



# Deviation of Repeated Trials



## Jacob D. Bernoulli (1659 – 1705)

Even the stupidest man---by some instinct of nature *per se* and by no previous instruction (this is truly amazing) -- knows for sure that the more observations ...that are taken, the less the danger will be of straying from the mark.

---*Ars Conjectandi* (The Art of Guessing), 1713\*

\*taken from Grinstead & Snell,  
[http://www.dartmouth.edu/~chance/teaching\\_aids/books\\_articles/probability\\_book/book.html](http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html)  
*Introduction to Probability*, American Mathematical Society, p. 310.



## Jacob D. Bernoulli (1659 – 1705)

It certainly remains to be inquired whether after the number of observations has been increased, **the probability...of obtaining the true ratio...finally exceeds any given degree of certainty**; **or** whether the problem has, so to speak, its own asymptote---that is, whether **some degree of certainty is given which one can never exceed**.



## Weak Law of Large Numbers

$A_n ::= \underline{\text{Average of } n \text{ trials}}$

$\mu ::= E[\text{trial}]$

$$\lim_{n \rightarrow \infty} \left[ \text{Pr} \{ |A_n - \mu| \leq \varepsilon \} \right] = 1$$



## Jacob D. Bernoulli (1659 – 1705)

Therefore, this is the problem which I now set forth and make known after I have pondered over it for **twenty years**. Both its **novelty** and its very **great usefulness**, coupled with its just as **great difficulty**, can exceed in weight and value all the remaining chapters of this thesis.



## Deviation from the Mean

$\Pr\{\text{observed value far from expected value}\}$   
is *SMALL*



## Deviation from the Mean

*Observed value* means random variable,  $R$ .  
*far from* may mean distance or amount above (or below)



## Chebychev Bound

$$\Pr\{\underbrace{|R-\mu|}_{\text{distance}} > \underbrace{a}_{\text{far}}\} \leq \frac{\sigma^2}{\underbrace{a^2}_{\text{small}}}$$



## Chebychev Bound

Requires  $\sigma$ . Can find  $\sigma$  by:

- nice theory (pairwise independent variance additivity)
- experimental sampling



## Repeated Trials

$X_1, \dots, X_n$  independent  
with mean,  $\mu$ , and variance  $\sigma^2$   
 $A_n ::= (X_1 + \dots + X_n)/n$   
 $E[A_n] = n\mu/n = \mu$



## Repeated Trials

$\text{Var}[X_1 + \dots + X_n] = n\sigma^2$   
(by independence)  
 $\text{Var}[A_n] = n\sigma^2/n^2 = \frac{\sigma^2}{n}$   
*decreases with # trials*



## Repeated Trials

So by Chebychev

$$\Pr\{|A_n - \mu| > \varepsilon\} \leq (\sigma/\varepsilon)^2 \cdot \frac{1}{n}$$

$$\Pr\{|A_n - \mu| \leq \varepsilon\} \geq 1 - \underbrace{\frac{(\sigma/\varepsilon)^2}{n}}_{\rightarrow 0}$$



## Weak Law of Large Numbers

Therefore

$$\lim_{n \rightarrow \infty} \left[ \Pr\{|A_n - \mu| \leq \varepsilon\} \right] = 1$$

QED



## In-Class Problem

# Problem 1



## Polling/Sampling

Estimate % contaminated fish in Charles River?



*Procedure:* catch  $n$  fish, test each, use % contaminated in catch as estimate



## Two Typical Questions



- Catch 100 fish; probability that estimate is within 10% of actual %?
- To be 95% confident that estimate within 4%, how many fish to catch?



## Model as Coin Tosses



Say actual fraction contaminated is  $p$ .

Fish tested: coin toss with bias  $p$ .

Catching  $n$  fish: tossing  $n$  coins

$A_{n,p} ::=$  fraction contaminated in sample



### Bernoulli Trials



$F_i ::=$  indicator for contamination of  $i$ th fish caught

$$B_{n,p} = F_1 + \dots + F_n$$

$$A_{n,p} = \frac{B_{n,p}}{n}$$



### Bernoulli Trials



$$E[F_i] = p, \text{ Var}[F_i] = pq$$

$$E[B_{n,p}] = nE[F_i] = np$$

$$\text{Var}[B_{n,p}] = n\text{Var}[F_i] = npq$$

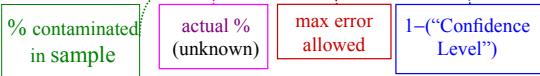
$$q ::= 1 - p$$



### Probability of Deviation

#### Chebychev Bound

$$\Pr\{|A_{n,p} - p| > x\} \leq \frac{\text{Var}[A_n]}{x^2}$$



### The Variance of $A_{n,p}$

We know

$$\text{Var}[A_{n,p}] = \frac{npq}{n^2} = \frac{pq}{n}$$

with  $n = 100$ ,  $x = 0.1$ ,

$$\Pr\{\text{deviation} > 10\%\} \leq \frac{pq}{(0.1)^2 100} = pq$$



### Confidence Level with Sample of 100

*Circularity:* estimating  $p$  using  $p$ ??

