DISTRIBUTIONS, VARIANCE, INEQUALITIES, CONFIDENCE INTERVALS 8

COMPUTER SCIENCE MENTORS 70

March 20 - March 14, 2017

1 Distributions

1.1 Introduction

Geometric Distribution: Geom(p) Number of trials required to obtain the first success. Each trial has probability of success equal to p. The probability of the first success happening at trial k is:

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

The expectation of a geometric distribution is:

$$\mathrm{E}(X) = \frac{1}{p}$$

The variance of a geometric distribution is:

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

Binomial Distribution: Bin(n, p) Number of successes when we do n independent trials. Each trial has a probability p of success. The probability of having k successes:

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

The expectation of a binomial distribution is:

$$\mathrm{E}(X) = np$$

The variance of a binomial distribution is:

$$Var(X) = np(1-p)$$

Poisson Distribution: Pois(λ) This is an approximation to the binomial distribution. Let the number of trials approach infinity, let the probability of success approach 0, such that $\mathrm{E}(X) = np = \lambda$. This is an accepted model for rare events. The probability of having k successes:

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

The expectation of a poisson distribution is:

$$\mathrm{E}(X) = \lambda$$

The variance of a poisson distribution is:

$$Var(X) = \lambda$$

1.2 Questions

1. You are Eve, and as usual, you are trying to break RSA. You are trying to guess the factorization of N, from Bobs public key. You know that N is approximately 1,000,000,000,000. To find the primes p and q, you decide to try random numbers from 2 to $1,000,000 \approx \sqrt{N}$, and see if they divide N.

   To do this, you roll a 999,999-sided die to choose the number, and see if it divides N using your calculator, which takes five seconds. Of course, there will be one number in this range that does divide Nnamely, the smaller of p and q.

(a) What kind of distribution would you use to model this?

(b) What is the expected amount of time until you guess the correct answer, if it takes five seconds per guess (you only have a calculator)? Answer in days.

2. Now you are trying to guess the 6-digit factorization digit by digit. Lets assume that when you finish putting these digits together, you can figure out how many digits you got right. Use zeros for blank spaces. For example, to guess 25, you would put 000025

   (a) What kind of distribution would you use to model this?

   (b) What is the probability that you get exactly 4 digits right?

   (c) What is the probability that you get less than 3 correct?

3. You are Alice, and you have a high-quality RSA-based security system. However, Eve is often successful at hacking your system. You know that the number of security breaches averages 3 a day, but varies greatly.

   (a) What kind of distribution would you use to model this?

   (b) What is the probability you experience exactly seven attacks tomorrow? At least seven (no need to simplify your answer)?

   (c) What is the probability that, on some day in April, you experience exactly six attacks?

2 Variance

2.1 Introduction

For a random variable X with expectation $\mathrm{E}(X) = \mu$, the variance of X is:

$$\text{Var}(X) = \mathrm{E}((X - \mu)^2)$$

The square root of Var(X) is called the standard deviation of X

Theorem: For a random variable X with expectation $\mathrm{E}(X) = \mu$ and a constant c,

$$\text{Var}(X) = \mathrm{E}(X^2) - \mu^2$$

$$\text{Var}(cX) = c^2 * \text{Var}(X)$$

Theorem: For a random variable X, expectation $\mathrm{E}(X) =$

$$\sum_a a * Pr[X = A]$$

2.2 Questions

1. Let's consider the classic problems of flipping coins and rolling dice. Let X be a random variable for the number of coins that land on heads and Y be the value of the die roll.

   (a) What is the expected value of X after flipping 3 coins? What is the variance of X?

   (b) Let Y be the sum of rolling a dice 1 time. What is the expected value of Y?

   (c) What is the variance of Y?

2. Say you're playing a game with a coin and die, where you flip the coin 3 times and roll the die once. In this game, your score is given by the number of heads that show multiplied with the die result. What is the expected value of your score? Whats the variance?

3. You are at a party with n people where you have prepared a red solo cup labeled with their name. Before handing red cups to your friends, you pick up each cup and put a sticker on it with probability $\frac{1}{2}$ (independently of the other cups). Then you hand back the cups according to a uniformly random permutation. Let X be the number of people who get their own cup back AND it has a sticker on it.

   (a) Compute the expectation $\mathrm{E}(X)$.

   (b) Compute the variance $\mathrm{Var}(X)$

4. a. Prove that for independent random variables X and Y, $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$.

   b. Is the above result true for non-independent random variables? Prove or give a counterexample.

5. Consider the random variable $X = X_1 + \ldots + X_n$, where X_i equals i with probability $\frac{1}{i}$ and 0 otherwise.

   (a) What is the variance of X? (Assume that X_i and X_j are independent for $i \neq j$)

   (b) For what value of n does $\text{E}(X) = \text{Var}(X)$?

   (c) For what value of n does $E(X) = SD(X) * \sqrt{2} + 100$?

6. An urn contains n balls numbered $1, 2, \ldots, n$. We remove k balls at random (without replacement) and add up their numbers. Find the mean and variance of the total.

3 Weak Law of Large Numbers

1. As our number of fair coin flipping trials goes to infinity, what does the probability that the proportion of heads is not $\frac{1}{2}$ go to?

4 Markov, Chebyshev

4.1 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 $\alpha > 0$:

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

4.2 Questions

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

2. **Squirrel Standard Deviation**
   As 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?

3. **Bound It**
   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 $\mathrm{P}[X \geq 20]$.

4. Give a distribution for a random variable where the expectation is 1,000,000 and the probability that the random variable is zero is 99%.

5. 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 $\frac{1}{3}$.

6. Let X be the sum of 20 i.i.d. Poisson random variables $X_1, \ldots, X_{20}$ with $\mathrm{E}(X_i) = 1$. Find an upper bound of $\mathrm{P}[X \geq 26]$ using,

   (a) Markov's inequality:

   (b) Chebyshev's inequality:

5 Confidence Intervals

5.1 Questions

1. Define i. i. d. variables $A_k \sim \mathrm{Bern}(p)$ where $k \in [1, n]$. Assume we can declare that $\mathrm{P}[|\frac{1}{n}\sum_k A_k - p| > 0.25] = 0.01$.

   (a) Please give a 99% confidence interval for p if given A_k.

(b) We know that the variables X_i, for i from 1 to n, are i.i.d. random variables and have variance. We also have a value (an observation) of $A_n = \frac{X_1+...+X_n}{n}$. We want to guess the mean, μ, of each X_i.

Prove that we have 95% confidence μ lies in the interval $\left[A_n - 4.5\frac{\sigma}{\sqrt{n}}, A_n + 4.5\frac{\sigma}{\sqrt{n}}\right]$

That is, $\mathrm{P}\left[\mu \in \left[A_n - 4.5\frac{\sigma}{\sqrt{n}}, A_n + 4.5\frac{\sigma}{\sqrt{n}}\right]\right] \geq 95\%$

(c) Give the 99% confidence interval for μ:

2. We have a die whose 6 faces are values of consecutive integers, but we dont know where it starts (it is shifted over by some value k; for example, if $k = 6$, the die faces would take on the values $7, 8, 9, 10, 11, 12$). If we observe that the average of the n samples (n is large enough) is 15.5, develop a 99% confidence interval for the value of k.