ORIE 4580/5580: Simulation Modeling and Analysis

ORIE 5581: Monte Carlo Simulation

Unit 3: Intro to Monte Carlo Simulation

Sid Banerjee

School of ORIE, Cornell University

expectation and variance of sums of rvs

linearity of expectation

for any rvs X and Y, and any constants $a, b \in \mathbb{R}$

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

note: no assumptions! (in particular, does not need independence)

- for general X, Y

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(XY)$$

'negatively correlated' can be negative!

- when X and Y are independent

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$$

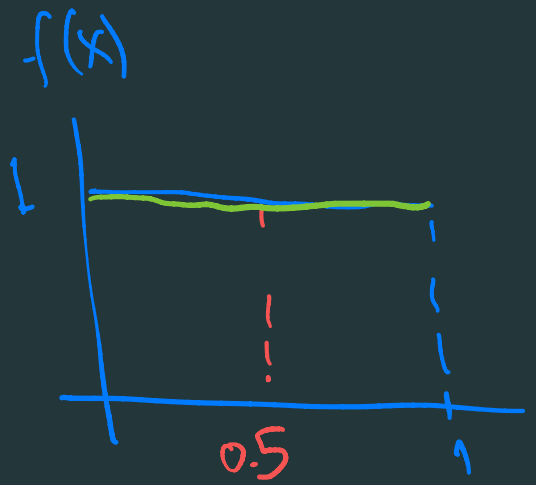
$\Rightarrow Var(X+Y)$ can be $\leq Var(X)+Var(Y)$ (useful for variance reduction)

law of large numbers

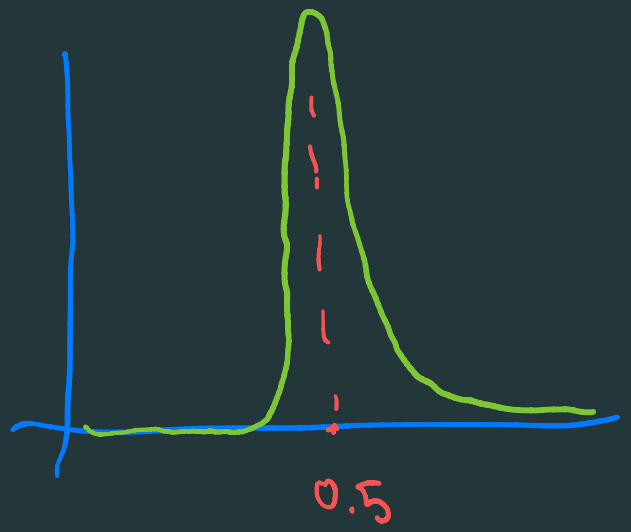
let $X_1, X_2, \ldots$ be a sequence of independent rvs with $\mathbb{E}[X_i] = \mu$ for all i
then, "almost" always (almost sure convergence)

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu \quad , \quad \text{as } n \to \infty$$

note: for any finite n, $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ is still a random variable



$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$



central limit theorem

let $X_1, X_2, \ldots$ be a sequence of independent rvs with

$$\mathbb{E}[X_i] = \mu, \; Var(X_i) = \sigma^2 < \infty \text{ for all } i$$

then,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i - \mu) = \sqrt{n}\left(\frac{\sum x_i}{n} - \mu\right)$$

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} \sigma\mathcal{N}(0,1) = \mathcal{N}(0, \sigma^2) \quad , \quad \text{as } n \to \infty$$

'convergence in distribution'

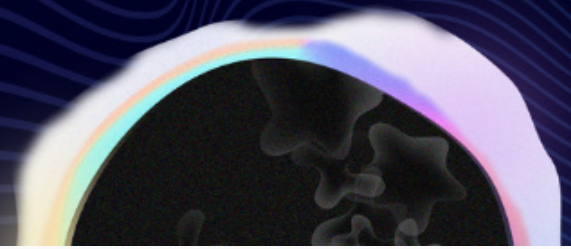
approximations for large n,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{D}{\approx} \mathcal{N}\left(\mu, \sigma^2/n\right)$$

$$S_n = \sum_{i=1}^{n} X_i \overset{D}{\approx} \mathcal{N}\left(n\mu, n\sigma^2\right)$$

QUESTION OF THE DAY

How can we tell if we are not already in a computer simulation?



SingularityNET

community.singularitynet.io

#AGICHAT

clicker question: Staffing a Food Bank

a food bank depends on volunteers for its labor pool
on any given day, the number of workers who show up is $\mathit{Uniform}(\{1, 2, \ldots, 9\})$,
while the number of donations needed to be collected is $\mathit{Uniform}(\{1, 2, \ldots, 29\})$
assuming the work is equally divided among each worker, what is the average load for each worker?

22 (a) 3

45 (b) > 3

9 (c) < 3

24 (d) can be any of the above on different days

staffing a food bank

a food bank depends on volunteers for its labor pool

on any given day, the number of workers who show up is $\mathit{Uniform}(\{1, 2, \ldots, 9\})$, while the number of donations needed to be collected is $\mathit{Uniform}(\{1, 2, \ldots, 29\})$

assuming the work is equally divided among each worker, what is the average load for each worker?

$\mathbb{E}[X] = \frac{1}{9}\sum_{i=1}^{9} i = \frac{9 \cdot 10}{2 \cdot 9} = 5$, $\mathbb{E}[Y] = \frac{1}{29} \cdot \left(\frac{29 \times 30}{2}\right) = 15$

- let X = number of workers, Y = number of donations
- we have $\frac{\mathbb{E}[Y]}{\mathbb{E}[X]} = \frac{30/2}{10/2} = 3$
- on the other hand, $\mathbb{E}[\mathit{Load}] = \mathbb{E}\left[\frac{Y}{X}\right]$; is this also 3?
- let us simulate and check!

