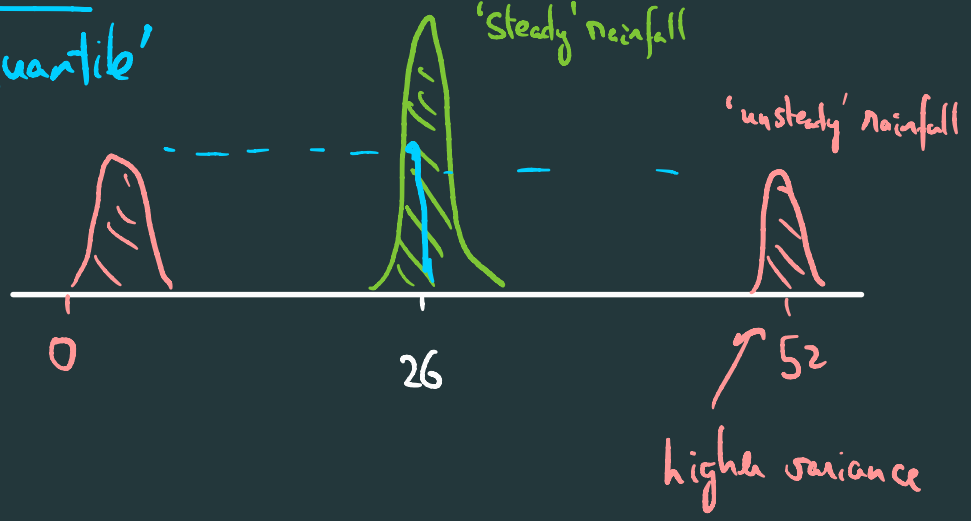
class poll: (mis)interpreting CI

a 95% confidence interval for the mean annual rainfall in Milford Sound, New Zealand is 26 feet plus or minus 1 foot (95% CI) ← CI for my mean estimate
therefore, the rainfall next year will be between 25 feet and 27 feet with 95% probability.

what you want is a 'quantile'



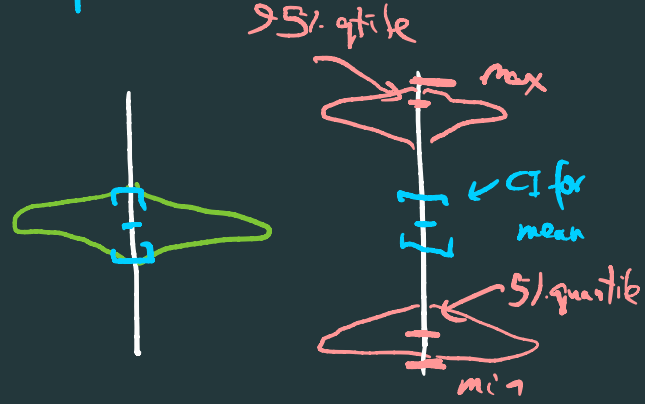
(a) true

(b) false ✓

(c) where is New Zealand?

Both these examples have the same mean!

Better way of representing data - Violin plot



Last class - 95% CI from data $(X_1, X_2, \ldots, X_n)$

$$\approx \left[\bar{X}_n - 1.96 \sqrt{\mathrm{Var}(\bar{X}_n)},\ \bar{X}_n + \underbrace{1.96}_{\approx 2} \sqrt{\mathrm{Var}(\bar{X}_n)} \right]$$

from the Gaussian CLT

$$\frac{\text{std-dev}(X_i)}{\sqrt{n}} \approx \frac{1}{\sqrt{n}} \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2}$$

empirical std-dev

- Use 'pilot runs' to estimate $\mathrm{Var}(X_i)$, decide # of replicates n

three advanced probability tools

Today

- 3 advanced tools - 'better' understand CIs
- Unit 5 - How to generate complex random variables

1) How can we 'guess' $\underbrace{\mathbb{E}[f(X)]}_{\text{may be unknown}}$ using $\underbrace{\mathbb{E}[X]}_{\text{known}}$?

- Jensen's Inequality (when is $\mathbb{E}[f(X)]$ 'related to' $f(\mathbb{E}[X])$)

2) How we deal with multiple CI (how to combine them)?

