CS 112: Modeling Uncertainty in Information Systems

Prof. Jenn Wortman Vaughan

May 14, 2012

Lecture 12

Reminders & Announcements

- Homework 3 is due this Friday, May 18
- Minor updates to the skeleton code and sample files have been posted on Piazza

Today…

- Bounding random variables
  - Markov Inequality
  - Chebyshev Inequality

- The Law of Large Numbers

- The Central Limit Theorem

Motivation: The Sample Mean

Suppose we would like to estimate the president's approval rating. We ask n random voters whether or not they approve of the president, and use the fraction of voters who say that they approve as our estimate.

How accurate is our estimate as a function of n?

Motivation: The Sample Mean

Suppose we would like to estimate the president's approval rating. We ask n random voters whether or not they approve of the president, and use the fraction of voters who say that they approve as our estimate.

How accurate is our estimate as a function of n?

What if we want to know more than just mean and variance?

Bounding Random Variables

Markov Inequality

Theorem: If a random variable X can only take nonnegative values, then for all $a > 0$,

$$P(X \geq a) \leq \frac{E[X]}{a}$$

Markov Inequality

Theorem: If a random variable X can only take nonnegative values, then for all $a > 0$,

$$P(X \geq a) \leq \frac{E[X]}{a}$$

Translation: If a nonnegative random variable has a small mean, then the probability that it takes on a large value must also be small.

Shuffle Mode

Suppose you have n songs on your MP3 player. In shuffle mode, songs are picked uniformly at random. Let X be a random variable representing the number of songs that play in shuffle mode before you have heard each of the n songs once.

Shuffle Mode

Suppose you have *n* songs on your MP3 player. In shuffle mode, songs are picked uniformly at random. Let X be a random variable representing the number of songs that play in shuffle mode before you have heard each of the *n* songs once.

What is E[X]?

Shuffle Mode

Suppose you have n songs on your MP3 player. In shuffle mode, songs are picked uniformly at random. Let X be a random variable representing the number of songs that play in shuffle mode before you have heard each of the n songs once.

What is E[X]?

What is the probability that $X \geq a$?

Shuffle Mode

Suppose $n = 1000$

Then $E[X] \approx n \ln(n) = 6908$, and we have…

$$P(X \geq 10{,}000) \leq 0.7$$

$$P(X \geq 20{,}000) \leq 0.35$$

$$P(X \geq 50{,}000) \leq 0.14$$

Markov Inequality Is Not Tight

Suppose X is uniformly distributed in [0,4]

- What does the Markov inequality say about $P(X \geq 2)$?
- What about $P(X \geq 3)$ and $P(X \geq 4)$?

Markov Inequality Is Not Tight

Suppose X is uniformly distributed in [0,4]

- What does the Markov inequality say about $P(X \geq 2)$?
- What about $P(X \geq 3)$ and $P(X \geq 4)$?

But it can still be useful!

Chebyshev Inequality

Theorem: If X is a random variable with mean μ and variance σ^2, then for any $c > 0$,

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

Chebyshev Inequality

Theorem: If X is a random variable with mean μ and variance σ^2, then for any $c > 0$,

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

Translation: If a random variable has small variance, then the probability that it takes a value far from its mean must also be small.

Chebyshev Inequality

Theorem: If X is a random variable with mean μ and variance σ^2, then for any $c > 0$,

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

Translation: If a random variable has small variance, then the probability that it takes a value far from its mean must also be small.

(Note that X does not have to be nonnegative here)

Back to Shuffle Mode

Suppose you have n songs on your MP3 player. In shuffle mode, songs are picked uniformly at random. Let X be a random variable representing the number of songs that play in shuffle mode before you have heard each of the n songs once.

How likely is it that $|X - E[X]| \geq c$?

Shuffle Mode

Suppose $n = 1000$

Then $E[X] \approx n \ln(n) = 6908$, and we have…

$$P(|X-E[X]| \geq 2000) \leq 0.42$$

$$P(|X-E[X]| \geq 5000) \leq 0.07$$

Back to Shuffle Mode

Suppose you have n songs on your MP3 player. In shuffle mode, songs are picked uniformly at random. Let X be a random variable representing the number of songs that play in shuffle mode before you have heard each of the n songs once.

How likely is it that $|X - E[X]| \geq c$?

Can use this to get much tighter bounds on $P(X \geq a)$

Shuffle Mode

Suppose $n = 1000$

Then $E[X] \approx n \ln(n) = 6908$, and we have…

$$P(|X-E[X]| \geq 2000) \leq 0.42$$

$$P(|X-E[X]| \geq 5000) \leq 0.07$$

$$P(X \geq 10{,}000) \leq P(|X-E[X]| \geq 3092) \leq 0.17$$

$$P(X \geq 20{,}000) \leq P(|X-E[X]| \geq 13{,}092) \leq 0.001$$

Much tighter than Markov in this case…

Chebyshev is Still Not Tight

Suppose X is uniformly distributed in [0,4]

- What does Chebyshev say about $P(|X - 2| \geq 1)$?

The Sample Mean

Suppose we would like to estimate the president's approval rating. We ask n random voters whether or not they approve of the president, and use the fraction of voters who say that they approve as our estimate.

- What is the probability that our estimate differs from the true approval rating by more than ε?

The Sample Mean

Suppose we would like to estimate the president's approval rating. We ask n random voters whether or not they approve of the president, and use the fraction of voters who say that they approve as our estimate.

- What is the probability that our estimate differs from the true approval rating by more than ε?

- If we would like to have high confidence (say, 95%) that our estimate is very accurate (say, within 0.01 of the true approval rating), how many random voters must we poll?

The Sample Mean

We can generalize the idea of the sample mean to other sequences of independent identically distributed random variables X_1, X_2, … , X_n too…

Law of Large Numbers

Theorem: Let X_1, X_2, … be independent identically distributed random variables with mean μ. For every $\varepsilon > 0$,

$$P\left(\left| \frac{X_1 + \ldots + X_n}{n} - \mu \right| \geq \varepsilon \right) \to 0 \quad \text{as} \quad n \to \infty$$

Law of Large Numbers

Theorem: Let X_1, X_2, … be independent identically distributed random variables with mean μ. For every $\varepsilon > 0$,

$$P\left(\left| \frac{X_1 + \ldots + X_n}{n} - \mu \right| \geq \varepsilon \right) \to 0 \quad \text{as} \quad n \to \infty$$

Translation: As the size of our sample gets very large, the probability that the sample mean is very close to the true mean goes to 1.

Law of Large Numbers

Theorem: Let X_1, X_2, … be independent identically distributed random variables with mean μ. For every $\varepsilon > 0$,

$$P\left(\left| \frac{X_1 + \ldots + X_n}{n} - \mu \right| \geq \varepsilon \right) \to 0 \quad \text{as} \quad n \to \infty$$

Translation: As the size of our sample gets very large, the probability that the sample mean is very close to the true mean goes to 1.

(Note that we can make ε as small as we want.)

The Central Limit Theorem

Loosely stated: Let X_1, X_2, … be independent identically distributed random variables with mean μ. As $n \to \infty$, the cumulative distribution function of the sample mean M_n approaches the cumulative distribution function of a normal random variable.

Translation: As n gets large, the average (or sum) of n i.i.d. random variables is approximately normal.