↑ 'Assuming X, Y independent'

$\underbrace{\mathbb{E}[Y]}_{15} \cdot \underbrace{\mathbb{E}\left[\frac{1}{X}\right]}_{?}$

understanding what happened

for random variables X and Y, and function $g(\cdot, \cdot)$,

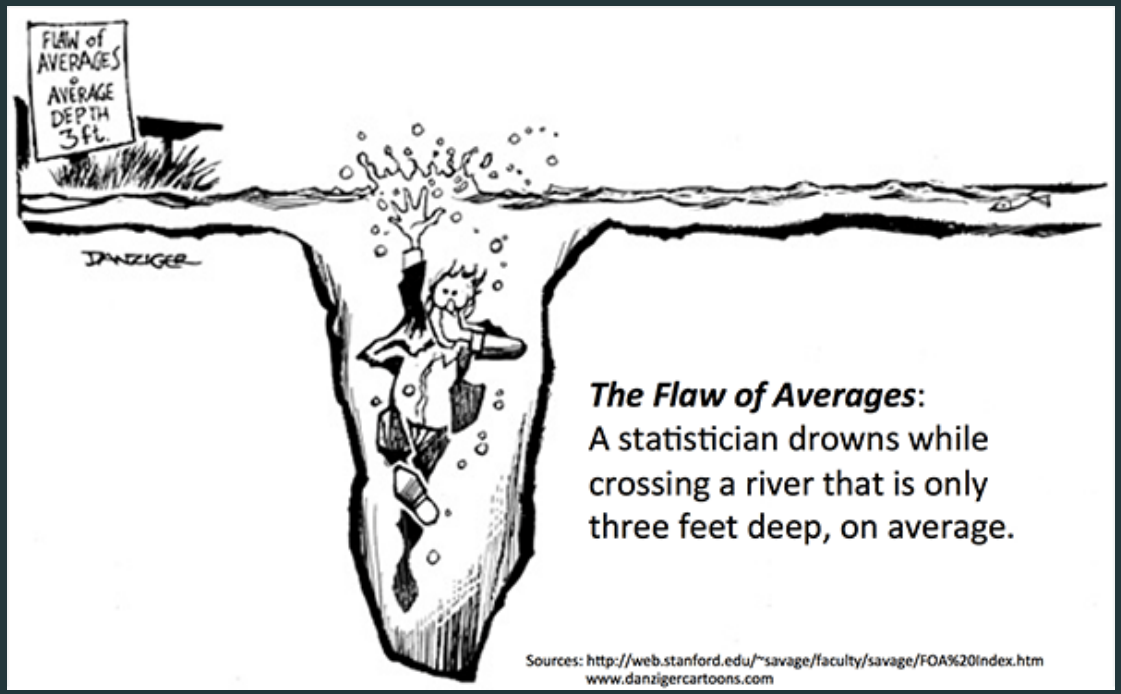
$$\mathbb{E}[g(X, Y)] \neq g(\mathbb{E}[X], \mathbb{E}[Y])!$$

- Equality only if g is linear (LOI)

flaw of averages

moral: in most settings, average inputs don't give average outputs!

$$\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$$



what can go wrong?

- non-linearities
- correlations between rvs
- 'inspection paradox' (buses take longer to arrive than they should!)
- ...

simulation allows us to avoid such problems!

- Ways of quantifying error

'frequentist' **confidence intervals**

(alternate - Bayesian 'credible' interval)

how many replications?

simulating food bank for many days (samples/replications) = distribution of loads

as number of replications increases, sample average $\to$ average load (by LLN)

question: every time we run the simulation model with some fixed number of replications, our estimate of $\mathbb{E}[load]$ changes.
how 'confident' can we be in our estimate?

- *answer*: use CLT to build a confidence interval!

want $-$ true avg $\in$ sim avg $\pm \delta$ with 95% prob

confidence intervals

let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean $\mathbb{E}[X_1] = \mu$ and variance $Var(X_1) = \sigma^2 < \infty$

want to measure μ from simulations

confidence interval: attempt 1...

an interval $[a, b]$ is called a 95% confidence interval for $\mathbb{E}[X_1]$ if

$$\mathbb{P}[a \leq \mathbb{E}[X_1] \leq b] \geq 0.95$$

Q: what is wrong with this?

