Nov 19, 2025

Sampling / Statistical inference

☆ Goal: to infer something about the population at large using a random sample.

a lot of math only works if the sample is randomly drawn from the population.

	Sample version	Population version
Mean	$\hat{\mu} = \frac{1}{n}\sum X_i = \bar{X} \longrightarrow$ "sample mean"	$\mu = \mathbb{E}(X)$ "population mean"
Variance	Version 1: Biased / naive $s^2 = \frac{1}{n}\sum (X-\bar{X})^2$	$\sigma^2 = \mathbb{E}\left[(X-\mathbb{E}(X))^2\right]$
	version 2: Unbiased / Corrected $s^2 = \hat{\sigma}^2 = \frac{1}{n-1}\sum (X-\bar{X})^2$	$= \mathbb{E}(X^2) - \mathbb{E}(X)^2$
	R's default behav. `var(x)`	$= Var(X)$ / $\mathbb{V}(X)$.

Covariance $\widehat{Cov(X,Y)}$
$= \frac{1}{n-1} \sum (X-\bar{X})(Y-\bar{Y})$.
↑
R: `cov(x, y)`

$Cov(X,Y) = \mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))]$
$= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$

Statistical inference:
- We want to know
  1. A certain characteristic of the population (eg: average life expectancy)

  Solution: use sample statistics to infer
  → Sample mean → population mean
  $\bar{X}$ → μ

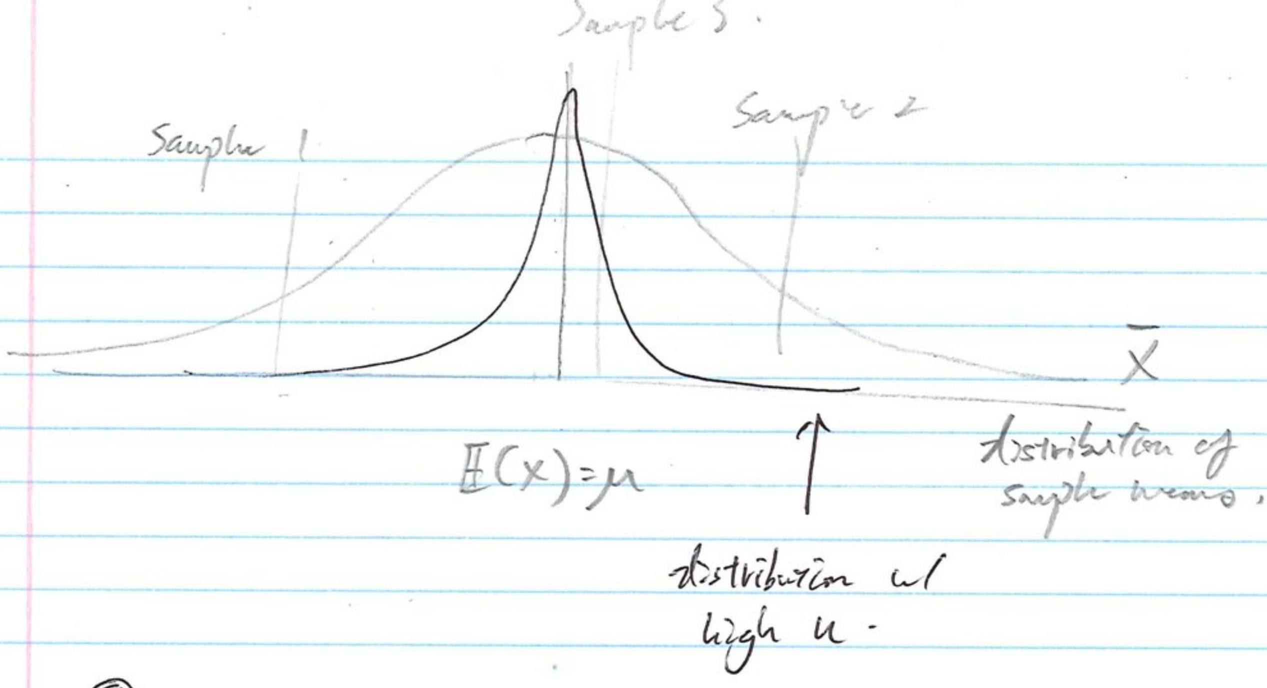
⇒ 2. How confident we are in our estimate, given that we only have 1 sample.

  Solution: Central limit theorem (CLT) + std. errors / confidence intervals.

