Expectation of geometric distribution

What is the probability that \( X \) is finite?

\[
\sum_{k=1}^{\infty} f_X(k) = \sum_{k=1}^{\infty} (1-p)^{k-1} p = p \sum_{k=0}^{\infty} (1-p)^k = \frac{p}{1-(1-p)} = 1
\]

Can now compute \( E(X) \):

\[
E(X) = \sum_{k=1}^{\infty} k \cdot (1-p)^{k-1} p
= p \sum_{k=1}^{\infty} (1-p)^{k-1} + \sum_{k=2}^{\infty} (1-p)^{k-1} + \sum_{k=3}^{\infty} (1-p)^{k-1} + \cdots
= p\left(\frac{1}{p} + \frac{1-p}{p^2} + \frac{(1-p)^2}{p^3} + \cdots\right)
= 1 + (1-p) + (1-p)^2 + \cdots
= \frac{1}{p}
\]

So, for example, if the success probability \( p \) is 1/3, it will take on average 3 trials to get a success.

• All this computation for a result that was intuitively clear all along . . .

Why not use \(|X(s) - E(X)|\) as the measure of distance instead of variance?

• \((X(s) - E(X))^2\) turns out to have nicer mathematical properties.

• In \( \mathbb{R}^n \), the distance between \((x_1, \ldots, x_n)\) and \((y_1, \ldots, y_n)\) is \(\sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}\).

Example:

• The variance of distribution 1 is
  \[
  \frac{1}{4}(51 - 50)^2 + \frac{1}{4}(50 - 50)^2 + \frac{1}{4}(49 - 50)^2 = \frac{1}{2}
  \]

• The variance of distribution 2 is
  \[
  \frac{1}{3}(100 - 50)^2 + \frac{1}{3}(50 - 50)^2 + \frac{1}{3}(0 - 50)^2 = \frac{5000}{3}
  \]

Expectation and variance are two ways of compactly describing a distribution.

• They don’t completely describe the distribution

• But they’re still useful!

Variance and Standard Deviation

Expectation summarizes a lot of information about a random variable as a single number. But no single number can tell it all.

Compare these two distributions:

• Distribution 1: \( \Pr(49) = \Pr(51) = 1/4; \ \Pr(50) = 1/2. \)

• Distribution 2: \( \Pr(0) = \Pr(50) = \Pr(100) = 1/3. \)

Both have the same expectation: 50. But the first is much less ‘dispersed’ than the second. We want a measure of dispersion.

• One measure of dispersion is how far things are from the mean, on average.

Given a random variable \( X \), \((X(s) - E(X))^2\) measures how far the value of \( s \) is from the mean value (the expectation) of \( X \). Define the variance of \( X \) to be

\[
\text{Var}(X) = E((X - E(X))^2) = \sum_{s \in S} \Pr(s)(X(s) - E(X))^2
\]

The standard deviation of \( X \) is

\[
\sigma_X = \sqrt{\text{Var}(X)} = \sqrt{\sum_{s \in S} \Pr(s)(X(s) - E(X))^2}
\]

Variance: Examples

Let \( X \) be Bernoulli, with probability \( p \) of success. Recall that \( E(X) = p. \)

\[
\text{Var}(X) = (0 - p)^2 \cdot (1 - p) + (1 - p)^2 \cdot p
= p(1-p)[p + (1-p)]
= p(1-p)
\]

Theorem: \( \text{Var}(X) = E(X^2) - E(X)^2. \)

Proof:

\[
E(X - E(X))^2 = E(X^2 - 2EXX + E(X)^2)
= E(X^2) - 2EXX + E(X)^2
= E(X^2) - 2E(X)^2 + E(X)^2
= E(X^2) - E(X)^2
\]

Think of this as \( E((X - c)^2) \), then substitute \( E(X) \) for \( c. \)

Example: Suppose \( X \) is the outcome of a roll of a fair die.