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n$$

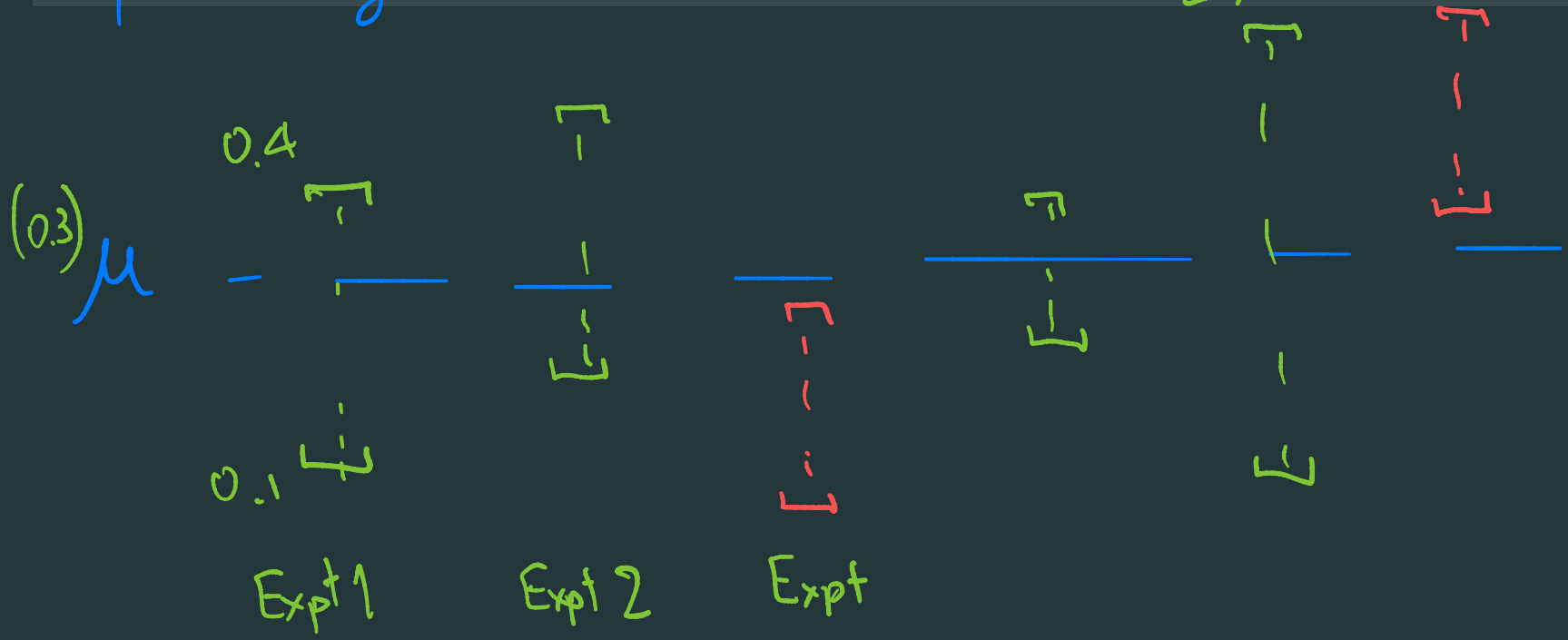
confidence intervals: definition

function of the simulation

an random interval $[A, B]$ (computed from data/experiments) is called a 95% confidence interval for some (deterministic) quantity μ if

$$\mathbb{P}[A \leq \mu \leq B] \geq 0.95$$

operationally - want smallest interval $[A, B]$



confidence intervals for population mean

$X_1, X_2, \ldots$ are i.i.d. rvs with $\mathbb{E}[X_1] = \mu$ and $Var(X_1) = \sigma^2 < \infty$; $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

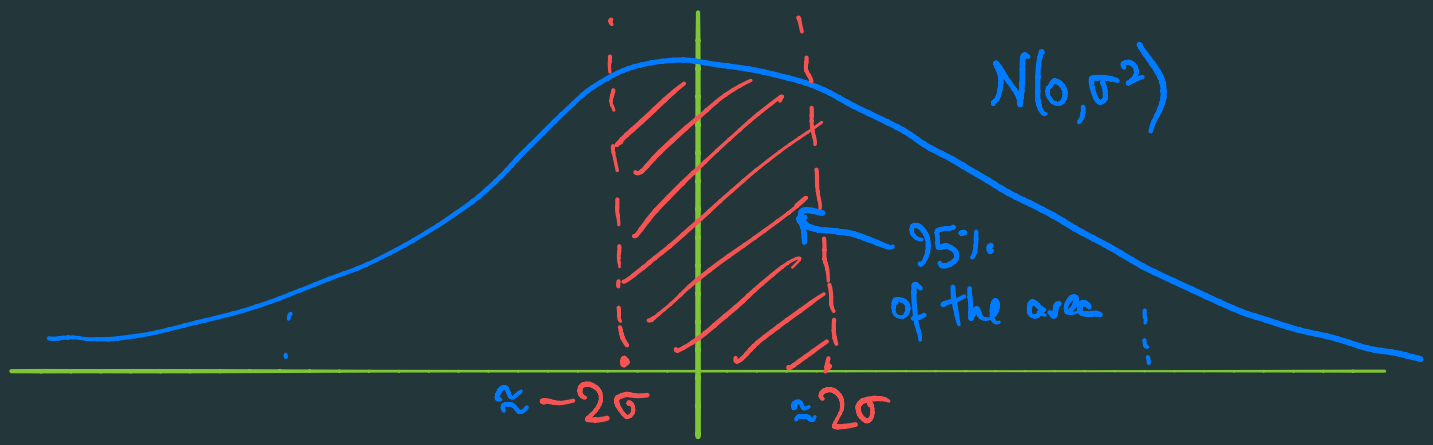
pop mean — want to measure — unbiased statistic

- from the central limit theorem:

$$\sqrt{n}(\bar{X}_n - \mu) \overset{D}{\approx} \sigma\mathcal{N}(0,1) = \mathcal{N}(0,\sigma^2)$$

- from the inverse cdf of $\mathcal{N}(0,1)$, we can compute

$$\mathbb{P}[-1.96 \leq \mathcal{N}(0,1) \leq 1.96] \geq 0.95.$$



confidence intervals

want to measure $\mu = \mathbb{E}[X_1]$ from simulations

- from the central limit theorem: $\sqrt{n}(\bar{X}_n - \mu) \overset{D}{\approx} \sigma\mathcal{N}(0,1)$
- from the cdf of $\mathcal{N}(0,1)$, we have $\mathbb{P}[-1.96 \leq \mathcal{N}(0,1) \leq 1.96] \geq 0.95$

putting these together, we have:

$$\mathbb{P}\left[\bar{X}_n - \frac{1.96\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{1.96\sigma}{\sqrt{n}}\right] \geqslant 0.95$$

random — constant (unknown) — random

confidence intervals: problems

- the confidence interval is approximate because

$\sqrt{n}(\bar{X}_n - \mu)$ is only approximately Gaussian

- the confidence interval is 'exact' when

$\sqrt{n}(\bar{X}_n - \mu)$ is 'exactly' Gaussian
(i.e., $X_i \sim N(\mu, \sigma^2)$)

