INEQUALITIES, DISTRIBUTIONS 10

COMPUTER SCIENCE MENTORS 70

November 5 - 9, 2018

1 Markov, Chebyshev

1.1 Introduction

Markov's Inequality:

For a non-negative random variable X with expectation $\mathrm{E}(X) = \mu$, and any $\alpha > 0$:

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

Chebyshev's Inequality

For a random variable X with expectation $\mathrm{E}(X) = \mu$, and any $\alpha > 0$:

$$\mathrm{P}[|X - \mu| \geq \alpha] \leq \frac{\mathrm{Var}(X)}{\alpha^2}$$

Chebyshev's Inequality can be used to estimate the mean of an unknown distribution. Often times we do not know the true mean, so we take many samples $X_1, X_2, ..., X_n$. Our sample mean is thus $S_n = \frac{X_1+X_2+...+X_n}{n}$

We can upper-bound the probability of that our sample mean deviates too much from our true mean as:

$$P(|\hat{\mu} - \mu| > \epsilon) \leq \delta$$

where ϵ is known as our error and δ is known as our confidence.

1.2 Questions

1. Use Markov's to prove Chebyshev's Inequality.

2. A random variable X is always strictly larger than -100. You know that $\mathrm{E}(X) = -60$. Give the best upper bound you can on $\mathrm{P}[X \geq -20]$.

3. **A**s we all know, Berkeley squirrels are extremely fat and cute. The average squirrel is 40% body fat. The standard deviation of body fat is 5%. Provide an upper bound on the probability that a randomly trapped squirrel is either too skinny or too fat? A skinny squirrel has less than 27.5% body fat, and a fat squirrel has more than 52.5% body fat?

4. Consider a random variable Y with expectation μ whose maximum value is $\frac{3\mu}{2}$, prove that the probability that Y is 0 is at most $\frac{1}{3}$.

2 Distributions

2.1 Bernoulli Distribution

Bernoulli Distribution: Bernoulli(p)
Random variable X has the Bernoulli distribution if it takes on value 1 with probability p, and value 0 with probability $1-p$. With the Bernoulli distribution we can model a single countable event, i.e. a single coin flip.
Expectation:

$$E(X) = 0 * (1-p) + 1 * p = p$$

Variance:

$$var(X) = E(X^2) - E(X)^2 = 0^2 * (1-p) + 1^2 * p - p^2 = p(1-p)$$

2.2 Binomial Distribution

Binomial Distribution: Bin(n, p)
The binomial distribution is used to count the number of successes in n independent trials. Each trial has a probability p of success. For this reason, we can think of the binomial distribution as a sum of n independent Bernoulli trials, each with probability p.
The probability of having k successes:

$$\mathrm{P}[X = k] = \binom{n}{k} * p^k * (1-p)^{n-k}$$

Expectation:

$$E(X) = E(X_1 + \ldots + X_n) = E(X_1) + \ldots + E(X_n) = np$$

Variance:
Since our sum of Bernoulli trials is independent, we can do the following:

$$var(X) = var(X_1 + X_2 + ... + X_n) = var(X_1) + var(X_2) + ... + var(X_n) = np(1-p)$$

1. Say that the Cleveland Browns have probability $p = 0.15$ chance of winning a football game. Assume that each game is independent from the last. The football regular season has 16 games.

(a) Write an expression for the probability that they win between 6 and 8 games during the course of a season.

(b) Find the probability that they win at least one game during the course of a season.

2.3 Poisson Distribution

Poisson Distribution: Pois(λ)

The Poisson distribution is an approximation of the binomial distribution under two conditions:

- n is very large
- p is very small

Let $\lambda = np$ represent the "rate" at which some event occurs. We usually use this distribution when these events are rare.

The probability of k occurrences is

$$\mathrm{P}[X = k] = \frac{e^{-\lambda} * \lambda^k}{k!}$$

Expectation:

Since $\mathrm{E}(X) = np$ for the binomial distribution, and we set $\lambda = np$, our expectation is $\mathrm{E}(X) = \lambda$.

Variance:

For a binomial distribution, Var$(X) = np(1-p)$, which looks like $\lambda(1-p)$. However, we started with the assumption that p is very small, so we can assume $(1-p) \approx 1$ and thus $\lambda(1-p) \approx \lambda$. Therefore, Var$(X) = \lambda$.

1. On the UC Berkeley meme page, on average, 3 good memes are posted a week. What is the probability that in a given week:

   (a) 6 good memes are posted?

   (b) No good memes are posted?

   (c) More than 1 good meme is posted?

2.4 Geometric Distribution

Geometric Distribution: Geom(p)
With the geometric distribution, we count the number of failures until the first success. The probability that the first success occurs on trial k is:

$$\mathrm{P}[X = k] = (1-p)^{k-1} * p, k > 0$$

Expectation:

We can derive the geometric distribution from the binomial distribution. $\mathrm{E}(X)$ is the number of trials until the first success occurs, including that first success. There are two cases:

1. The first success occurs, with probability p.
2. We obtain a failure, with probability $1-p$, meaning that we are back where we started but already used one trial.

Putting this together, we get:

$$E(X) = p * 1 + (1-p) * (1 + E(X)) \implies E(X) = \frac{1}{p}$$

Variance:

$$var(X) = \frac{1-p}{p^2}$$

1. Andy Go-es to class 20% of days. What is the probability that:
   (a) the first time he goes to class is the 5^{th} day of school?

   (b) the first time he goes to class is after the 5^{th} day?

   (c) He goes to class on the 5^{th} day or before?

2.5 Questions

1. In this problem, we will explore how we can apply multiple distributions to the same problem.

   Suppose you are a professor doing research in *machine learning*. On average, you receive 12 emails a day from students wanting to do research in your lab, but this number varies greatly.

(a) Which distribution would you use to model the number of emails you receive from students on any one day?

(b) What is the probability that you receive 7 emails tomorrow? At least 7?

(c) Now, let's look at the month of April, in which lots of students are emailing you to secure a summer position. What is the probability that the first day in April that you receive exactly 15 emails is April 8th? *Hint: Break this problem down into parts, and assign your result to the first part to the variable* p.

(d) Now, calculate the probability that April 8th is the first day that we receive **at least** 15 emails.

(e) What is the probability that you receive at least 15 emails on 10 different days in April?