Mathematics for Computer Science

MIT 6.042J/18.062J

Binomial Distribution & Sampling

What is a Random Variable?

Formally,

$$R : S \to \mathbb{R}$$

Sample space (usually)

Independent Variables

Random variables R,S are independent iff [R = a], [S = b] are independent *events* for all numbers a, b.

Independent Variables

Alternative version:

$\Pr\{R = a \text{ and } S = b\} = \Pr\{R = a\} \cdot \Pr\{S = b\}.$

Mutually Independent RV's

Mutual Independence of random vars $A_1, A_2, \ldots, A_n$:

$\Pr\{A_1=a_1 \text{ and } A_2=a_2 \text{ and } \ldots A_n=a_n\} = \Pr\{A_1=a_1\} \cdot \Pr\{A_2=a_2\} \cdots \Pr\{A_n=a_n\}.$

Independent Variables

k-wise Independence: any k of the variables are mutually independent

(2-wise = pairwise)

Independent Variables

Pairwise Independence sufficient for major applications (in later lecture) which is useful since pairwise holds in important cases where mutual does not.

Density & Distribution

The Probability Density Function of random variable R,

$$PDF_R(a) ::= Pr\{R=a\}$$

Cumulative Distribution Function of R,

$$CDF_R(a) ::= Pr\{R \le a\}$$

Indicator Variables

Indicator variable for event A:

$$I_A ::= \begin{cases} 1 & \text{if } A \text{ occurs,} \\ 0 & \text{if } \overline{A} \text{ occurs.} \end{cases}$$

Distributions

Example:

H_i ::= indicator for a head on the ith coin flip.

Coin may be *biased:*

$pr\{H_i = 1\} = p \neq 1/2$

Binomial Distribution

$H_{n,p}$::= # heads in n mutually independent flips of a p-biased coin.

$$H_{n,p} = H_1 + H_2 + \cdots + H_n$$

Probability space: the 2^n sequences of n H's and T's.

$pr\{Q\} ::= p^{\#H\text{'s in } Q} \cdot (1-p)^{\#T\text{'s}}$

Binomial Distribution

pr{k Heads} =

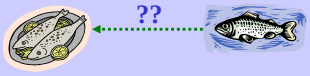
(#k head seqs)

· pr{seq with k H's}

$$PDF_{H_{n,p}}(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Polling & Sampling

Estimate % contaminated fish in Charles River?



Procedure: catch n fish, test each, use %contaminated in catch as *estimate* of %contaminated in whole river

Sampling Questions

Catch 100 fish; what is probability that estimate is within 10% of actual%?

Model as Coin Tosses

p ::= fraction contaminated in river

Fish tested: coin toss with bias p.

Catching n fish: tossing n coins

A ::= fraction contaminated in the sample of 100

Polling using Binomial PDF

A = #"heads"/100 within 10% of p?

$$\Pr\{|A - p| \leq 0.1\} = \Pr\{|H_{100,p} - 100p| \leq 10\}$$

Polling using Binomial PDF

How do we bound this probability when we don't know p?

Lemma: $\Pr\{|H_{n,p} - 100p| \leq 10\}$ is min for $p = 1/2$

Compute the exact probability

$$\Pr\{|A - p| \leq 0.1\} \geq$$

$$\Pr\{|H_{100,1/2} - 50| \leq 10\}$$

$$= \sum_{h=40}^{60} \binom{100}{h} 2^{-100} \geq 0.96$$

Confidence

We can be 96% *confident* that our estimated fraction is with 0.1 of the actual fraction of contaminated fish in the whole river.

Sample size for better estimate

Suppose we want an estimate of the fraction that will be 4% ($\pm$ 0.04) accurate for 95% of the time?
Similar calculation implies need to sample 589 fish.

Confidence – not Probable Reality

Now suppose we sample 589 fish and discover 47 are contaminated.
So we estimate p is 47/589.
It's tempting to say
"the *probability* that
$p = 47/589 \pm 0.04$
is at least 95%"
--Technically not correct!

Confidence

p is the actual fraction of bad fish in the river.

p is *unknown*,

but not a random variable!

Confidence

The possible outcomes of our *sampling procedure* is a random variable. We can say that "the *probability* that our sample fraction will be within $\pm$ 0.04 of the true fraction is at least 95%"

Confidence

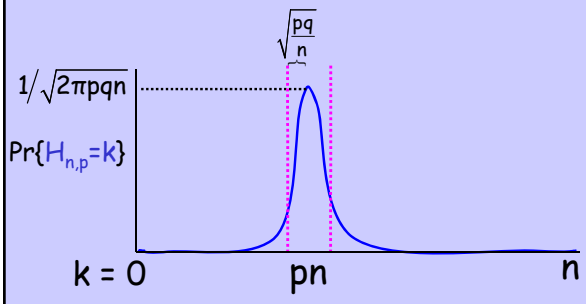
For simplicity we say that

$p = 47/589 \pm 0.04$ at the 95% confidence level

Binomial Approximation

Numerical approximations for $PDF_{H_{n,p}}(\alpha\ n)$, $CDF_{H_{n,p}}(\alpha\ n)$, in Notes 13.

Distribution of Heads



Binomial Approximation

Messy formulas, but easy to compute.

Exact answers for n more than a few 1000 are impossible to compute

(requires arithmetic on million-digit numbers)

Team Problems

Problems 1&2