- the confidence interval above requires knowledge of σ^2

Soln 1) use upper bound for σ^2 (Eg - If $X \sim \text{Ber}(p)$, $\text{Var}(X) = p(1-p) \leq 1/4$ (≤ 1))

confidence intervals: problems

- the confidence interval is approximate because

- the confidence interval is 'exact' when

- the confidence interval above requires knowledge of σ^2

can replace σ^2 with its sample estimator

$$s_n^2 = \underbrace{\frac{1}{n-1}}_{\text{correction}} \sum_{i=1}^{n} (X_i - \underbrace{\bar{X}_n}_{\text{sample mean}})^2 .$$

In practice $n \approx n-1$

fixed sample-size: recipe for CI

approximate $100(1-\alpha)\%$ Gaussian CI for $\mathbb{E}X$

pilot

1. select a sample size N
2. generate N i.i.d. samples $X_1, X_2, \ldots, X_N$ of X (replicates)
3. compute the estimators $\bar{X}_N, s_N^2$

$$\bar{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n, \qquad s_N^2 = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \bar{X}_N)^2$$

4. look up the value of $z_{\alpha/2}$ such that

$$\mathbb{P}[-z_{\alpha/2} \leq N(0,1) \leq z_{\alpha/2}] = 1-\alpha$$

For this course
$\alpha = 0.05$
$z_{\alpha/2} = 1.96 \approx 2$

5. the approximate $100(1-\alpha)\%$ CI for $\mathbb{E}X$ is given by

$$\bar{X}_N \mp z_{\alpha/2}\frac{s_N}{\sqrt{N}}$$

selecting the sample size

how large should N be so that the resulting $100(1-\alpha)\%$ confidence interval will have a pre-specified width?

- CI $\implies \bar{X}_N \mp z_{\alpha/2}\left(\sigma/\sqrt{N}\right)$
- half-width $\implies z_{\alpha/2}\frac{\sigma}{\sqrt{N}} = \ell$
- ℓ be the desired half-width
- Set $N = \frac{4\sigma^2}{\ell^2}$ $\left(\frac{(z_{\alpha/2})^2\sigma^2}{\ell^2}\right)$

selecting the sample size

- problem: σ^2 is unknown!
- estimating σ^2 through s_N^2 requires simulation!
- solution: 'pilot runs'
  perform k simulation runs to get $[X'_n : n = 1, \ldots, k]$ as outcomes
  compute

$$\tilde{X}_k = \frac{1}{k} \sum_{n=1}^{k} X'_n, \qquad \tilde{s}_k^2 = \frac{1}{k-1} \sum_{n=1}^{k} \left(X'_n - \tilde{X}_k \right)^2$$

'correction' empirical variance

  use $\tilde{s}_k^2$ to estimate σ^2

for confidence level α, half-width ℓ, set

$$N = \left\lceil \frac{z_{\alpha/2}^2 \tilde{s}_k^2}{\ell^2} \right\rceil$$

clicker question

we want to estimate $\mathbb{E}[X]$ with ∓ 0.01 accuracy with 95% confidence from 10 replications, $\tilde{s}_{10}^2 = 0.5$

how many replications should we use?

A 200

38 **B** 2,000

48 **C** 20,000

D 40,000

E 200,000

clicker question

we want to estimate $\mathbb{E}[X]$ with ∓ 0.01 accuracy with 95% confidence from 10 replications, $\tilde{s}_{10}^2 = 0.5$

thus we have

$$\left\lceil \frac{z_{\alpha/2}^2 \tilde{s}_k^2}{\ell^2} \right\rceil = \frac{2^2 \times 0.5}{0.01^2} = 20,000$$

how many replications should we use?

A 200

B 2,000

C 20,000

D 40,000

E 200,000

basic simulation workflow

Typical Sim : Choose 'statistic' which is an 'unbiased estimator'

- perform pilot run of k simulations (sufficient but not large k) – compute $\tilde{s}_k$
- compute required sample-size N for desired confidence interval
- run N additional simulations $\Longrightarrow$ production runs
- form fixed-sample confidence intervals from these N samples
- note: final CI may be different than desired, because it is constructed by using s_N^2 (may be larger/smaller than $\tilde{s}_k^2$)
- for the final confidence interval, discard the information from the trial runs
  not a problem, since $N \gg k$ usually

clicker question: abusing CI (?)

in a homework assignment in Simulation, students were asked to do a Monte Carlo simulation to compute π up to 2 decimal places

there were 100 homework submissions, and *all* submissions reported 95% CIs which included 3.14

the probability that this happened 'by chance' is approximately

5 (a) 0.4

44 (b) 0.05

20 (c) 0.006 ✓

8 (d) 0.0007

23 (e) 0

$$\mathbb{P}[3.14 \in CI] \approx 0.95$$

$$\Rightarrow \mathbb{P}[\text{All students have } 3.14 \in CI] \approx (0.95)^{100}$$

Reasons –

- Everyone chooses same 'seed'
- Everyone has 'prior knowledge'

confidence intervals as a social contract

a random interval $[A, B]$ (computed from data/experiments) is a 95% confidence interval for some unknown μ if before the experiment is done

$$\mathbb{P}[A \leq \mu \leq B] \geq 0.95$$

Good practices

- Announce questions before hand
- Design exp^{t}, let someone else perform it

clicker question: (mis)interpreting CI

a 95% confidence interval for the mean annual rainfall in Milford Sound, New Zealand is 26 feet plus or minus 1 foot

Avg rain $\in$ [25ft, 27ft]

therefore, the rainfall next year will be between 25 feet and 27 feet with 95% probability.

