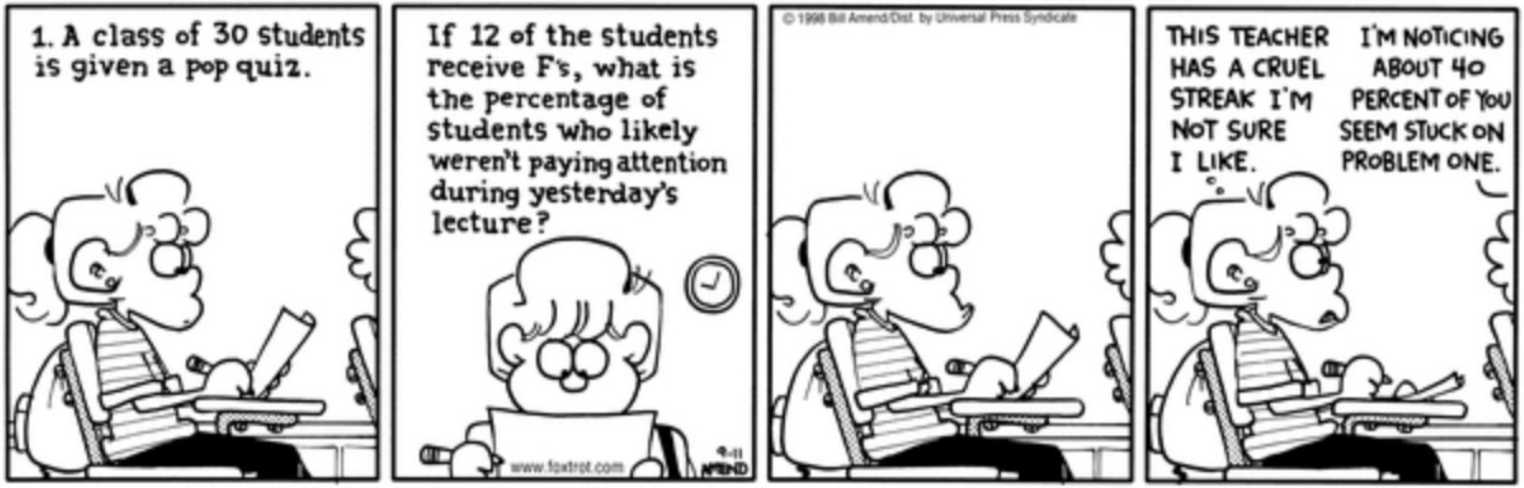
Stat 88: Probability & Mathematical Statistics in Data Science



Lecture 17: 3/17/2022

Finishing up the bounding of tail probabilities, Variance of a sum of random variables

6.4, 7.1, 7.2

Review: Variance and SD

$D = X - E(X)$

1. The variance of a rv $Var(X) = \overset{E(D^2)}{E[(X - E(X))^2]} = \underbrace{E(X^2) - (E(X))^2}$ computational formula.
2. The SD or standard deviation is given by $SD(X) = \sqrt{Var(X)}$
3. We denote the variance of X by σ^2 and the SD by σ. Note that $\sigma \geq 0$.
4. $SD\ (c) = 0$ $\quad SD(X+c) = \cancel{SD(X) + SD(c)} = SD(X)$
5. $SD(X + c) = SD(X)$, and $SD(cX) = |c|SD(X)$, c is a constant

$SD(-X) = SD(X)$

Exercise

Starting with the computational formula show that $Var(X+c) = Var(X)$

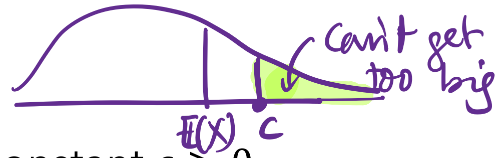
If X and Y are *independent*, then

6. $Var(X + Y) = Var(X) + Var(Y)$
7. $SD(X + Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$

$Var(X-Y) = Var(X) + Var(Y)$

$Var(X) = \sigma_X^2$ & if X is understood we write σ^2

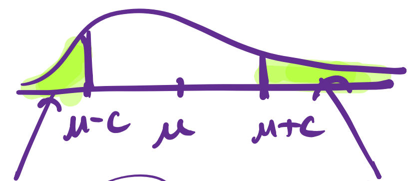
Review: Bounding the tail probabilities



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

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

- When to use which? If possible try both & use better one
- **Chebyshev's inequality**: For any random variable X (not necessarily non-negative), with mean μ and standard deviation σ, for any positive constant $c > 0$, we have: $P(|X-\mu| \geq c)$



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

- What is a "better bound"?

