# STAT 513/413: Lecture 12
# Classical Monte Carlo

(probabilistic theorems)

# Recall: classical Monte Carlo

Classical Monte Carlo deals with the estimation of expected values, with probabilities also falling under this case. This particular problem allows for more than merely intuitive approach: it is possible to assess the precision more rigorously via existing mathematical theorems.

Notation: we are still after quantity $\vartheta$



we still compute its estimate $T_N$

And the subscript N still indicates how many random numbers were used in the computation of $T_N$: the higher N, the more numbers used



But now we have $\vartheta = E(T_N)$

Or even more precisely: we are after $\vartheta = E(X)$ where X has the same distribution as random numbers $X_1, X_2, \ldots, X_N$; that is

$$\vartheta = E(X_1) = E(X_2) = \ldots = E(X_N)$$

This particular case covers quite a lot of instances; in particular...

# Probabilities come as a special case of this scheme

We are interested in $p = P(E)$, the probability of some event $E$

We define an *indicator random variable* $Y$ to be

$$Y = \begin{cases} 1 & \text{if event } E \text{ happens} \\ 0 & \text{if event } E \text{ does not happen} \end{cases}$$

The expected value of $Y$ is



$$E(Y) = 1P[Y = 1] + 0P[Y = 0] = P[Y = 1] = p$$



Clear?

# We estimate expected values by averages

(or means by means - but this formulation may confuse some)

We make (that is, generate) N replications of X

$$X_1, X_2, \ldots, X_N$$

are considered independent and having the same distribution as X

Then



$$\vartheta = E(X_1) = E(X_2) = \ldots = E(X_N)$$



and



$$T_N = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Note that then $T_N$ is an unbiased estimator of $\vartheta$

$$E(T_N) = E(\bar{X}) = \frac{1}{N}\sum_{i=1}^{n} E(X_i) = \frac{1}{N}(N\vartheta) = \vartheta$$

Why does Monte Carlo work in this case? A classical result:

# Jacob Bernoulli (1655-1705), Ars Conjectandi (1713)

**Law of large numbers** (in the version proved later by Kolmogorov)

If $X_1, X_2, \ldots, X_N$ are independent random variables with the same distribution, such that the expected value of all of them (or of one: same distribution!) [exists and] is finite

$$\mu = \mathsf{E}(X_1) = \mathsf{E}(X_2) = \ldots$$



then the average $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$



converges to the mean $\mu$ for $n \to \infty$, *in probability*:



for every $\varepsilon > 0$, $\quad \mathsf{P}\left[|\bar{X} - \mu| \geqslant \varepsilon\right] \to 0$ with $N \to \infty$

If our estimated value is $\vartheta = \mu$ and its estimate is $T_N = \bar{X}$, then this establishes consistency, right?

(A fine point: mere $T_N$ being unbiased could not be enough, we still need something more... and there are also another fine points for math enthusiasts)

# For math enthusiasts

# Probabilities are again a special case

We are interested in $p = P(E)$, the probability of some event $E$

We define *indicator random variables* to be

$$Y_i = \begin{cases} 1 & \text{if event } E \text{ happens in the } i\text{-th repetition (trial)} \\ 0 & \text{if event } E \text{ does not happen in the } i\text{-th repetition} \end{cases}$$

If the repetitions are independent, then also $Y_i$ are independent



and their common mean is



$$E(Y_i) = 1P[Y_i = 1] + 0P[Y_i = 0] = P[Y_i = 1] = p$$



So, then we calculate $$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{\text{card}\{i : Y_i = 1\}}{N}$$

to estimate (that is, compute) $p$

The consistency follows from the Law of large numbers

# Let us try it on toy examples

Recall: what is the probability of obtaining 2 or less heads in 3 coin tosses? (Answer: 0.875)

A problem for expected values proper (not probabilities): what is the expected value of heads in 3 coin tosses?

(Well, everybody knows that 2 - but we need a simple problem to start with.)