Central Limit Theorem:
Let $\mu = \mathbb{E}(X)$ and $\sigma^2 = Var(X)$,
the sample statistics $\bar{X}$ is distributed acc. to a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$

" $\sqrt{n}(\bar{X}-\mu) \xrightarrow{d} N(0, \sigma^2)$ "

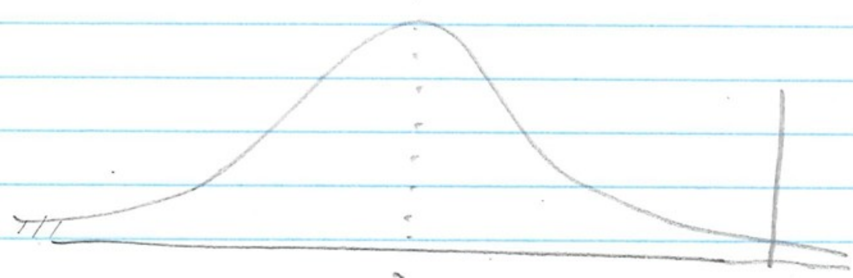
"Standard error" = std. dev. of this dist = $\sqrt{\frac{\sigma^2}{n}}$



Quantifying "confidence"

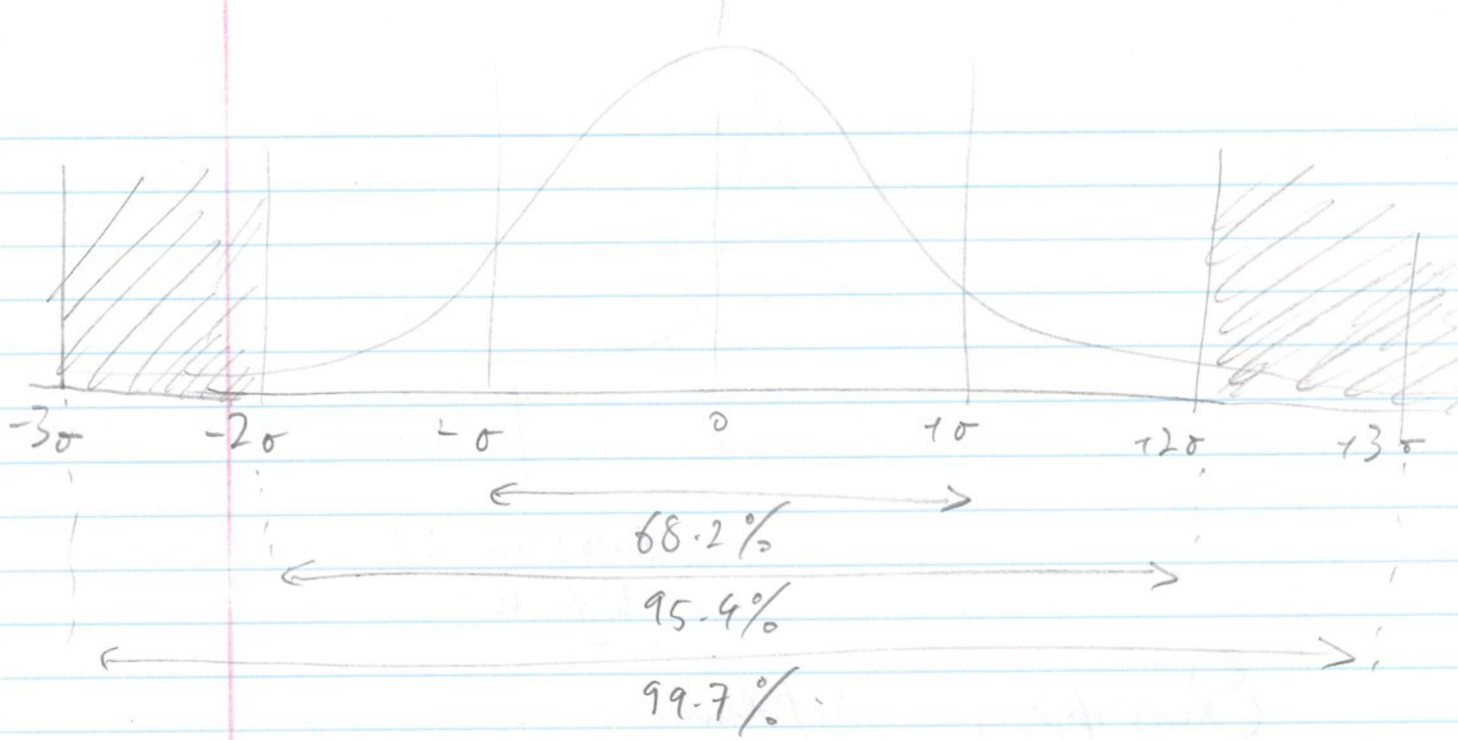
↳ eg. SDO – we want to know whether the SDO we have is meaningfully different from 0.

→ Apply CLT: if the true effect is 0, then the sample SDOs should be normally distributed w/ mean 0, and variance $\frac{\sigma^2}{n}$



- Confidence: how lucky do we have to be to observe an SDO that far in the dist. by sheer luck.

→ $= \Pr(Z \geq SDO)$ ← we know this because the norm dist is standard!!



Convention: we treat a statistic as "significant" if it falls outside the 95% threshold, i.e.

$SDO \geq 1.96$ s.e. or $SDO \leq -1.96$ s.e.

↳ ~5% false positive rate.

Variants: Confidence intervals!

95% = [statistic − 1.96 s.e, statistic + 1.96 s.e]

↳ if interval "crosses" 0, the statistic is not significant.

P-value: probability of getting a number that big by chance

→ significant if $p < 0.05$ (i.e. 5%).