Eg - Testing multiple hypothesis $\equiv$ Union bound

3) Can we get valid CIs if we do not believe in the CLT?

$$CI \equiv \bar{X}_n \mp 1.96 \sqrt{Var(\bar{X}_n)}$$

Gaussian approximation

More generally $\sigma^2 = Var(\bar{X}_n)$

$$\mathbb{P}\left[|\bar{X}_n - \mu| > c\sigma\right] \leq \frac{1}{c^2}$$

$\Rightarrow$ Can always use CIs of the form $\bar{X}_n \mp \underbrace{c}_{\approx 4 \dots \text{ for } 95\%} \sqrt{Var(\bar{X}_n)}$

staffing a food bank (example from 2 classes back)

a food bank depends on volunteers for its labor pool
on any given day, the number of workers who show up is $\mathit{Uniform}(\{1, 2, \ldots, 9\})$, X
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? Y

i.e, Number of donations they need to pick up

- 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!

$\mathbb{E}[X] = \frac{1}{9}\sum_{i=1}^{9} i = \frac{9 \cdot 10}{9 \cdot 2}$

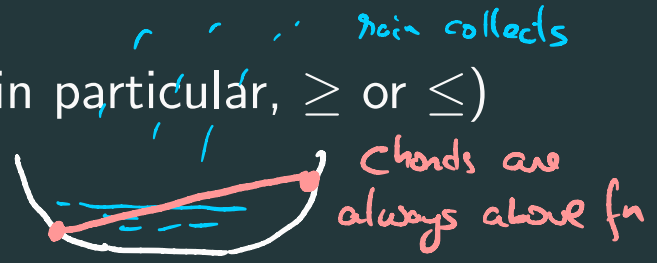
$\mathbb{E}[Y] = \frac{1}{29}\sum_{i=1}^{29} i = \frac{30}{2}$

$= \mathbb{E}\left[\frac{Y}{X}\right]$ load

found $\mathbb{E}\left[\frac{Y}{X}\right] > 3 = \frac{\mathbb{E}[Y]}{\mathbb{E}[X]}$

tool 1: Jensen's inequality

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



- $f(x)$ is convex in x if $\left(\frac{d^2f}{dx^2} \geq 0\right)$

Jensen's inequality

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

g is a convex fn

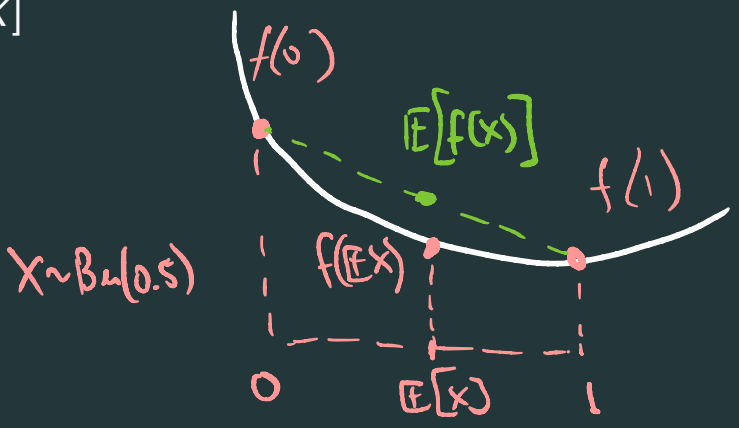
$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

$$\mathbb{E}[g(x)] \leq f(\mathbb{E}[g(x)])$$

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

$$\mathbb{E}\left[\frac{Y}{X}\right] \overset{X \perp\!\!\!\perp Y}{=} \mathbb{E}[Y]\mathbb{E}\left[\frac{1}{X}\right] \geq \frac{\mathbb{E}[Y]}{\mathbb{E}[X]}$$