# First: it works for estimating a probability

(We should have kept the seed. Just in case.)

```
> mean(apply(replicate(1000,rbinom(3,1,0.5)),2,sum) <= 2)
[1] 0.858
> mean(apply(replicate(1000,rbinom(3,1,0.5)),2,sum) <= 2)
[1] 0.854
> mean(apply(replicate(1000,rbinom(3,1,0.5)),2,sum) <= 2)
[1] 0.888
> mean(apply(replicate(10000,rbinom(3,1,0.5)),2,sum) <= 2)
[1] 0.8806
> mean(apply(replicate(10000,rbinom(3,1,0.5)),2,sum) <= 2)
[1] 0.8761
> mean(apply(replicate(1000000,rbinom(3,1,0.5)),2,sum) <= 2)
[1] 0.875105
> mean(apply(replicate(1000000,rbinom(3,1,0.5)),2,sum) <= 2)
[1] 0.874815
> mean(apply(replicate(1000000,rbinom(3,1,0.5)),2,sum) <= 2)
[1] 0.875326
> mean(apply(replicate(10000000,rbinom(3,1,0.5)),2,sum) <= 2)
[1] 0.8748622
```

# Works for the expected values as well

# The evaluation of precision

Classical Monte Carlo has one big advantage over the intuitive approach: we can evaluate the precision in an exact way

For starters, let us - *tentatively!* - measure it once again by the mean squared error

$$\mathsf{E}[(\mathsf{T}_\mathsf{N} - \vartheta)^2]$$



(Why? Because it is quite reasonable and we are able to handle the mathematics - compared to other possible choices.)



An important case is when $\mathsf{E}(\mathsf{T}_\mathsf{N}) = \vartheta$ for all $\vartheta$ under consideration

Recall: such estimators are called *unbiased*



For unbiased estimator, the mean squared error is equal to their

variance $\quad \mathsf{E}[(\mathsf{T}_\mathsf{N} - \vartheta)^2] = \mathsf{E}[(\mathsf{T}_\mathsf{N} - \mathsf{E}(\mathsf{T}_\mathsf{N}))^2] = \mathsf{Var}(\mathsf{T}_\mathsf{N})$

Sometimes, we may also look at the square root of variance

standard deviation $\quad \sqrt{\mathsf{Var}(\mathsf{T}_\mathsf{N})}$

# Performance guarantees II: precision via variance

Variance has direct consequences for the precision via the **Chebyshev(-Bienaymé) inequality**, which in its general form says that for any random variable $Z$ with both $E(Z)$ and $\mathrm{Var}(Z)$ finite

$$P\left[|Z - E(Z)| \geqslant \varepsilon\right] \leqslant \frac{\mathrm{Var}(Z)}{\varepsilon^2}$$



Thus, if $T_N$ is an unbiased estimator of $\vartheta$, then for every $\varepsilon > 0$,



$$P\left[|T_N - \vartheta| \geqslant \varepsilon\right] \leqslant \frac{\mathrm{Var}(T_N))}{\varepsilon^2}$$



So, the variance (or its square root, standard deviation) is an important indicator of the precision of the estimate $T_N$ of $\vartheta$

And it is evaluated *explicitly* via Chebyshev inequality, so *no mean square error anymore!* We can do better here!

A side product: if $T_N$ is unbiased and we can show mathematically that $\mathrm{Var}(T_N) \to 0$ for $N \to \infty$, , then consistency for $T_N$ immediately follows

# So, again the canonical case

Let $X_1, X_2, \ldots, X_N$ be random variables

with the same distribution

they (all) have (the same) expected value $\mu$

and they also have the same variance $\sigma^2$



What is the variance of $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$? Standard calculation...