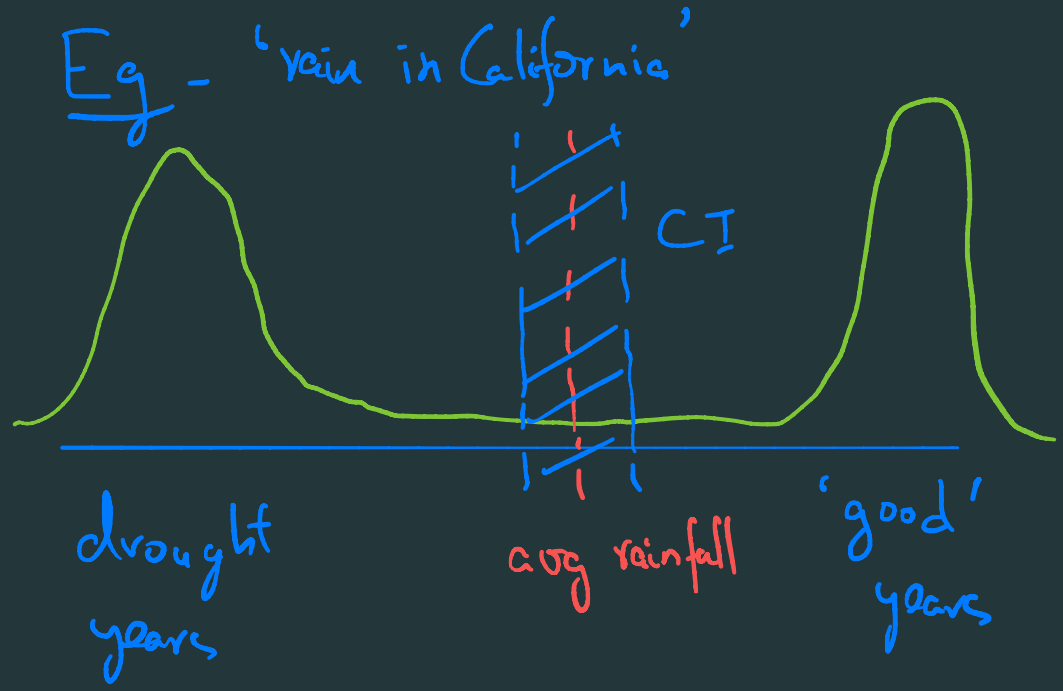
32%

(a) true

63%

(b) false

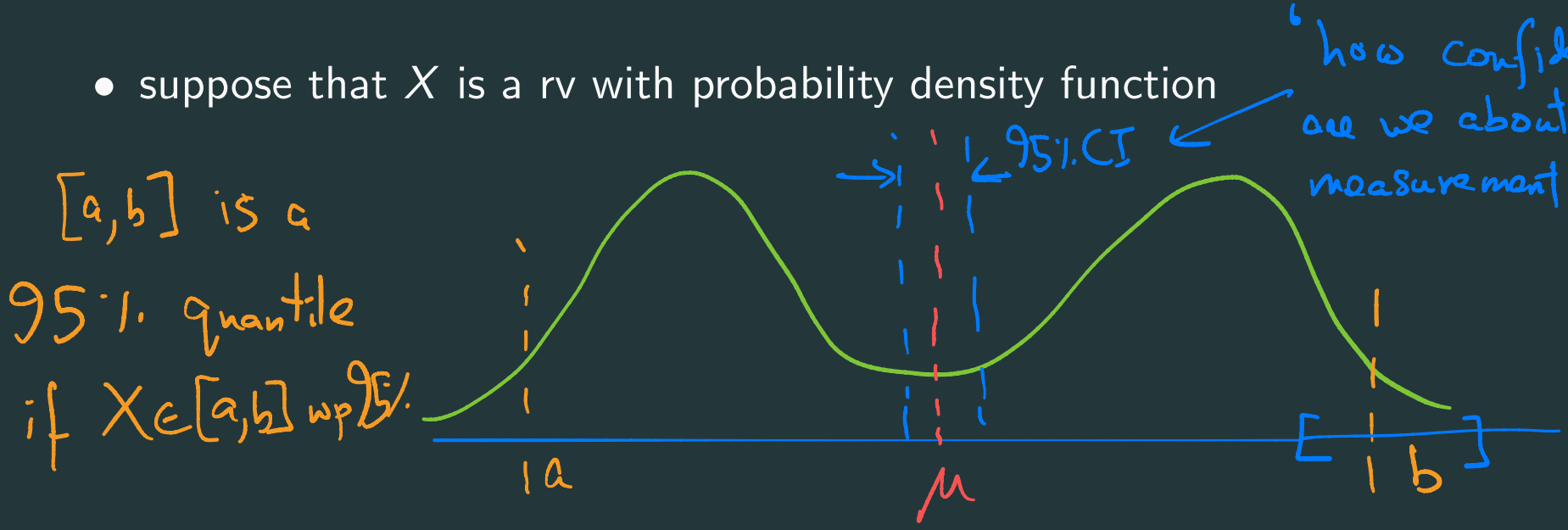
(c) where is New Zealand?



confidence intervals vs quantiles

the CI for the mean is NOT the same as the quantiles of a random variable.

- suppose that X is a rv with probability density function



- we can select q_1 and q_2 so that

$$\mathbb{P}[q_1 \leq X \leq q_2] = 0.95,$$

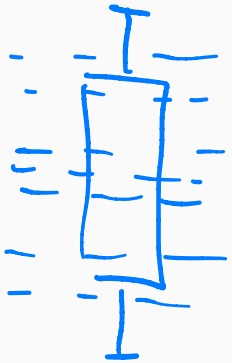
but $[q_1, q_2]$ is not a 95% confidence interval for $\mathbb{E}X$.

b = 95% quantile

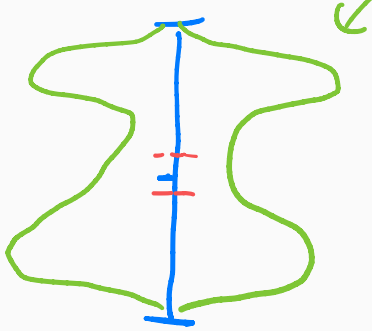
∈ [b−δ, b+δ]