Convex fn



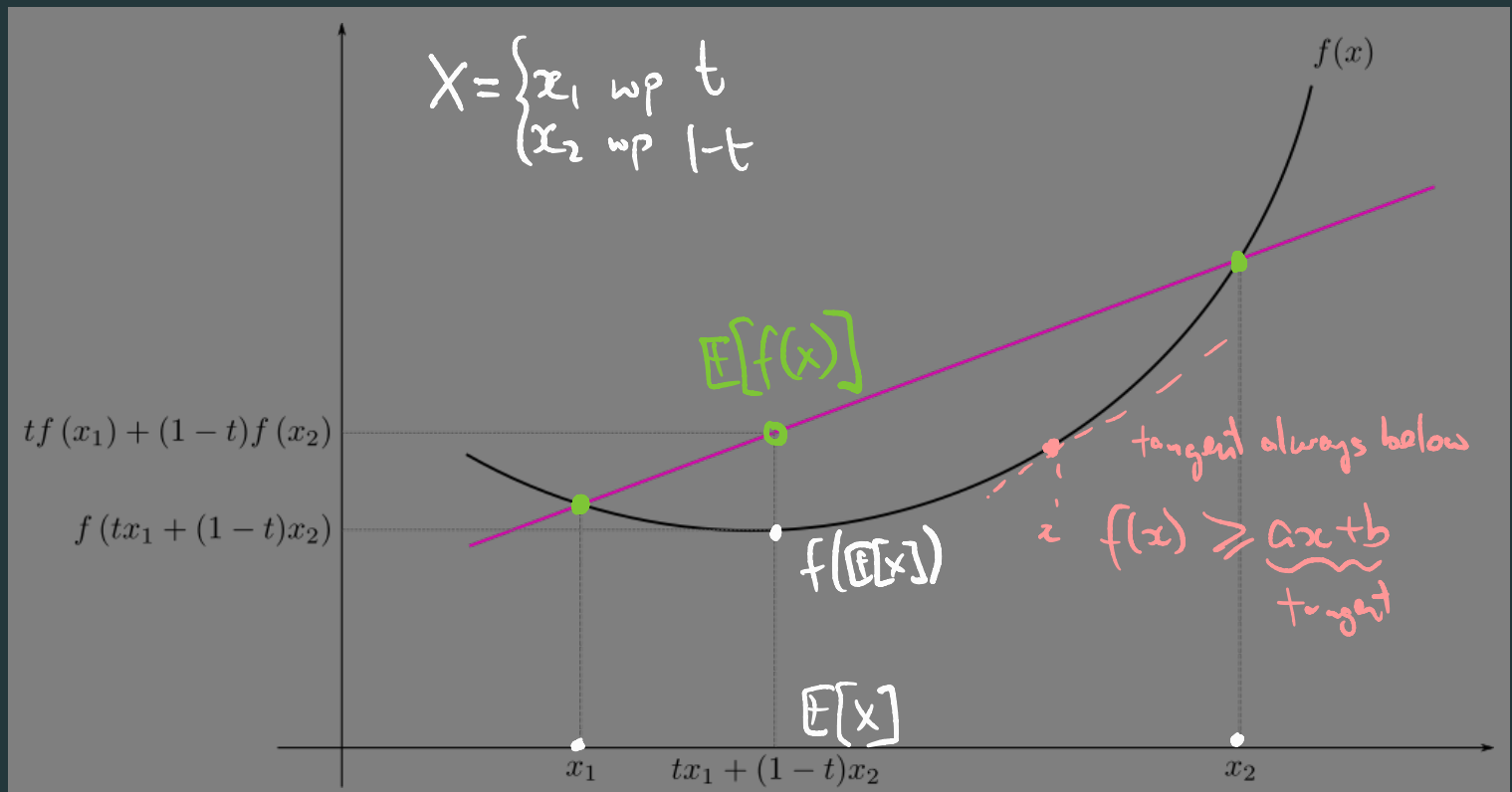
tool 1: Jensen's inequality

Jensen's inequality

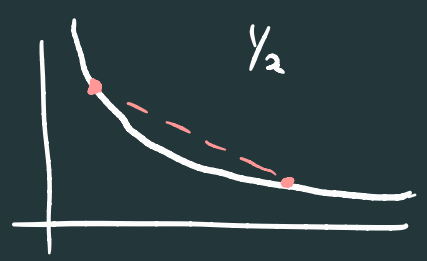
if X is a random variable and f is a convex function, then

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

proof sketch (plus way to remember)