$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{N^2} N\,\sigma^2 = \frac{\sigma^2}{N}$$

This may serve as the measure of precision

- only we need to know $\sigma^2$

(which we may not, however)

# Performance guarantees II: $\bar{X}$ via Chebyshev

Once we know $\sigma^2$, we have a bound on the probability that our results deviates from $\mu$ by more than $\varepsilon$:

Knowing that $E(\bar{X}) = \vartheta$, we apply the Chebyshev inequality to $\bar{X}$

$$P[|\bar{X} - \vartheta| \geqslant \varepsilon] \leqslant \frac{\mathrm{Var}(\bar{X})}{\varepsilon^2} = \frac{\sigma^2}{N\varepsilon^2}$$



This gives us a lucid expression of precision (as long as we know $\sigma^2$)



Also, note that, as $\dfrac{\sigma^2}{N\varepsilon^2} \to 0$ with $N \to \infty$ (whatever $\sigma^2$ is)



we have automatically consistency here

Return to precision: what if $\sigma^2$ is not known (as is often the case)?

# An upper bound can help

First, recall “the other formula” for the variance: this yields

$$\mathrm{Var}(Z) = \mathrm{E}(Z^2) - (\mathrm{E}(Z))^2 \leqslant \mathrm{E}(Z^2)$$

Note: as $\mathrm{E}(Z)$ is usually finite (we are estimating $\mathrm{E}(Z)$, so if it would not be properly defined, it would be hopeless), the crux is in $\mathrm{E}(Z^2)$



And geeks know: $\mathrm{E}(Z^2) < \infty$ implies not only $\mathrm{Var}(Z) < \infty$, but also that $\mathrm{E}(|Z|) < \infty$, that is, $\mathrm{E}(Z)$ exists and is finite



Recall some very simple, but also very useful principles: if



for random variables $U$ and $V$ we have $U \leqslant V$ (with probability one)

then $\mathrm{E}(U) \leqslant \mathrm{E}(Y)$

In particular, if $X^2 \leqslant Y^2$ (with probability one)

then $\mathrm{E}(X^2) \leqslant \mathrm{E}(Y^2)$

(Beware: nothing like that necessarily follows for variances)

# A very often used upper bound on variance

If a random variable $Z$ has values in $[a, b]$ (with probability one)

$$\text{then} \qquad \mathrm{Var}(Z) \leqslant \frac{(b-a)^2}{4} \qquad \text{(try to prove that!)}$$

Combined with the Chebyshev inequality for $\bar{X}$, we get



$$P[|\bar{X} - \vartheta| \geqslant \varepsilon] \leqslant \frac{\mathrm{Var}(\bar{X})}{\varepsilon^2} \leqslant \frac{(b-a)^2}{4N\varepsilon^2}$$





whenever $\vartheta = E(\bar{X})$ and all $X_i$ forming $\bar{X}$ have values in $[a, b]$

(with probability one)

# Probabilities are a special case

In particular, for probabilities:

we define $Y_i$ with values 0 and 1 as above

these $Y_i$ happen to have expected value equal to $p$ and

$$\mathrm{Var}(Y_i) = (1-p)(0-p)^2 + p(1-p)^2 = p(1-p)^2 - p^2(1-p) = p(1-p)$$



Given that $p$ is in $[0,1]$, we have an inequality $p(1-p) \leqslant 1/4$



And therefore $\mathrm{Var}(\bar{Y}) = \dfrac{p(1-p)}{N} \leqslant \dfrac{1}{4N}$



So, for 0-1 $Y_i$, we get now from the Chebyshev inequality

$$P[|\bar{Y} - \mu| \geqslant \varepsilon] \leqslant \frac{\sigma^2}{N\varepsilon^2} \leqslant \frac{1}{4N\varepsilon^2}$$

# Return to precision: how is this used?

We first select $\varepsilon$ to express precision we would like to achieve

for instance, if we want 2 decimal digits, then $\varepsilon = 0.005$

(or slightly less)

Then, we select "margin of error" $\delta$: if we are to be outside the bounds with some probability, what probability we would tolerate

for instance, $\delta = 0.01$



We are interested in $N$ that will guarantee us chosen $\varepsilon$ and $\delta$



