Mathematics for Computer Science

MIT 6.042J/18.062J

Sampling & Confidence

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 500 fish; what is probability that estimate is within 0.1 of the actual fraction?

Model as Coin Tosses

p ::= fraction contaminated in river

test a fish $\leftrightarrow$ toss bias p coin

catch n fish $\leftrightarrow$ toss n coins

A_n ::= fraction contaminated in the sample of n

Pairwise Independent Sampling

$$\Pr\left\{\left|A_{500} - p\right| > 0.1\right\} \leq \frac{1}{500}\left(\frac{1/2}{0.1}\right)^2$$

$$n = 500, \quad \mu = p, \quad \delta = 0.1$$

$$\text{worst } \sigma = \frac{1}{2}$$

Pairwise Independent Sampling

$$\Pr\left\{\left|A_{500} - p\right| > 0.1\right\} \leq \frac{1}{500}\left(\frac{1/2}{0.1}\right)^2$$

$$n = 500, \quad \mu = p, \quad \delta = 0.1$$

$$\Pr\left\{\left|A_{500} - p\right| \leq 0.1\right\} > 0.95$$

Confidence in our estimate

With probability 0.95 our estimated fraction will be within 0.1 of the actual fraction of contaminated fish in the whole river.

Sampling using Binomial PDF

Better estimate: A_n is $\frac{B_{n,p}}{n}$

$$\Pr\left\{\left|A_n - p\right| \le \delta\right\} = \Pr\left\{\left|B_{n,p} - np\right| \le \delta n\right\}$$

Sampling using Binomial PDF

Better estimate:

$$n = 500, \quad \delta = 0.06$$

$$\Pr\left\{\left|B_{500,p} - 500p\right| \le 30\right\}$$

Sampling using Binomial PDF

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

Lemma: $\Pr\left\{\left|B_{n,p} - np\right| \le \delta n\right\}$ is min when p = 1/2

Sampling using Binomial PDF

$$\Pr\left\{220 \le B_{500,1/2} \le 280\right\}$$

$$\Pr\left\{\left|B_{500,1/2} - 250\right| \le 30\right\}$$

Sampling using Binomial PDF

$$\Pr\left\{220 \le B_{500,1/2} \le 280\right\}$$

$$= \sum_{i=220}^{280} \binom{500}{i} 2^{-500}$$

$$\ge 0.99$$

Confidence in our estimate

We can actually be 99% confident that our estimated fraction is with 0.06 of the true fraction of contaminated fish in the whole river.

Confidence –not Probable Reality

Now suppose we sample 500 fish and discover 230 are contaminated.
So we estimate p is 230/500 = 0.46
It's tempting to say

"the probability that
$p = 0.46 \pm 0.06$
is at least 0.99"

--technically wrong!

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 sampling process will yield a fraction that is $\pm$ 0.06 of the true fraction at least 0.99"

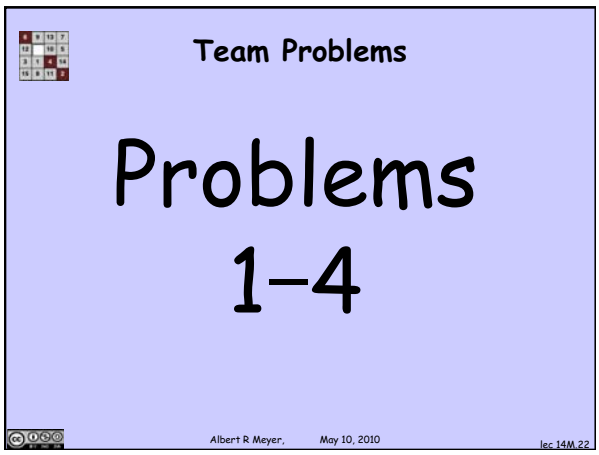
Confidence

for simplicity we say that

$p = 0.46 \pm 0.06$
at the
99% confidence level

Confidence

Moral: when you are told that some fact holds at a high confidence level, remember that a random experiment lies behind this claim. Ask yourself "what experiment?"


Team Problems
Problems
1–4
Albert R Meyer, May 10, 2010
lec 14M.22

6.042J / 18.062J Mathematics for Computer Science
Spring 2010