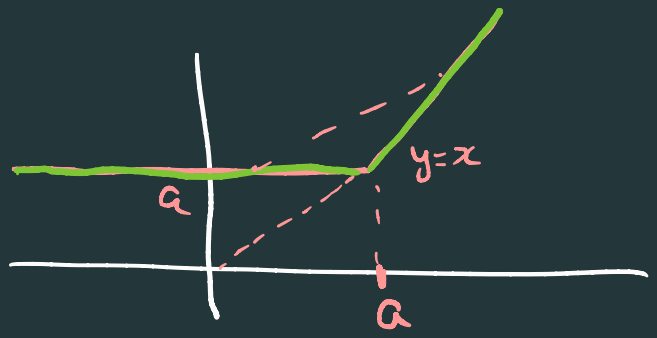
Convex fns



- $f(x) = x^2$

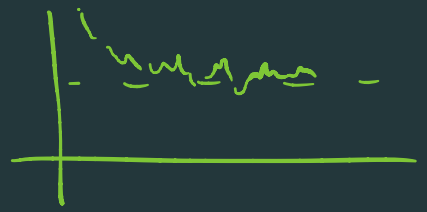
$f(x) = \frac{1}{x}$, $x \geq 0$

$f(x) = \max\{x, a\}$



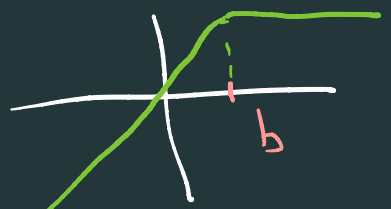
Eg- $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$ $(\Rightarrow \mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \geq 0)$

$\mathbb{E}[1/X] \geq 1/\mathbb{E}[X]$ ← showed up for Buffon's needle



$\mathbb{E}[\max(X, a)] \geq \max(\mathbb{E}[X], a)$

$\mathbb{E}[\min(X, b)] \leq \min(\mathbb{E}[X], b)$



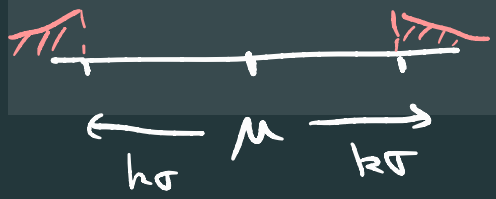
tool 2: Chebyshev's Inequality (similar to linearity of indep in that NO FEW ASSUMPTIONS)

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 What we need for CLT

let X be any rv with finite mean μ and finite variance $\sigma^2 > 0$ ie $|E[X]| = |\mu| < \infty$, $Var(X) = \sigma^2 < \infty$

then for any $k > 0$,



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

absolute deviation from mean

- Gaussian CI $\equiv \bar{X}_n \mp 1.95\sqrt{Var(\bar{X}_n)}$ (for 95%)

Always valid CI $\equiv \bar{X}_n \mp 5\sqrt{Var(\bar{X}_n)}$ for 95% CI

- Is Chebyshev tight? YES if X is Bernoulli (Extreme values)

Why is this useful

random interval

$$-\ (1-\alpha)\text{-CI} \equiv \mathbb{P}\left[\mu \in [A, B]\right] \geq 1-\alpha$$

From Chebyshev -

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

true mean

$$\Rightarrow \mathbb{P}\left[|\bar{X}_n - \mu| \geq k\sqrt{\mathrm{Var}(\bar{X}_n)}\right]$$

ALWAYS TRUE

$$= \mathbb{P}\left[\mu \notin \left[\bar{X}_n - k\sqrt{\mathrm{Var}(\bar{X}_n)}, \bar{X}_n + k\sqrt{\mathrm{Var}(\bar{X}_n)}\right]\right] \leq \underbrace{1/k^2}_{\text{at most } \alpha}$$

For $(1-\alpha)$-CI, can choose $k \geq 1/\sqrt{\alpha}$

Eg - 75% CI $\Rightarrow \mathbb{P}[\ldots] \geq 0.75 \Rightarrow k = 1/\sqrt{0.25} = 2$