The Chebyshev inequality says that it will be equal to or larger than



$N$ satisfying $\dfrac{\sigma^2}{N\varepsilon^2} = \delta$ which is $N = \dfrac{\sigma^2}{\delta\varepsilon^2}$

if estimating a probability we need $N \geqslant \dfrac{1}{4\delta\varepsilon^2} = 1000000$

Note: to increase precision by one decimal digit

we have to divide $\varepsilon$ by 10, hence multiply $N$ by 100

and to decrease "margin of error" by one decimal digit

we have to divide $\delta$ by 10, hence multiply $N$ by 100

# Performance guarantees II: $\bar{X}$ for bounded $X_i$

**Hoeffding inequality**. Suppose that $X_1, X_2, \ldots, X_N$ are independent random variables, each $X_i$ bounded by an interval $[a_i, b_i]$ (that is, $X_i$ takes values in $[a_i, b_i]$ with probability one).

If $\vartheta = \mathsf{E}(\bar{X})$, then

$$P[|\bar{X} - \vartheta| \geqslant \varepsilon] \leqslant 2e^{-\frac{2N^2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$



In particular, if all $[a_i, b_i] = [a, b]$, then



$$P[|\bar{X} - \vartheta| \geqslant \varepsilon] \leqslant 2e^{-\frac{2N\varepsilon^2}{(b-a)^2}}$$



And then, if $[a, b] = [0, 1]$,

$$P[|\bar{X} - \vartheta| \geqslant \varepsilon] \leqslant 2e^{-2N\varepsilon^2}$$

(the latter case is also known as the **Chernoff bound**)

# The use of this one

Suppose that we have $[a_i, b_i] = [a, b]$ - and as before

we want precision $\varepsilon$ and the "margin of error" $\delta$

To have $N$ guaranteeing this, it has to be equal or greater than

the one satisfying $$N = -\frac{(b-a)^2}{2\varepsilon^2} \log\left(\frac{\delta}{2}\right)$$

for indicator variables it is $$N = -\frac{1}{2\varepsilon^2} \log\left(\frac{\delta}{2}\right)$$





So, for $\delta = 0.01$ and $\varepsilon = 0.005$ we get $N \geqslant 105967$



Note: to increase precision by one decimal digit

we have to divide $\varepsilon$ by 10, hence multiply $N$ by 100

this is the same of the Chebyshev

But, decreasing "margin of error" by one decimal digit:

dividing $\delta$ by 10 means *adding* $\dfrac{\log 10}{2\varepsilon^2} \approx \dfrac{1.1513}{\varepsilon^2}$ to $N$

That looks cheaper...

# Some remarks

Note first:

All the mentioned inequalities bound some *probability* from above

thus, to be nontrivial, the bound has to be $< 1$

Some examples now:

For $\delta = 0.01$ and $\varepsilon = 0.005$ we get $N \geqslant 105967$ by the Hoeffding



In the same situation, for $N = 1000000$ and $\varepsilon = 0.005$, Hoeffding says we $\delta \leqslant 3.8575 \times 10^{-22}$



Great, this is way much we hoped for, but nonetheless fine... Can we expect also better precision here - maybe with slightly higher probability of error?



If we try $\varepsilon = 0.0005$, still with $N = 1000000$, we actually obtain the bound $\delta \leqslant 1.213$ - which does not say anything nontrivial: it is a bound on a probability which is always $< 1$

To achieve $\delta = 0.01$ for $\varepsilon = 0.0005$, we need $N \geqslant 10596634$

And also: Hoeffding works for *bounded* $X_i$

but in that case we also have an upper bound on variance!