For an upper bound, the smaller / for a lower bound, larger

$X \leq \mu - c$

$X - \mu \leq -c$

$X \geq \mu + c$

$X - \mu \geq c$

- Ex: Is it possible that half of US flights have delay times at least 3 times the national average? Have to use Markov's ineq

X: delay time

$$P(X \geq c) \leq \frac{\mu}{c} \rightarrow P(X \geq \underbrace{3\mu}_{c}) \leq \frac{\mu}{3\mu} = \frac{1}{3}$$

No. The fraction of flights w/ delay times at least 3μ is at most $1/3$, so can't be $\frac{1}{2} > \frac{1}{3}$

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

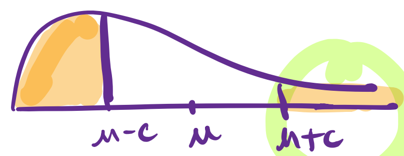
Chebyshev's inequality interpreted as distances

Let $c = k\sigma$

$$\frac{\sigma^2}{c^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

- 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:

$$\boxed{P(|X-\mu| \ge k \cdot \sigma) \le \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}}$$



- What if we are only interested in one tail (one side)?

Example: 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?

$\mu = 10000,\ \sigma = 550,\ c$

$a, b \ge 0$

$0 \le a + b \le 7$

$0 \le a \le 7$

$0 \le b \le 7$

$P(|X-\mu| \ge c) \leftarrow$ sum of two tails

$= P(X-\mu \le -c) + P(X-\mu \ge c)$

$P(|X-\mu| \ge c) \le \frac{\sigma^2}{c^2} \longrightarrow P(X-\mu \ge c) \le \frac{\sigma^2}{c^2}$

X = lifetime of bulb

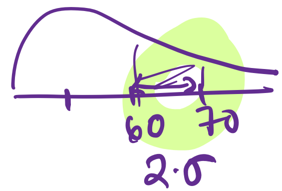
$$P(X \geq 11{,}980) = P(X - \mu \geq 11980 - \mu)$$

Chebyshev or Markov?

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

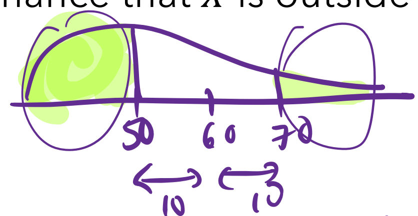
- Suppose X is a non-negative random variable with expectation 60 and SD 5.

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



$$P(X \geq 70) = P(X - \mu \geq 70 - \mu)$$
$$= P(X - \mu \geq 70 - 60)$$
$$= P(X - \mu \geq 10) \leq P(|X - \mu| \geq \underbrace{10}_{c}) \leq \frac{\sigma^2}{c^2} = \frac{25}{100} = \frac{1}{4}$$

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



$$P(|X - \mu| > 10) \leq P(|X - \mu| \geq 10) \leq \frac{\sigma^2}{10^2}$$

$$P(|X - \mu| > 10) \leq \frac{25}{100} = \frac{1}{4}$$

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

$$P(X \in (50, 70)) = 1 - \underbrace{P(X \leq 50 \text{ or } X \geq 70)}_{\leq \frac{1}{4}}$$

$$P(X \in (50, 70)) \geq 1 - \frac{1}{4} = \frac{3}{4}$$

Exercise 6.5.6 from text

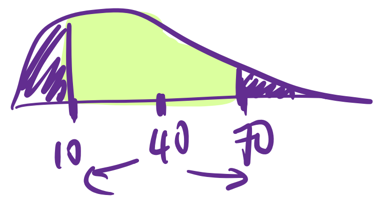
Ages in a population have a mean of 40 years. Let X be the age of a person picked at random from the population.

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.

Markov's Inequality: $P(X \geq 80) \leq \frac{\mu}{80} = \frac{1}{2}$

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

$P(10 < X < 70) = P(|X-40| < 30)$



$1 - P(|X-40| \geq \underset{2\sigma}{30}) \geq 1 - \left(\frac{1}{4}\right) = \frac{3}{4}$

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

Exercise.

Examples

Let X = # of students admitted

$E(X) = 15000, \quad \sigma_X = 5000$

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.