Avoid by using *worst case* variance (at  $p = 1/2$ ). So

$$\Pr\{\text{deviation} \leq 10\%\} \geq 0.75$$



### Confidence Level with Sample of 100

$$\Pr\{\text{deviation} \leq 10\%\} \geq 0.75:$$

We can be **75% confident** that estimate is with  $\pm 10$  of actual % contaminated in Charles River.



### How Big a Sample for 95% Confidence?

$$\begin{aligned} \Pr\{\text{deviation} > 4\%\} &\leq \frac{pq}{(0.04)^2 n} \\ &\leq \frac{\frac{1}{2}(1-\frac{1}{2})}{(0.04)^2 n} \\ &= \frac{156.25}{n} \end{aligned}$$



### How Big a Sample for 95% Confidence?

To be 95% confident, need

$$\begin{aligned} \frac{156.25}{n} &\leq 0.05 \\ n &\geq 3125 \end{aligned}$$



### Confidence – not Probable Reality

OK, sample 3200 fish and discover 640 are contaminated: estimate  $p$  is  $640/3200 = 0.2$

It's tempting to say

~~$$\Pr\{|p - 0.2| \leq 0.04\} \geq 0.95$$~~

but that is a **misstatement**:

We **can't** talk about the **probability** that  $p$  has a particular value.



### Confidence – not Probable Reality

$p$  is the *actual* % contaminated

$p$  is *unknown*,

but **not a random variable!**

Our *estimate* is a random variable:

we *can* talk about the probability (**confidence**) that *it* is correct.



### In Class Problem

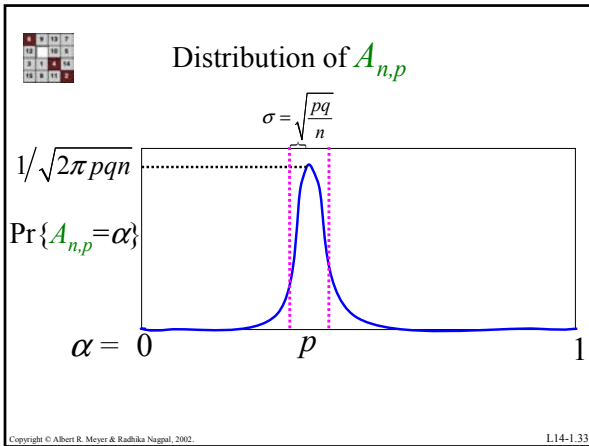
## Problems 2 & 3



### Better Polling using Binomial PDF

We know the PDF of  $A_n$   $A_{n,p} = \frac{B_{n,p}}{n}$

$$\begin{aligned} \Pr\left\{A_{n,p} = \frac{k}{n}\right\} &= \Pr\{B_{n,p} = k\} \\ &= \binom{n}{k} p^k q^{n-k} \end{aligned}$$



Compute the exact probabilities

For  $n = 100$ ,

$$\Pr \{\text{within 10\% of mean}\}$$

$$= \Pr \{|A_{n,p} - p| < 0.1\}$$

$$= \Pr \{|B_{n,p} - pn| < (0.1)n\}$$

$$= \Pr \{n(p - 0.1) \leq B_{n,p} \leq n(p + 0.1)\}$$

Copyright © Albert R. Meyer & Radhika Nagpal, 2002. L14-1.34

Exact Confidence with **Sample of 100**

*Same circularity: estimate  $p$  using  $p$ ??*

Again, worst case is at  $p = 1/2$ . So

$$\Pr \{\text{within 10\% of mean}\}$$

$$\leq \sum_{k=40}^{60} \binom{100}{k} 2^{-n} \geq 0.96$$

Copyright © Albert R. Meyer & Radhika Nagpal, 2002. L14-1.35

Exact Confidence with **Sample of 100**

So, actual confidence of being with 10% with sample of **100 fish** is  $\geq 96\%$  instead of **75%** just using Chebychev.

Copyright © Albert R. Meyer & Radhika Nagpal, 2002. L14-1.36

Exact Sample Size for 95% Confidence

Similarly, to be **95%** confident of within **4%** need sample size  $n \geq 599$  instead of **3125** just using Chebychev.

Copyright © Albert R. Meyer & Radhika Nagpal, 2002. L14-1.37

In Class Problem

# Problem 4

Copyright © Albert R. Meyer & Radhika Nagpal, 2002. L14-1.38