# Chebyshev strikes back

If $X_i$ is within $[a, b]$ (with probability one)

then Chebyshev and the inequality on variance together yield

$$P[|\bar{X} - \vartheta| \geqslant \varepsilon] \leqslant \frac{\mathrm{Var}(\bar{X})}{\varepsilon^2} \leqslant \frac{(b-a)^2}{4N\varepsilon^2}$$

and Hoeffding inequality gives

$$P[|\bar{X} - \vartheta| \geqslant \varepsilon] \leqslant 2e^{-\frac{2N\varepsilon^2}{(b-a)^2}} = e^{-\frac{1}{2}\frac{4N\varepsilon^2}{(b-a)^2}}$$

So, the bounds are respectively $\delta = \dfrac{(b-a)^2}{4N\varepsilon^2}$ and $2e^{-\frac{1}{2\delta}}$



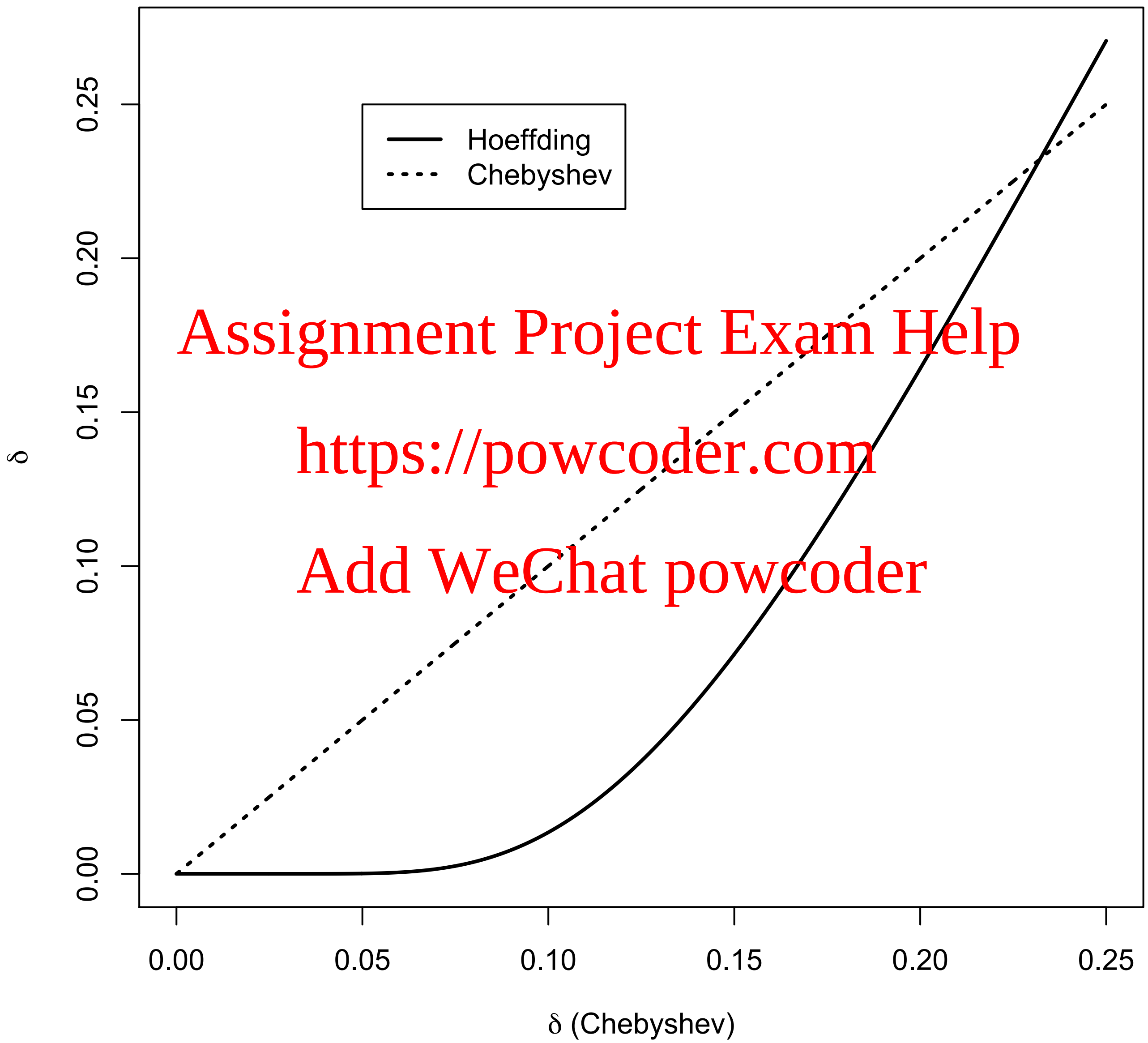
Analyzing this, we figure out that for small $\delta$, the second bound is better; the situation reverses at about $\delta \approx 0.2322026$