Useful visualizations

SMORE plot



Violin plot



advanced probability tools

tool 1: Jensen's inequality

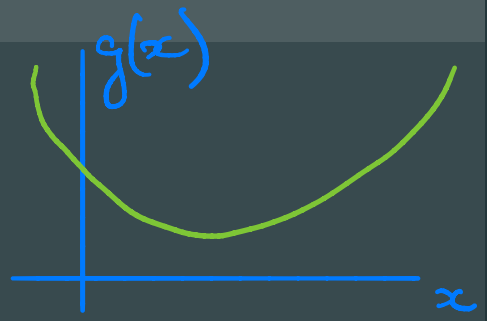
Recall - LOI - $\mathbb{E}[g(X)] = g(\mathbb{E}[X])$ if g is linear

Q: can we say something about $\mathbb{E}[f(X)]$ vs $f(\mathbb{E}[X])$ (in particular, $\geq$ or $\leq$) without simulating?

Jensen's inequality

if X is a random variable and f is a convex function, then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$



Example – in our food-bank staffing problem, since $f(x) = \frac{1}{x}$, $x > 0$ is convex:

$$\mathbb{E}[Y]\mathbb{E}\left[\frac{1}{X}\right] \geq \frac{\mathbb{E}[Y]}{\mathbb{E}[X]}$$

$g(x) = 1/x$ if $x > 0$

$\Rightarrow \mathbb{E}[1/X] \geq \frac{1}{\mathbb{E}[X]}$

Eg – $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$

- If g is concave (ie $\cap$), then $\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$

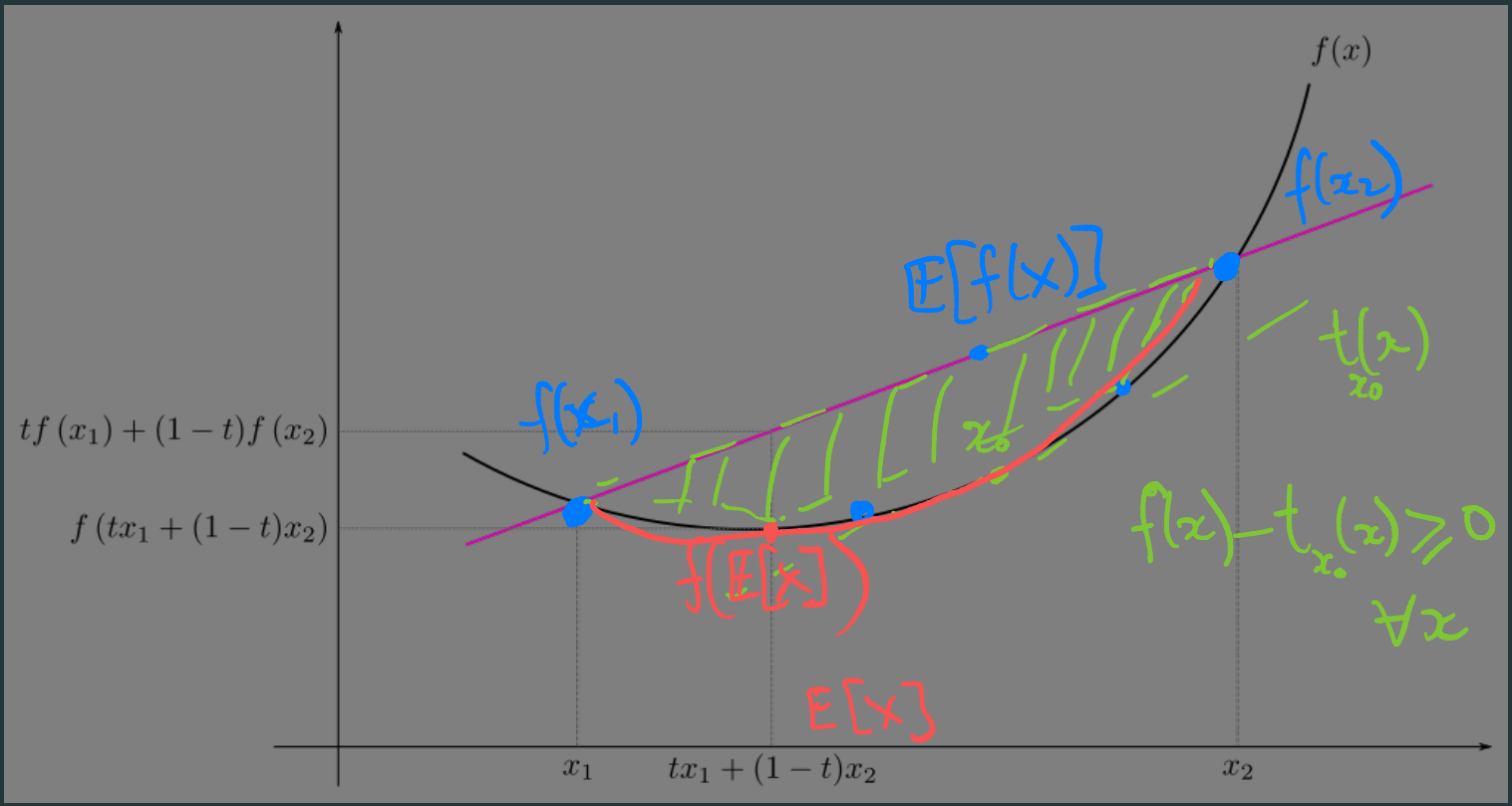
tool 1: Jensen's inequality

Jensen's inequality

if X is a random variable and f is a convex function, then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

proof sketch (plus way to remember)