Exercise use both Markov's & Chebyshev's inequalities

$P(X \geq 22{,}500)$

Markov $P(X \geq 22500) \leq \frac{15000}{22500} = \frac{2}{3}$

Chebyshev

$$P(X \geq 22500)$$

$$= P(X - \mu \geq \underbrace{22500 - 15000}_{7500 \; c}) \leq \frac{\sigma^2}{c^2}$$

$$= \frac{5000^2}{7500^2} \approx \frac{4}{9}$$

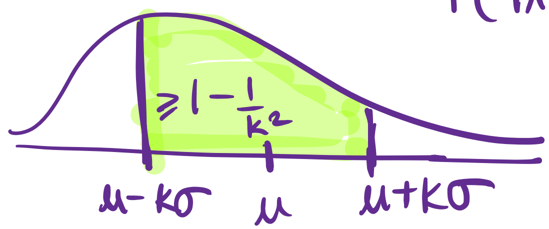
Example

Exercise

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?

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

We want $1 - \frac{1}{k^2} = 50\%$



Solve for k

$k = \sqrt{2}$

Example

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.

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?

$$Var(X) = E(X^2) - (E(X))^2$$

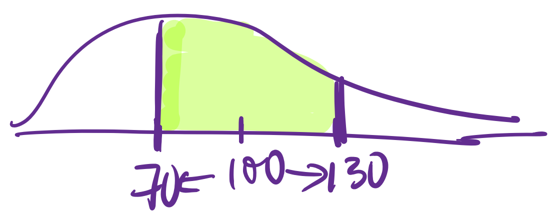
$$E(X^2) = Var(X) + (E(X))^2$$

$$= 100 + 100^2 = 10100$$

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

No

$$P(|X-\mu| < 30)$$



7.1: Sums of Independent Random Variables

- Recall that expectation is additive, which we used many times.
$$(E(X+Y) = E(X) + E(Y))$$

- What about $Var(X+Y)$? Well, it depends.

$H + T = \text{fixed } \#$

- Consider tossing a fair coin 10 times. Let H be the number of heads and T be the number of tails in 10 tosses. Then $H + T = 10$. Note that $\underbrace{Var(H)}_{\neq 0}, \underbrace{Var(T)}_{\neq 0} \neq 0$, but $Var(H+T) = Var(10) = \cancel{0!}\ 0$

10 tosses

- But now let H_1 be the number of heads in the first 5 tosses, and H_2 the number of heads in the last 5 tosses. Will we have that $Var(H_1 + H_2) = 0$?

$\underbrace{OOOOO}_{H_1}\ \underbrace{OOOOO}_{H_2}$

$H_1 \sim Bin(5, \frac{1}{2})$

$H_2 \sim Bin(5, \frac{1}{2})$

- Recall that if X and Y are ***independent***, then we have that
$$Var(X+Y) = Var(X) + Var(Y)$$

$$Var(H_1 + H_2) = Var(H_1) + Var(H_2)$$

In general $Var(X_1 + X_2 + \cdots + X_n) = Var(X_1) + Var(X_2) + \cdots + Var(X_n)$

if $X_1, X_2, \ldots X_n$ are indep.

Sums of iid random variables

S_n : sum of n r.v.

- Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with mean μ and variance σ^2. Define S_n to be their sum:

$\mathbb{E}(X_k) = \mu$

$Var(X_k) = \sigma^2$

$$\mathbb{E}(S_n) = \mathbb{E}(X_1 + X_2 + \cdots + X_n.)$$

$= \mu + \mu + \mu - - \cdot$

$\mathbb{E}(S_n) = n\mu$

$Var(S_n) = n \cdot \sigma^2$

- We already know that $E(S_n) = \sum E(X_k) = n\mu.$

- Now we can further say that:

because of INDEPENDENCE

$$Var(S_n) = Var(X_1 + X_2 + \cdots + X_n) = Var(X_1) + \cdots + Var(X_n) = n\sigma^2$$

$Var(S_n) = n\sigma^2$

$$\boldsymbol{SD(S_n) = \sqrt{n}\,\sigma} = \sqrt{n\sigma^2} = \sqrt{n} \cdot \sigma$$

SQUARE ROOT LAW.

- Notice that the expected value grows as n, but the sd grows as $\sqrt{n}$.)

