CHEBYSHEV'S THEOREM

I. MOTIVATIONS

In class, we covered lots of examples where we know that a dataset follows a specific distribution, both in continuous and discrete cases. Sometimes, we do not have this information, yet still want to be able to draw statistical inferences from a dataset without looking at individual values. Also, it's on the course description online, so you should probably know it!

In particular, given the mean and variance of a dataset, we want to predict the probability that a randomly selected sample falls within a specified interval around the mean. Of course, doing this exactly is not possible — what we are after here is a lower bound.

II. DEFINITIONS

RANDOM VARIABLE (DEF 1): Given a sample space S, a random variable X is a variable which represents an unknown value in S in accordance with some probability/frequency P(X = x) = f(x), where f is a probability mass or density function S → |R (Not Stanford University).

EXPECTED VALUE OF A RANDOM VARIABLE: The expected value of a random variable X where P(X = x) = f(x) is the weighted average of all possible values of X. In the continuous case,

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx \quad \text{(DEF 2).}$$

Additionally, given a function g(X), the expected value of g(X) is

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx \quad \text{(DEF 3).}$$

Definitions may be adapted for discrete random variables with sums over S.

VARIANCE OF A RANDOM VARIABLE (Definition 4): The variance of a random variable X~f(x), denoted Var[X] or σ^2 , represents the weighted mean euclidean distance between each datapoint and the mean. In the continuous case,

$$\begin{aligned} Var[X] &= E[(X - E(X))^2 \\ &= E[X]^2 - E[X^2] \end{aligned}$$

by the linearity of expectation — proof is left as an exercise to reader, and follows trivially from expanding definition three for $g(x) = (X - E(X))^2$.

III. CHEBYSHEV'S INEQUALITY (DEF 5)

Let X (integrable) be a random variable with finite non-zero variance σ^2 (and finite expected value μ).[7] Then for any real number k > 1,

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

PROOF:

Let X be a continuous random variable with $E[X] = \mu$ and $Var[X] = \sigma^2 > 0$.

Claim 1: For a continuous random variable Y with P(Y = y) = f(y) and finite, positive real number a, $P(Y \geq a) \leq E[Y]/a$ (Markov).

Proof of claim:

$$\begin{aligned} E[Y] &= \int_{-\infty}^{\infty} yf(y)dy \\ &\geq \int_{a}^{\infty} yf(y)dy \\ &\geq \int_{a}^{\infty} af(y)dy \\ &= a\int_{a}^{\infty} f(y)dy \\ &= a * P(Y \geq a). \end{aligned}$$

Dividing both sides by a yields $P(Y \geq a) \leq E[Y]/a$.

Now set $Y = (X - \mu)^2$ and $a = k^2\sigma^2$. Then

$$\begin{aligned} P(|X - \mu| \geq k\sigma) &= P((X - \mu)^2 \geq k^2\sigma^2) \\ &\leq E[(X - \mu)^2] / k^2\sigma^2 \\ &= \sigma^2 / (k^2\sigma^2) \\ &= 1 / (k^2). \end{aligned}$$

Subtracting both sides of the inequality from 1, it follows that

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

as desired.

IV. EXAMPLES:

1. Suppose that the mean time spent by students in Dr. John McCuan's Math 3215 on their homework over the entire semester is 36 hours with variance 2 hours. Using Chebyshev's inequality, find the strongest lower-bound on the probability that a randomly selected student spends between 32 and 40 hours on their homework.

Solution:
Let X represent the number of hours that a student takes on their homework. Note SD(X) = sqrt(Var(x)) = sqrt(2). We know 40-36 = 36-32 = 4 = sqrt(2)* 2sqrt(2), so we pick k = 2*sqrt(2).
Then P(32 < X < 40) = P(36 - ksqrt(2) < X < 36 + ksqrt(2))
>= 1 - 1/(k^2)
= 1 - 1/((2sqrt(2))^2)
= 1 - ⅛ = **0.875.**

2. A variation on a variation of someone who lectured before's problem: Jeremy and John are competing in a 7 round integration bee. In each round, Jeremy wins with probability .9 and loses with probability 0.1. The first to 4 wins is the victor.
    a. Calculate the exact probability that Jeremy wins the match in between 4 and 6, inclusive, rounds.
    b. Calculate a lower-bound for the quantity in (a) using Chebyshev's theorem
    c. Compare them!

Solution: Let X be the number of rounds that it takes for Jeremy to win. Then with Y = X - 4, Y~NegativeBinomial(r=4, p=0.9).

a) P(4 <= X <= 6) = P (0 <= Y <= 2) = f(0) + f(1) + f(2) = .656 + .262 + .066 = .997
b) Note that E(Y) = 4(0.1)/(0.9) = 0.444, Var(Y) = 4(0.1)/(0.9*0.9) = .494, so E(X) = 5.444, Var(X) = .494, and SD(X) = .703.

Then P(4 <= X <= 6) = P(|X - 5| <= SD(X) * 1/SD(X))
= P(|X - 5| <= SD(X) * 1.422)
>= 1 - 1/(1.422^2)
= 0.505%

c) The lower bound from Chebyshev's theorem is pretty bad.
    i) Even supposing we double the size of the interval: even then, we have a lower bound of .8722, which is still drastically less than the real answer of 0.99999

V. PROBLEM SET

1. (BERKELEY) Let f(x) = $5/x^6$ for x ≥ 1 and 0 otherwise. What bound does Chebyshev's inequality give for the probability P(X ≥ 2.5)? For what value of a can we say P(X ≥ a) ≤ 15%?
2. Let f(x) be the uniform distribution on 0 ≤ x ≤ 20 and 0 everywhere else. Give a bound using Chebyshev's for P(4 ≤ X ≤ 16). Calculate the actual probability. How do they compare?
3. (UIUC) The number of items produced in a factory during a week is a random variable with mean 50. a) What can be said about the probability that this week's production is at least 100? b) If the variance of a week's production is known to equal 25, can we obtain a better bound for part (a)?
4. Let $f(x) = m/x^{m+1}$ where m is an integer more than 2, for x ≥ 1 and 0 everywhere else. Give a bound using Chebyshev's Inequality for P(1 ≤ X ≤ $(m + 1)/(m - 1)$).

This might take some thinking. Extends Chebyshev to any interval — based on a 1940 paper that still gives the strongest bounds we know of.

5. **CHALLENGE PROBLEM** (SELBERG, 1940): Suppose *X* is a random variable with mean μ and variance σ^2, and a and b are positive real numbers.
   a. Show that $P(\mu - a \leq X \leq \mu + b) \geq a^2/(a^2 + \sigma^2)$ when $a(b - a) \geq 2\sigma^2$
   b. Show that $P(\mu - a \leq X \leq \mu + b) \geq (4ab - 4\sigma^2)/(a + b)^2$ when $2ab \geq 2\sigma^2 \geq a(b - a)$
   c. Show that $P(\mu - a \leq X \leq \mu + b) \geq 0$ when $\sigma^2 \geq ab$
   d. Show that when $a = b$, this system of inequalities generated in (a, b, c) reduces to Chebyshev's inequality. [easier]