However, we are interested in *small* $\delta$

and in such situation is Hoeffding better

Finally: for given precision $\varepsilon$, we decrease the “margin of error” by increasing $N$ - but we may also do it by decreasing $(b-a)^2$ or $\sigma^2$

# A pictorial comparison of bounds



## The comparison of the required $N$'s

For a given precision $\varepsilon$, the required $N$ is greater or equal than

$$\frac{(b-a)^2}{4\delta\varepsilon^2} = \frac{(b-a)^2}{2\varepsilon^2}\left(\frac{1}{2\delta}\right) \qquad \text{by the Chebyshev and bound}$$

$$-\frac{(b-a)^2}{2\varepsilon^2}\log\left(\frac{\delta}{2}\right) \qquad \text{by the Hoeffding}$$



Once $\varepsilon$ is given, then $(b-a)^2/(2\varepsilon^2)$ is given as well



and we can compare $\dfrac{1}{2\delta}$ and $-\log\left(\dfrac{\delta}{2}\right)$

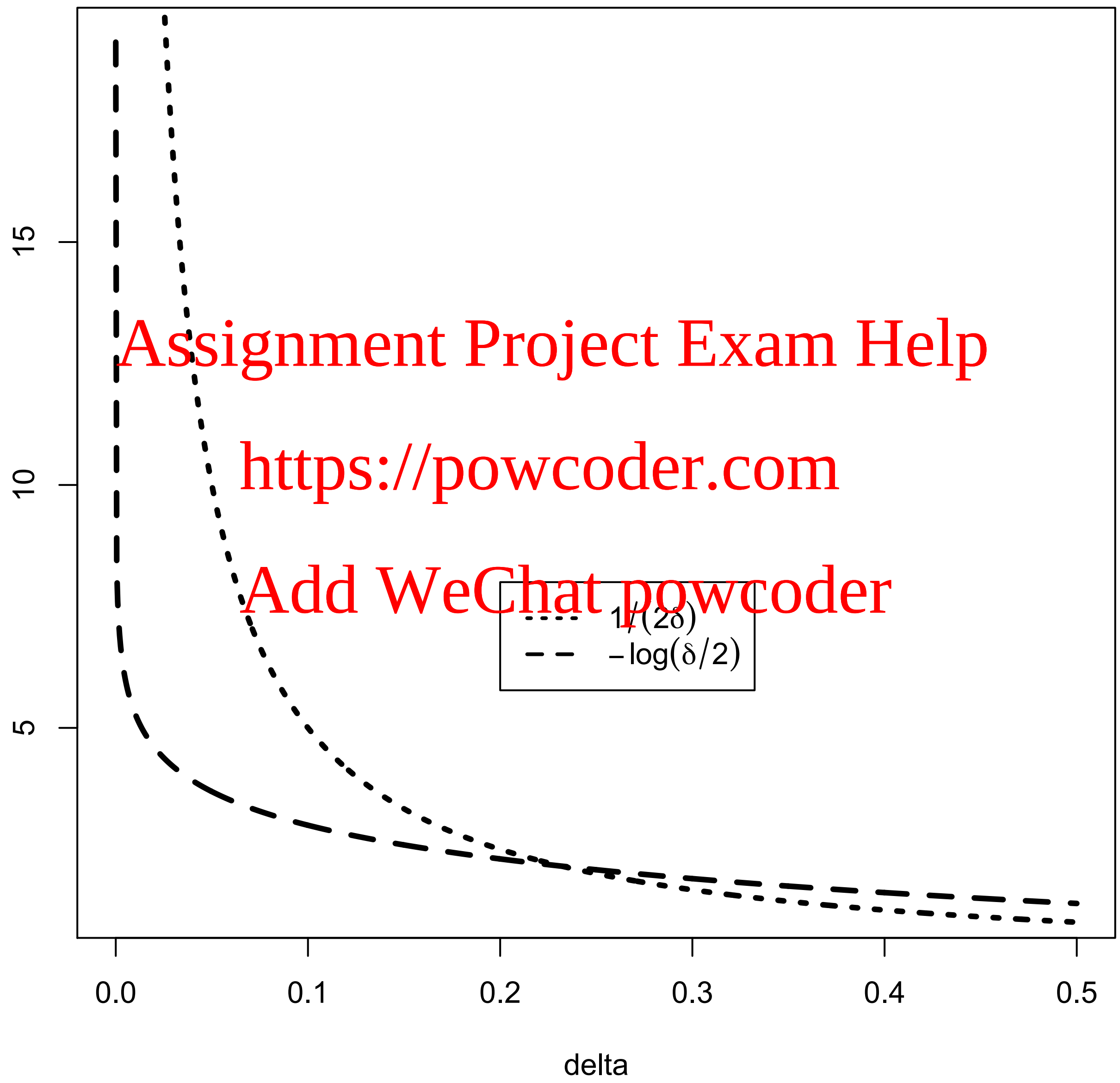


Again, they are equal for $\delta \approx 0.2322026$

and then for smaller $\delta$ Hoeffding gives smaller $N$

in fact, *much* smaller for very small $\delta$

# The comparison of required $N$ for probabilities



## Sometimes, however, $\sigma^2$ can only be estimated

The well-known estimate of $\sigma^2$ is $\quad s^2 = \frac{1}{N-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

and Var $\bar{X}$ is then estimated by $\quad \frac{s^2}{N} = \frac{\frac{1}{N-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}{N}$





Note: sometimes instead of $s^2 = \frac{1}{N-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$