$Var(X_1) = \sigma^2$ | $Var(X_1 + X_2) = 2\sigma^2$ | $Var(X_1 + X_2 + X_3 + X_4) = 4\sigma^2$

$SD(X_1) = \sigma$ | $SD(X_1 + X_2) = \sqrt{2}\sigma$ | $SD(X_1 + X_2 + X_3 + X_4) = 2\sigma$

Variance of the Binomial distribution

$$I_k = \begin{cases} 1 & \text{w.p. } p \ (\text{if } S) \\ 0 & \text{w.p. } (1-p) \ (\text{if } F) \end{cases}$$

$$E(I_k) = p$$

- Recall that a binomial random variable $X \sim Bin(n, p)$ is the sum of n iid Bernoulli(p) random variables $I_1, I_2, \dots, I_n$ where I_k is the indicator of success on the kth trial.
- What are the mean and variance of I_k? And therefore, what are the mean and variance of X? For what p will this variance be maximum?

$$Var(I_k) = E(I_k^2) - \underbrace{(E(I_k))^2}_{p^2}$$

$$I_k^2 = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases}$$

$$E(I_k^2) = p$$

$$= p - p^2 = p(1-p)$$

$$X = I_1 + I_2 + I_3 + \dots + I_n$$

$$E(X) = p + p + \dots + p = np$$

$$Var(X) = p(1-p) + p(1-p) + p(1-p) \dots + p(1-p)$$

$$= np(1-p) = npq \quad (q = 1-p)$$

$$SD(X) = \sqrt{npq}$$

$X \sim Pois(\mu)$ $P(X=k) = e^{-\mu} \frac{\mu^k}{k!}$, use defn of E(X) & formula for variance to derive $E(X) = Var(X) = \mu$

Variance of Poisson (μ) and geometric(p)

- Recall that one way to get the Poisson rv is by approximating the Binomial(n, p) distribution when n is large and p is small. ($\mu = np$)

- SD of the binomial distribution is $\sqrt{np(1-p)}$. $= \sqrt{np\,q}$
- Note that if p is small, $(1-p) \approx 1$, and we can say that $np(1-p) \approx np$.
- This makes it plausible that the SD of the Poisson(μ) distribution is $\sqrt{\mu}$

(can be derived using infinite series, optional exercise)

$X \sim Pois(\mu)$: $E(X) = Var(X) = \mu$

- Geometric(1/p) distribution:

Fact: the variance of the geometric distribution is $\frac{1-p}{p^2}$

- Ex: (Waiting till the 10th success) Suppose you roll a die until the 10th success. Let R be the number of rolls required. Find $SD(R)$.

$R = 10 \cdot T_1$ $SD(R) = 10 \cdot SD(T_1)$

Sampling without replacement

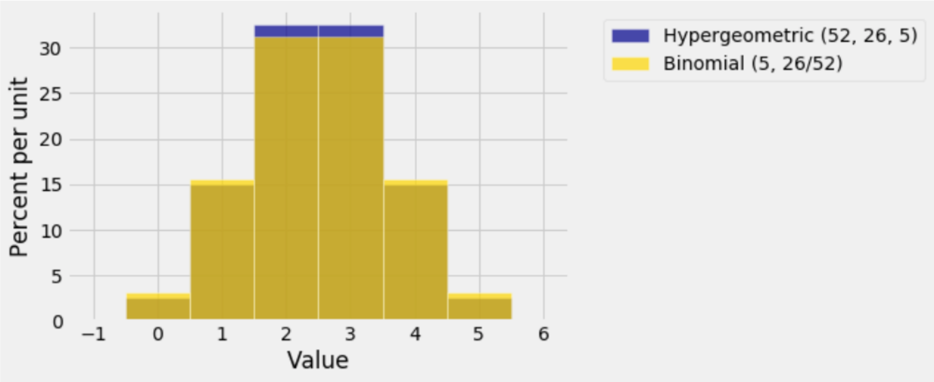
- When we have a simple random sample (SRS), the draws are without replacement (like drawing cards from a deck).
- The random variables are no longer independent
- So, how do we compute the variance of the sum of draws of a SRS?
- To begin with, let's look at the squares and products of indicators
- If I_A and I_B are indicator functions, what can we say about I_A^2 and $I_A I_B$?

Variance of a hypergeometric random variable

- Let $X \sim HG(N, G, n)$, then can write $X = I_1 + I_2 + \cdots + I_n$, where I_k is the indicator of the event that the kth draw is good.

- We can compute the expectation of X using symmetry: $E(X) \ =$
- But what about variance?

The finite population correction & the accuracy of SRS



From section 7.2

Accuracy of samples

Simple random samples of the same size of 625 people are taken in Berkeley (population: 121,485) and Los Angeles (population: 4 million). True or false, and explain your choice: The results from the Los Angeles poll will be substantially more accurate than those for Berkeley.