vs 95% CI for Gaussian CI

‘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 $k > 0$,

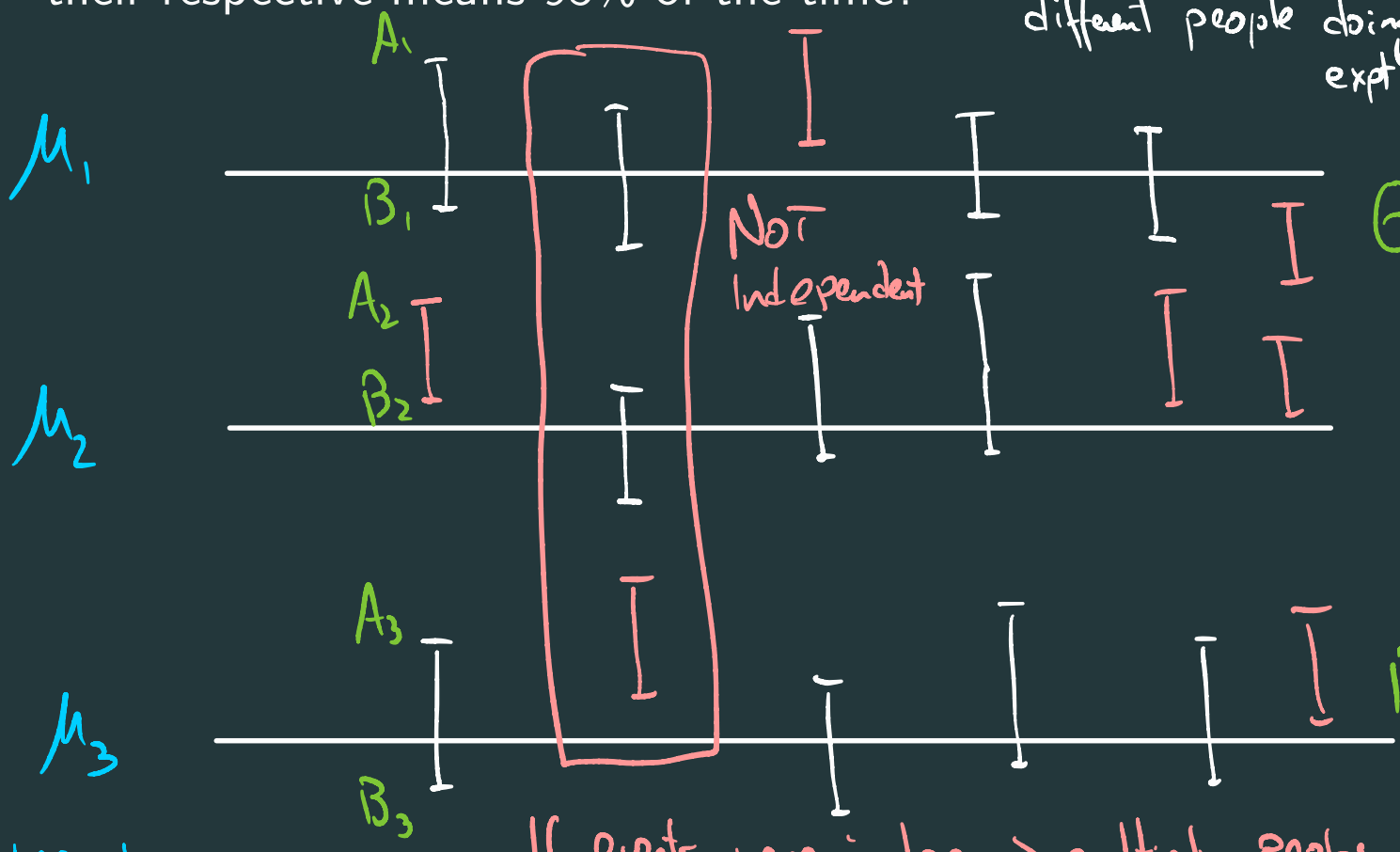
$$\mathbb{P}\left[|X - \mu| \geq k\sigma\right] \leq \frac{1}{k^2}$$

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

tool 3: the union bound

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

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



Q: What is the prob that ALL CIs for a single experiment are valid

ie

$$\mathbb{P}\left[\begin{array}{l}\mu_1 \in [A_1, B_1] \text{ AND} \\ \mu_2 \in [A_2, B_2] \text{ AND} \\ \mu_3 \in [A_3, B_3]\end{array}\right]$$

tool 3: the union bound

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

the union bound = $\mathbb{P}[A_1 \text{ OR } A_2 \text{ OR } A_3] \leq \mathbb{P}[A_1] + \mathbb{P}[A_2] + \mathbb{P}[A_3]$

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

same as $A_1 \cup A_2 \cup A_3$

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

← de Morgan's Law $(A \cup B)^c = A^c \cap B^c$

AND

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

$\mathbb{P}[\text{CI is BAD}]$

let A_i = event that the ith ($1-\alpha$) CI contains its true mean...

(Ex- 95% CI for μ_1, μ_2 and μ_3)

$$\mathbb{P}\left[\begin{array}{l}\mu_1 \in [A_1, B_1] \text{ AND} \\ \mu_2 \in [A_2, B_2] \text{ AND} \\ \mu_3 \in [A_3, B_3]\end{array}\right] \geq 1 - (\alpha + \alpha + \alpha) = 1 - 3\alpha$$

'bad probabilities add up'

Best case scenario = if all were independent, $(1-\alpha)^3 \geq 1 - 3\alpha$

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 $\mathbb{E}X$.