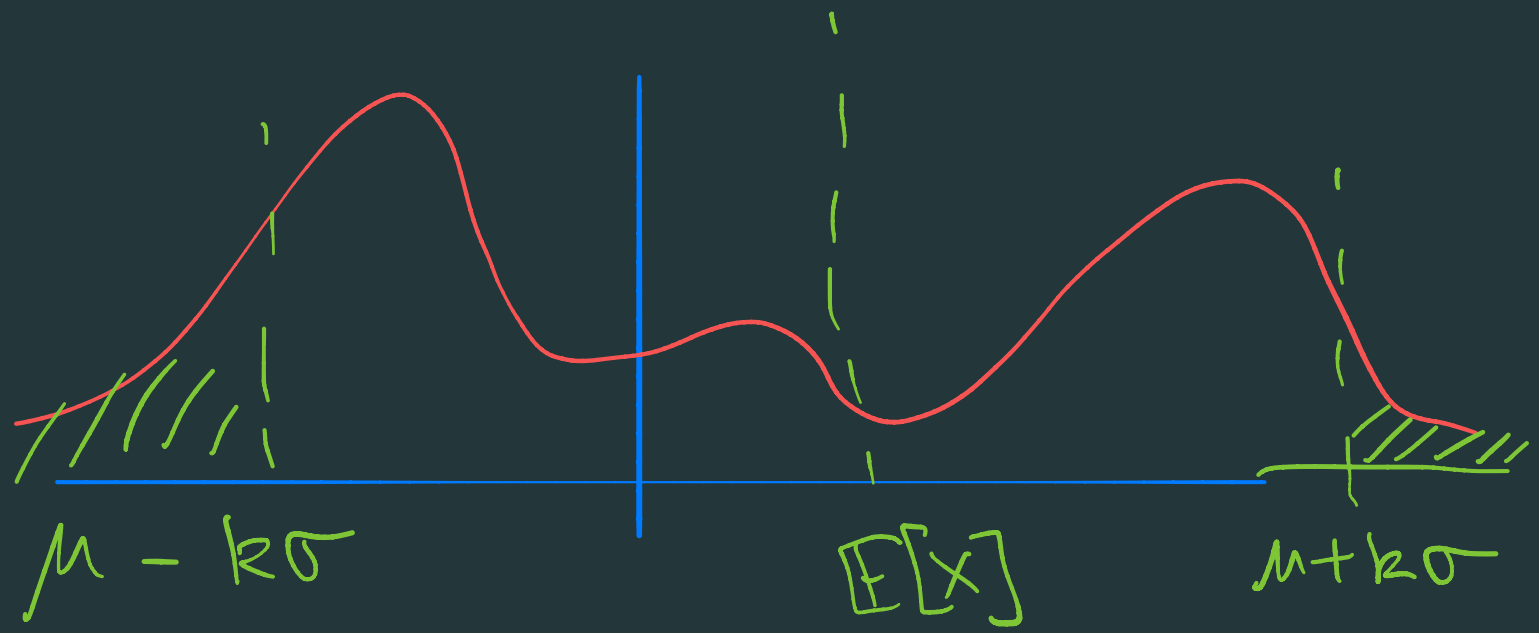
tool 2: Chebyshev's Inequality

Q: since the CLT convergence is faster/slower for different rvs, can we be sure that CIs based on variance always make sense?

Chebyshev's inequality

let X be any rv with finite mean μ and finite variance $\sigma^2 > 0$
then for any $k > 0$,

$$\mathbb{P}\left[|X - \mu| \geq k\sigma\right] \leq \frac{1}{k^2}$$



'always-valid' confidence intervals

Chebyshev's inequality

let X be any rv with finite mean μ and finite variance $\sigma^2 > 0$
then for any $k > 0$,

$$\mathbb{P}[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2} \quad = \quad \frac{1}{4}$$

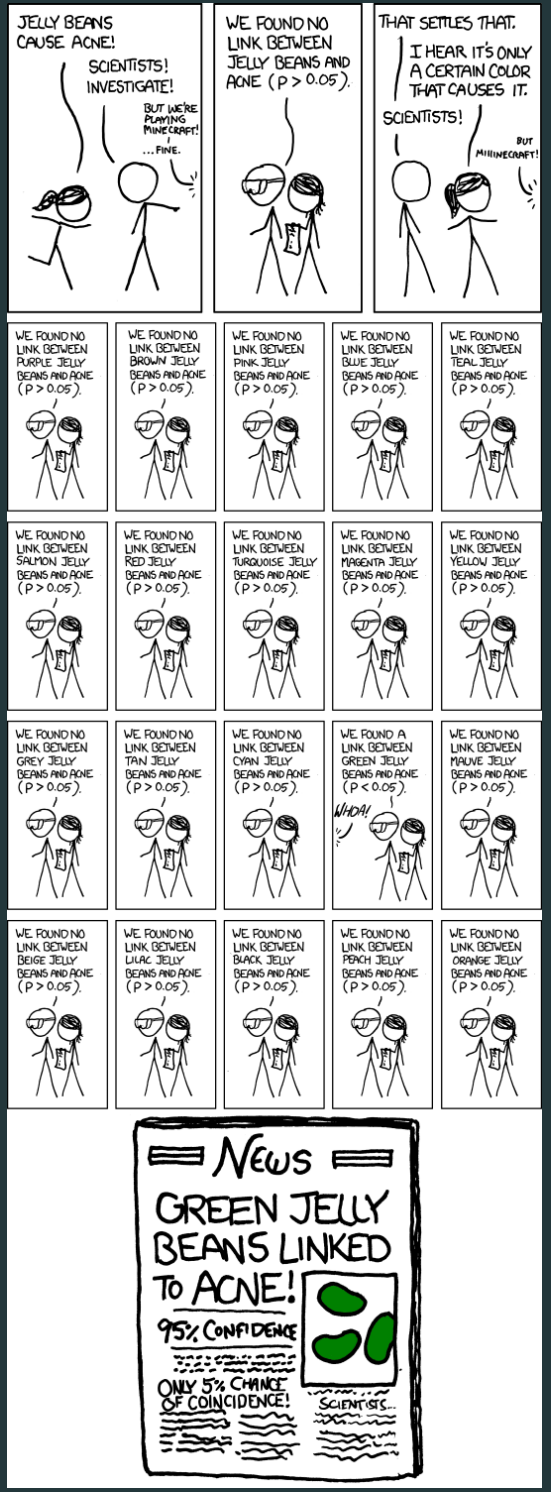
$k\sigma = 2\sigma$

worst-case CI: if we choose $k = 2$, then we always have 75% confidence intervals

