p, you should:

a) Assume a binomial distribution ✗

b) Use Markov's inequality.

c) Use Chebyshev's inequality

d) None of the above.

$P(X \geq 4) = p \leq ??$

$P(X \geq 4) \leq \frac{1}{4}$ ← Markov

$P(X - \mu \geq 4 - \mu)$

$= P(X - \mu \geq 3) \leq \frac{2^2}{3^2} = \frac{4}{9}$

Example

Let X be a non-negative random variable such that $E(X) = 100 = Var(X)$.

a) Can you find $E(X^2)$ exactly? If not, what can you say?

b) Can you find $P\ (70 < X < 130)$ exactly? If not, what can you say?