they use $\quad \frac{1}{N}\sum_{i=1}^{n}(X_i - \bar{X})^2$

(compare textbook, pp. 151-152 and also pp. 52-53)

The reason is that - in the present context - N is usually large...

...so $\frac{1}{N}$ differs very little from $\frac{1}{N-1}$

# In fact, it may work that way for probabilities

Suppose again that $Y_1, Y_2, \ldots, Y_N$ are independent random variables

such that $P[Y_1 = 1] = p$ and $P[Y_1 = 0] = 1 - p$

We know that $\vartheta = p = E(Y_i)$

and also that $\sigma^2 = \mathrm{Var}(Y_i) = p(1-p)$

We estimate $\vartheta$ by $\bar{Y} = \dfrac{1}{N}\sum_{i=1}^{N} Y_i$



So we could perhaps also estimate $\mathrm{Var}(\bar{Y})$ by $\bar{Y}(1-\bar{Y})$

but we could do it also by $\dfrac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$

and perhaps also by $\dfrac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$

Are those related in some way?

# Let us do a computational experiment first

```
> tosses = floor(2*runif(20)) ; tosses
 [1] 0 1 0 1 0 0 1 0 0 1 1 0 0 1 1 0 1 1 1 0
> mean(tosses)*(1-mean(tosses))
[1] 0.25
> var(tosses)
[1] 0.2631579
> var(tosses)*19/20
[1] 0.25
```



Maybe only a coincidence... OK, another one:



```
> tosses = floor(2*runif(30))
> tosses
 [1] 1 1 0 0 0 1 1 0 0 0 0 0 0 1 1 1 0 1 0 0 0 0 0 1 1 0 0 1 1 0
> mean(tosses)*(1-mean(tosses))
[1] 0.24
> var(tosses)
[1] 0.2482759
> var(tosses)*29/30
[1] 0.24
```



So, a rule here?

## Your theory for that: