THE CENTRAL LIMIT THEOREM

VIRAJ DSOUZA
10 MAY 2024
MATHEMATICS OF UNCERTAINTY

In simple words…

- **Regardless** of the Population distribution, as the Sample size **increases**,
  - Sample Mean tends to Normally Distribute around the Population Mean, and
  - Sample Standard Deviation shrinks
- Understanding this empirically (By Simulating Random samples)

POPULATION VS SAMPLE

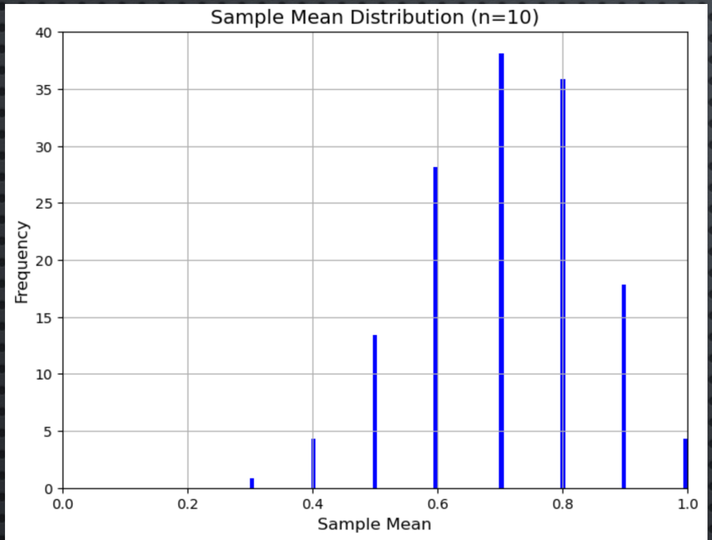
A POPULAR ARTICLE CLAIMS:

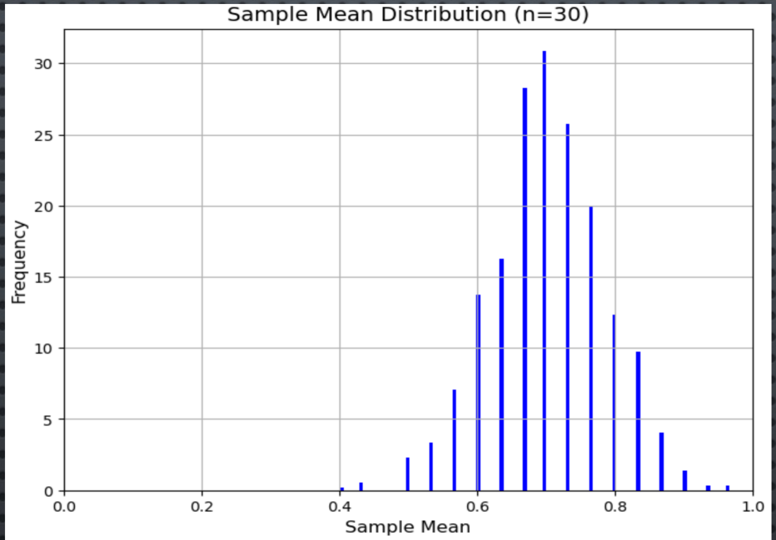
OUT OF THE POPULATION OF 10 MILLION PEOPLE WHO CAN VOTE, 70% SUPPORT PARTY **A** AND 30% SUPPORT PARTY **B**

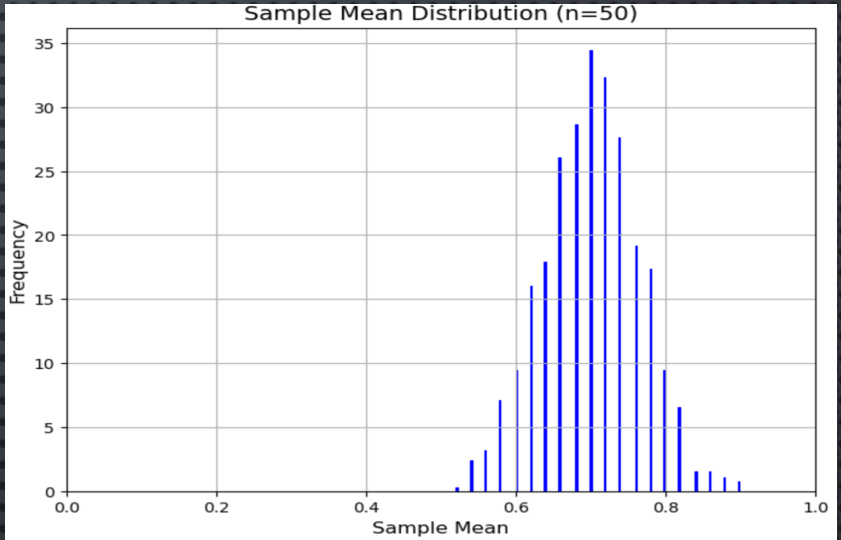
GIVEN YOU HAVE ACCESS TO THIS POPULATION (BUT LIMITED RESOURCES), HOW WOULD YOU VERIFY THE ABOVE CLAIM?

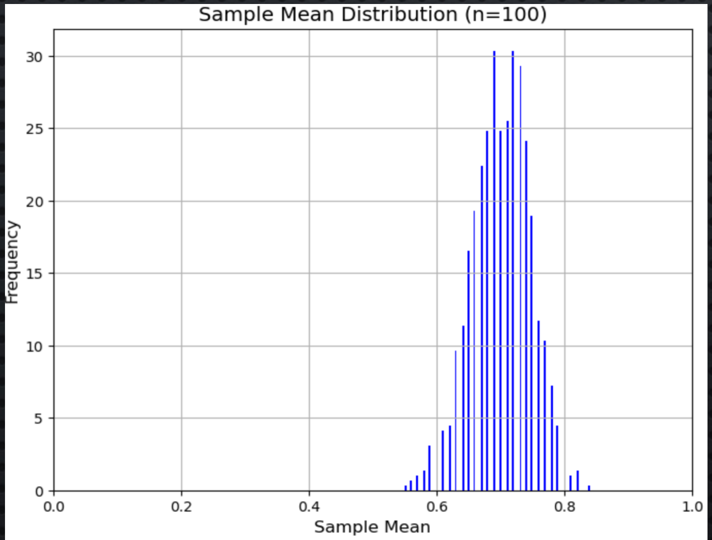
- RANDOM SAMPLING
- **ESTIMATE** THE SUPPORT FOR PARTY A → LAW OF LARGE NUMBERS

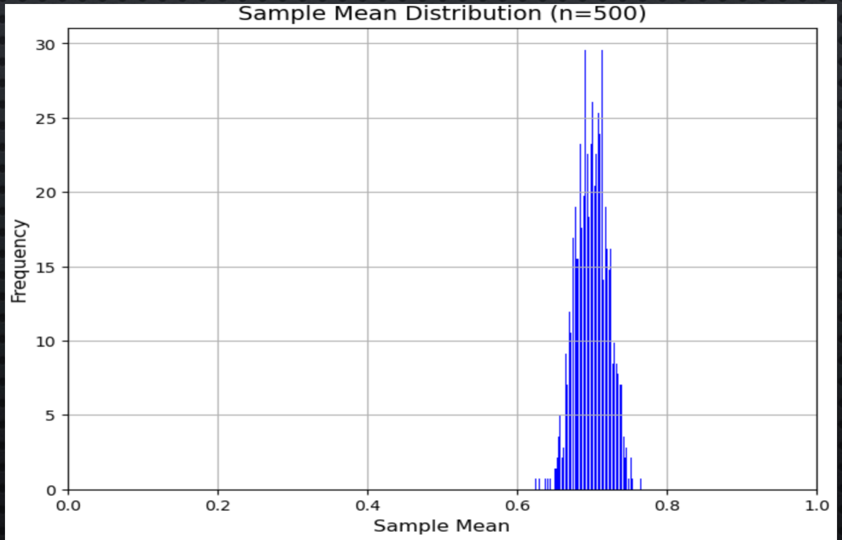
RANDOM SAMPLE SIMULATIONS











True Population mean= 0.7

Simulation results for various sample sizes ‘n’ is shown

FROM THE SIMULATIONS...

Variability in Sampling Distribution Decreased as Sample Size Increased

Estimate from a Larger Sample size → More Accurate (**Low sampling error)**

Mean of Sample Distribution Peaked close to the Population mean, for sufficiently large Sample sizes

Thoughtful Data collection → Randomizing samples (Minimizes **Bias)**

In real world, Sampling Distributions are almost Never observed

THE CENTRAL LIMIT THEOREM

LET $(\boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_n})$ BE A RANDOM SAMPLE OF SIZE $'n'$ FROM A DISTRIBUTION WITH A FINITE MEAN $(\boldsymbol{\mu})$ AND A FINITE VARIANCE $(\boldsymbol{\sigma^2})$. IF $'n'$ IS SUFFICIENTLY LARGE, THEN $\overline{X} = \frac{X_1+X_2+\dots+X_n}{n}$ IS APPROXIMATELY A

$\boldsymbol{N\left(\mu, \frac{\sigma^2}{n}\right)}$ NORMAL DISTRIBUTION

- SAMPLES MUST BE INDEPENDENT; $X_i's$ ARE INDEPENDENT RANDOM VARIABLES
- n : DEPENDS ON THE POPULATION DISTRIBUTION
- MEAN OF $\overline{X}$ $(say\ \mu_{\overline{X}})$ $tends\ to\ \mu\ if\ CLT\ conditions\ hold$
- SD OF $\overline{X}$ $(say\ \sigma_{\overline{X}})$ $tends\ to\ \frac{\sigma}{\sqrt{n}}$ (STANDARD ERROR) $if\ CLT\ conditions\ hold$

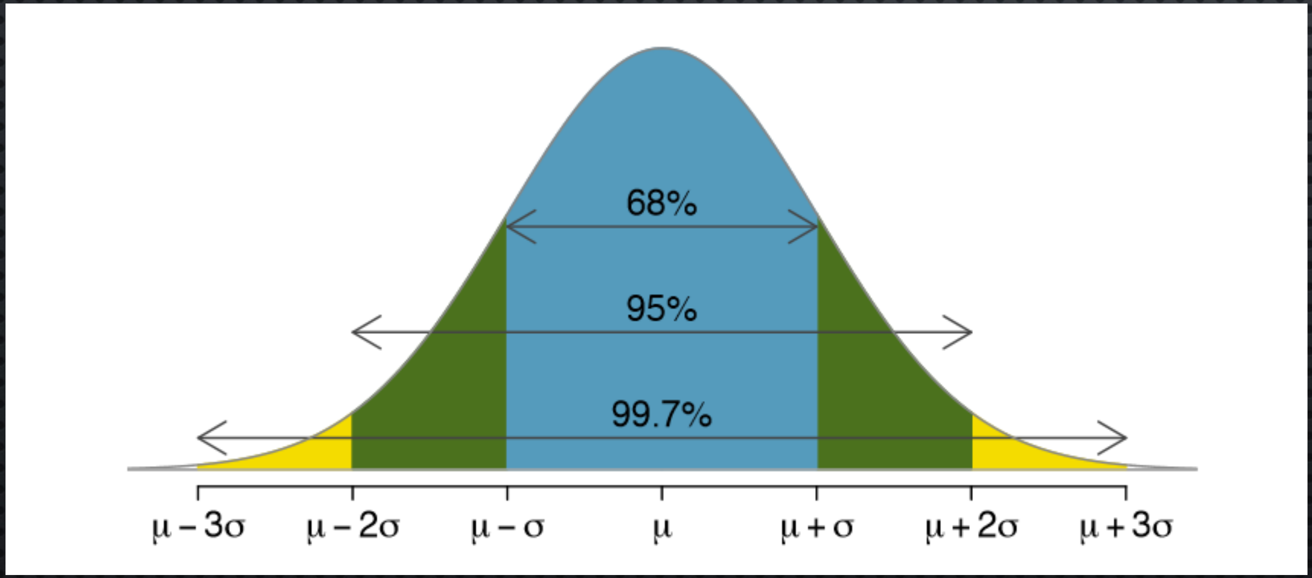
Point Estimates

NORMAL DISTRIBUTION (RECALL)

IF $\overline{X}$ IS A $N\left(\mu, \frac{\sigma^2}{n}\right)$ DISTRIBUTION. THE PDF OF $\overline{X}$:

$$pdf(\overline{X}) = \frac{1}{\sqrt{2\pi}\left(\frac{\sigma}{\sqrt{n}}\right)} e^{-\left(\frac{(\overline{X}-\mu)^2}{2\left(\frac{\sigma^2}{n}\right)}\right)}$$

$N(\mu, \sigma^2)$:



$$P\left(\mu - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) \leq \overline{X} \leq \mu + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)\right) \approx 0.95$$

Z-score

THE CLT IS VERY POWERFUL – QUITE LITERALLY

- If CLT conditions hold, Sampling Distribution (of $\bar{X}$) Is $N\left(\mu, \frac{\sigma^2}{n}\right)$
- Population Mean can Be **estimated** using sample mean $(\mu_{\bar{X}})$, but There is some $\sigma_{\bar{X}}$
  - How confident are you about the estimate based on samples You collected From the Population ?
    - **Example:** We are 95% of the times confident that the Population mean is between $(\mu - 1.96\,\sigma_{\bar{X}},\ \mu + 1.96\,\sigma_{\bar{X}})$

Interval Estimate

<u>CLT illustrates LAW of Large Numbers (LLN)</u>

Long History Short

- CLT is just a powerful extension of LLN.
- **LLN**: Cardano (16th Cent.) → Bernoulli, De Moivre(18th Cent.) → Poisson (19th Cent.) → Markov, Chebyshev, Kolmogorov, Borel… (Late 19th and 20th Cent.)
- Monte Carlo Methods (1940)

Careful: Gambler's Fallacy

Inappropriate use of LLN/CLT may lead to serious trouble

CLT in Action

- Political Polling, Product-Market Fit → Public opinion surveys
- Clinical Trials
- Forecasting Weather, Stock market
- Physics: Measurement errors, Diffusion equation (Recall: Random Walk)…

REFERENCES

- OpenIntro Statistics. Available Here
- Hogg, R.V. , Tanis, E. and Zimmerman, D. (2015) Probability and Statistical Inference. 9th Edition, Pearson, Upper Saddle River.
- For CLT applications: Investopedia