• Recall \( E(X) = 7/2. \)

• \( E(X^2) = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + \cdots + 6^2 \cdot \frac{1}{6} = \frac{91}{3} \)

• So \( \text{Var}(X) = \frac{91}{3} - (\frac{7}{2})^2 = \frac{17}{12}. \)
Markov’s Inequality

**Theorem:** Suppose $X$ is a nonnegative random variable and $\alpha > 0$. Then

$$\Pr(X \geq \alpha E(X)) \leq \frac{1}{\alpha}.$$  

**Proof:**

$E(X) = \sum_{x} x \cdot \Pr(X = x) \\
\geq \sum_{x \geq \alpha} x \cdot \Pr(X = x) \\
\geq \sum_{x \geq \alpha} \alpha E(X) \cdot \Pr(X = x) \\
= \alpha E(X) \sum_{x \geq \alpha} \Pr(X = x) \\
= \alpha E(X) \cdot \Pr(X \geq \alpha E(X)).$

**Example:** If $X$ is $B_{100,1/2}$, then

$$\Pr(X \geq 100) = \Pr(X \geq 2E(X)) \leq \frac{1}{2}.$$  

This is not a particularly useful estimate. In fact, $\Pr(X \geq 100) = 2^{-100} \sim 10^{-30}$.

Chebyshev’s Inequality: Example

Chebyshev’s inequality gives a lower bound on how well $X$ concentrated about its mean.

- Suppose $X$ is $B_{100,1/2}$ and we want a lower bound on $\Pr(40 < X < 60)$.
- $E(X) = 50$ and
  
  $$40 < X < 60 \text{ iff } |X - 50| < 10$$

  so

  $$\Pr(40 < X < 60) = \Pr(|X - 50| < 10) = 1 - \Pr(|X - 50| \geq 10).$$

  Now

  $$\Pr(|X - 50| \geq 10) \leq \frac{\text{Var}(X)}{100(1/2)^2} = \frac{1}{4}.$$  

  So

  $$\Pr(40 < X < 60) \geq 1 - \frac{1}{4} = \frac{3}{4}.$$  

This is not too bad: the correct answer is ~ 0.9611.

Chebyshev’s Inequality

**Theorem:** If $X$ is a random variable and $\beta > 0$, then

$$\Pr(|X - E(X)| \geq \beta \sigma_X) \leq \frac{1}{\beta^2}.$$  

**Proof:** Let $Y = (X - E(X))^2$. Then

$$|X - E(X)| \geq \beta \sigma_X \text{ iff } Y \geq \beta^2 \text{Var}(X).$$

I.e.,

$$\{s : |X(s) - E(X)| \geq \beta \sigma_X\} = \{s : Y(s) \geq \beta^2 \text{Var}(X)\}.$$  

In particular, the probabilities of these events are the same:

$$\Pr(|X - E(X)| \geq \beta \sigma_X) = \Pr(Y \geq \beta^2 \text{Var}(X)).$$

Note that $E(Y) = E[(X - E(X))^2] = \text{Var}(X)$, so

$$\Pr(Y \geq \beta^2 \text{Var}(X)) = \Pr(Y \geq \beta^2 E(Y)).$$

Since $Y \geq 0$, by Markov’s inequality

$$\Pr(|X - E(X)| \geq \beta \sigma_X) = \Pr(Y \geq \beta^2 E(Y)) \leq \frac{1}{\beta^2}.$$  

- Intuitively, the probability of a random variable being $k$ standard deviations from the mean is $\leq 1/k^2$.

CS Applications of Probability: Primality Testing

Recall idea of primality testing:

- Choose $b$ between 1 and $n$ at random
- Apply an easily computable (deterministic) test $T(b, n)$ such that
  - $T(b, n) = 1$ (for all $b$) if $n$ is prime.
  - There are lots of $b$’s for which $T(b, n) = 0$ if $n$ is not prime.
- In fact, for the standard test $T$, for at least 1/3 of the $b$’s between 1 and $n$, $T(b, n)$ is false if $n$ is composite

So here’s the algorithm:

**Input** $n$ [number whose primality is to be checked]

**Output** Prime [Want Prime = 1 iff $n$ is prime]

**Algorithm Primality**

- for $k$ from 1 to 100 do
  - Choose $b$ at random between 1 and $n$
  - If $T(b, n) = 0$ return Prime = 0
- endfor
- return Prime = 1.
Probabilistic Primality Testing: Analysis

If \( n \) is composite, what is the probability that algorithm returns \( \text{Prime} = 1 \)?

- \((2/3)^{100} < (2)^{25} \approx 10^{-18}\)
- I wouldn’t lose sleep over mistakes!
- if \( 10^{-18} \) is unacceptable, try 200 random choices.