For 95%, choose k s.t. $\frac{1}{k^2} = 0.05 = \frac{1}{20}$

$\Rightarrow k = \sqrt{20} \leq 5$

(ie, $\mu \in \left[\bar{X}_n - \frac{5\sigma}{\sqrt{n}}, \bar{X}_n + \frac{5\sigma}{\sqrt{n}}\right]$ wp $> 95\%$ always!

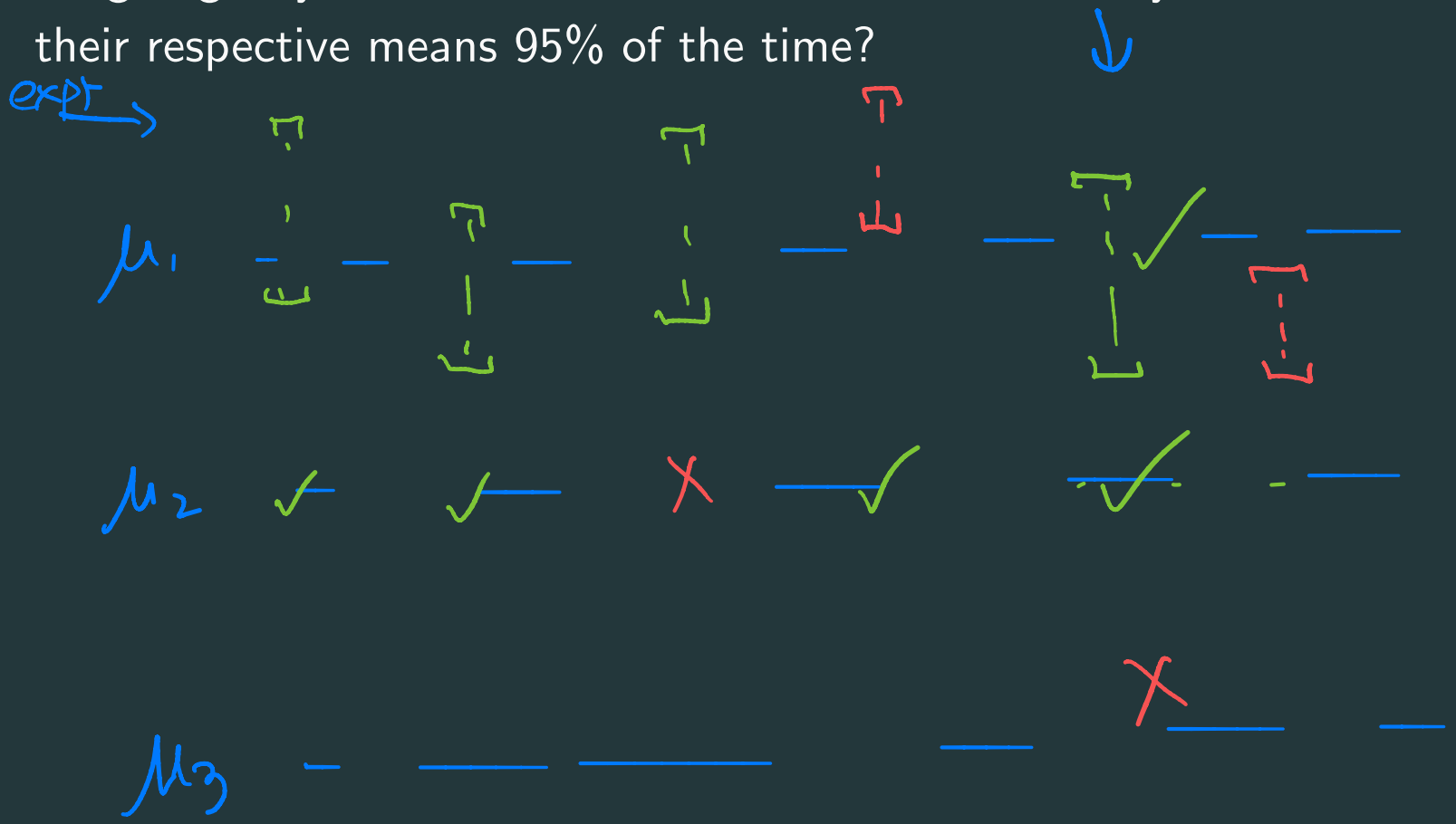


multiple hypothesis testing

tool 3: the union bound

Q: how can we get simultaneous confidence intervals for multiple hypothesis?

• Eg. I give you five 95% confidence intervals; do they simultaneously contain their respective means 95% of the time?



tool 3: the union bound

the union bound

let $A_1, A_2, \ldots, A_k$ be events; then

$$P(A_1 \cap A_2 \cap \cdots \cap A_k) = 1 - P(A_1^c \cup A_2^c \cup \cdots \cup A_k^c)$$

$$\geq 1 - (P(A_1^c) + P(A_2^c) + \cdots + P(A_k^c))$$

let A_i = event that the ith CI contains its true mean...

- α CI for each

$\mathbb{P}[\text{All } A_i \text{ hold simultaneously}] \geq 1 - k\alpha$

$\Rightarrow (1-\alpha)$ CI for each $A_i \longrightarrow (1-k\alpha)$ CI for all simultaneously