Mathematics for Computer Science

MIT 6.042J/18.062J

Sampling & Confidence

Pairwise Independent Sampling

Theorem:

Let $R_1,\ldots,R_n$ be pairwise independent random vars with the same finite mean μ and variance σ^2. Let $A_n ::= (R_1 + R_2 + \cdots + R_n)/n$. Then

$$\Pr[\,|A_n - \mu| > \delta\,] \le \frac{1}{n}\left(\frac{\sigma}{\delta}\right)^2$$

Sampling

coliform count in Charles River for swimming

EPA requires

average CMD $<$ 200

(Coliform Microbial Density)

Sampling Questions

Make 32 measurements of CMD at random times and locations

Sampling Questions

A few of the 32 counts turn out to be $>$ 200 but their average is 180. Convince the EPA that avg in whole river is $<$ 200?

Sampling Questions

That is, convince EPA that the estimate based on 32 samples is within 20 of the actual average?

Sampling parameters

c ::= actual average CMD in river

CMD sample $\leftrightarrow$ ran var with μ = c

n samples $\leftrightarrow$ n mutually indep ran vars with μ = c

A_n ::= avg of the n CMD samples

Pairwise Independent Sampling

$$\Pr[|A_n - \mu| > \delta] \leq \frac{1}{n}\left(\frac{\sigma}{\delta}\right)^2$$

$$n = 32, \quad \mu = c, \quad \delta = 20$$

Pairwise Independent Sampling

$$\Pr[|A_{32} - c| > 20] \le \frac{1}{32}\left(\frac{\sigma}{20}\right)^2$$

$n = 32, \quad \mu = c, \quad \delta = 20$

?? don't know σ

Bound for σ

$$\Pr[|A_{32} - c| > 20] \le \frac{1}{32}\left(\frac{\sigma}{20}\right)^2$$

$n = 32, \quad \mu = c, \quad \delta = 20$

suppose L is max possible difference of samples

worst $\sigma = \frac{L}{2} = 50$

Pairwise Independent Sampling

$$\Pr[|A_{32} - c| > 20] \le \frac{1}{32}\left(\frac{25}{20}\right)^2 < 0.05$$

$$\Pr[|A_{32} - c| \le 20] > 0.95$$

Confidence —not Probable Reality

tempting to say:

~~"the probability that $c = 180 \pm 20$ is at least 0.95"~~

--technically wrong!

Confidence

c is the actual average in the river.

c is unknown,

but not a random variable!

Confidence

The possible outcomes of our sampling process is a random variable. We can say that the "*probability* that our sampling process will yield an average that is $\pm$ 20 of the true average at least 0.95"

Confidence

Tell the EPA that with probability 0.95 our estimate method for avg CMD will be within 20 of the actual avg, c, in the river.

Confidence

For simplicity we say that

c = 180 $\pm$ 20 at the 95% 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?"

Confidence

Moral: Also ask "Why am I hearing about this particular experiment? How many others were tried and not reported?"

See http://xkcd.com/882/