How long will it take until we find a witness

- Expected number of steps is \( \leq 3 \)

What is the probability that it takes \( k \) steps to find a witness?

- \((2/3)^{k-1}(1/3)\)
- geometric distribution!

Bottom line: the algorithm is extremely fast and almost certainly gives the right results.

Finding the Median

Given a list \( S \) of \( n \) numbers, find the median.

- More general problem:
  \( \text{Sel}(S, k) \) — find the \( k \)th largest number in list \( S \)

One way to do it: sort \( S \), then find \( k \)th largest.

- Running time \( O(n \log n) \), since that’s how long it takes to sort

Can we do better?

- Can do \( \text{Sel}(S, 1) \) (max) and \( \text{Sel}(S, n) \) (min) in time \( O(n) \)

A Randomized Algorithm for \( \text{Sel}(S, k) \)

Given \( S = \{a_1, \ldots, a_n\} \) and \( k \), choose \( m \in \{1, \ldots, n\} \) at random:

- Split \( S \) into two sets
  - \( S^+ = \{a_j : a_j > a_m\}\)
  - \( S^- = \{a_j : a_j < a_m\}\)
- this can be done in time \( O(n) \)
- If \( |S^+| \geq k \), \( \text{Sel}(S, k) = \text{Sel}(S^+, k) \)
- If \( |S^+| = k - 1 \), \( \text{Sel}(S, k) = a_m \)
- If \( |S^+| < k - 1 \), \( \text{Sel}(S, k) = \text{Sel}(S^-, k - |S^+| - 1) \)

This is clearly correct and eventually terminates, since \( |S^+|, |S^-| < |S| \)

- What’s the running time for median (\( k = \lceil n/2 \rceil \)):
  - * Worst case \( O(n^2) \)
  - * Always choose smallest element, so \( |S^-| = 0 \), \( S^+ = |S| - 1 \).
  - * Best case \( O(n) \): select \( k \)th largest right away
  - What happens on average?

Selection Algorithm: Running Time

Let \( T(n) \) be the running time on a set of \( n \) elements:

- \( T(n) \) is a random variable,
- We want to compute \( E(T(n)) \)

Say that the algorithm is in phase \( j \) if it is currently working on a set with between \( n(3/4)^j \) and \( n(3/4)^{j+1} \) elements.

- Clearly the algorithm terminates after \( \leq \lceil \log_{3/4}(1/n) \rceil \) phases.
- Then you’re working on a set with 1 element
- A split in phase \( j \) involves \( \leq n(3/4)^j \) comparisons.

What’s the expected length of phase \( j \)?

- If an element between the 25th and 75th percentile is chosen, we move from phase \( j \) to phase \( j + 1 \)
- Thus, the average # of calls in phase \( j \) is 2, and each call in phase \( j \) involves at most \( n(3/4)^j \) comparisons, so
  \[ E(T(n)) \leq 2n \sum_{j=0}^{\lceil \log_{3/4}(1/n) \rceil} (3/4)^j \leq 8n \]

Bottom line: the expected running time is linear.

- Randomization can help!
Hashing Revisited

Remember hash functions:

- We have a set $S$ of $n$ elements indexed by ids in a large set $U$.
- Want to store information for element $s \in S$ in location $h(s)$ in a “small” table (size $\approx n$).
  - E.g., $U$ consists of $10^{10}$ social security numbers.
  - $S$ consists of 30,000 students.
  - Want to use a table of size, say, 40,000.
- $h$ is a “good” hash function if it minimizes collisions:
  - Don’t want $h(s) = h(t)$ for too many elements $t$.

How do we find a good hash function?

- Sometimes taking $h(s) = s \mod n$ for some suitable modulus $n$ works.
- Sometimes it doesn’t.

Key idea:

- Naive choice: choose $h(s) \in \{0, \ldots, n-1\}$ at random.
- The good news: $\Pr(h(s) = h(t)) = 1/n$.
- The bad news: how do you find item $s$ in the table?

**Theorem:** If $\mathcal{H}$ is universal and $|S| \leq n$, then $E(X_{u,S}) \leq 1$.

**Proof:** Let $X_{uv}(h) = 1$ if $h(u) = h(v)$; 0 otherwise.

- By Property 1 of universal sets of hash function,
  $$E(X_{uv}) = \Pr\{h \in \mathcal{H} : h(u) = h(v)\} = 1/n.$$  

$$X_{u,S} = \sum_{v \neq u} e_S X_{uv}.$$  

$$E(X_{u,S}) = \sum_{v \neq u} e_S E(X_{uv}) \leq |S|/n = 1$$

What this says:

- If we pick a hash function at random from a universal set of hash functions, then the expected number of collisions is as small as we could expect.
- A random hash function from a universal class is guaranteed to be good, no matter how the keys are distributed.

Universal Sets of Hash Functions

Want to choose a hash function $h$ from some set $\mathcal{H}$.

- Each $h \in \mathcal{H}$ maps $U$ to $\{0, \ldots, n-1\}$.

A set $\mathcal{H}$ of hash functions is universal if:

1. For all $u \neq v \in U$:
   $$\Pr\{h \in \mathcal{H} : h(u) = h(v)\} = 1/n.$$
   - The probability that two ids hash to the same thing is $1/n$.
   - Exactly as if you’d picked the hash function completely at random.

2. Each $h \in \mathcal{H}$ can be compactly represented; given $h \in \mathcal{H}$ and $u \in U$, we can compute $h(u)$ efficiently.
   - Otherwise it’s too hard to deal with $h$ in practice.

Why we care: For $u \in U$ and $S \subseteq U$, let

$$X_{u,S}(h) = |\{v \neq u \in S : h(v) = h(u)\}|$$

- $X_{u,S}(h)$ counts the number of collisions with $u$ and an element in $S$ for hash function $h$.
- $X_{u,S}$ is a random variable on $\mathcal{H}$.

We will show that $E(X_{u,S}) = |S|/n$.

Designing a Universal Set of Hash Functions

The theorem shows that if we choose a hash function at random from a universal set $\mathcal{H}$, then the expected number of collisions with an arbitrary element $u$ is 1.

- That motivates designing such a universal set.

Here’s one way of doing it, given $S$ and $U$:

- Let $p$ be a prime, $p \equiv n = |S|$, $p > n$.
  - Can find $p$ using primality testing.
  - Choose $r$ such that $p^r > |U|$.
    - $r \approx \log |U|/\log n$.

- Let $A = \{a_1, \ldots, a_r\} : 0 \leq a_i \leq p - 1$.
  - $|A| = p^r > |U|$.
  - Can identify elements of $U$ with vectors in $A$.

- Let $\mathcal{H} = \{h_\vec{a} : \vec{a} \in A\}$.
- If $\vec{x} = (x_1, \ldots, x_r)$ define
  $$h_\vec{a}(\vec{x}) = \left(\sum_{i=1}^{r} a_i x_i\right) \pmod{p}.$$
**Theorem:** $\mathcal{H}$ is universal.

**Proof:** Clearly there’s a compact representation for the elements of $\mathcal{H}$ – we can identify $\mathcal{H}$ with $\mathcal{A}$.

Computing $h_{\vec{a}}(\vec{x})$ is also easy: it’s the inner product of $\vec{a}$ and $\vec{x}$, mod $p$.

Now suppose that $\vec{x} \neq \vec{y}$.

- For simplicity suppose that $x_1 \neq y_1$
- Must show that $\Pr(\{h \in \mathcal{H} : h(\vec{x}) = h(\vec{y})\}) \leq 1/n$.
- Fix $a_j$ for $j \neq 1$
- For what choices of $a_1$ is $h_{\vec{a}}(\vec{x}) = h_{\vec{a}}(\vec{y})$?
  - Must have $a_1(y_1 - x_1) \equiv \sum_{j \neq 1} a_j(x_j - y_j) \pmod{p}$
  - Since we’ve fixed $a_2, \ldots, a_r$, the right-hand side is just a fixed number, say $M$.
  - There’s a unique $a_1$ that works:
    
    $$a_1 = M(y_1 - x_1)^{-1} \pmod{p}$$

- The probability of choosing this $a_1$ is $1/p < 1/n$.
- That’s true for every fixed choice of $a_2, \ldots, a_r$.

- Bottom line:
  
  $$\Pr(\{h \in \mathcal{H} : h(\vec{x}) = h(\vec{y})\}) \leq 1/n.$$ 

This material is in the Kleinberg-Tardos book (reference on web site).