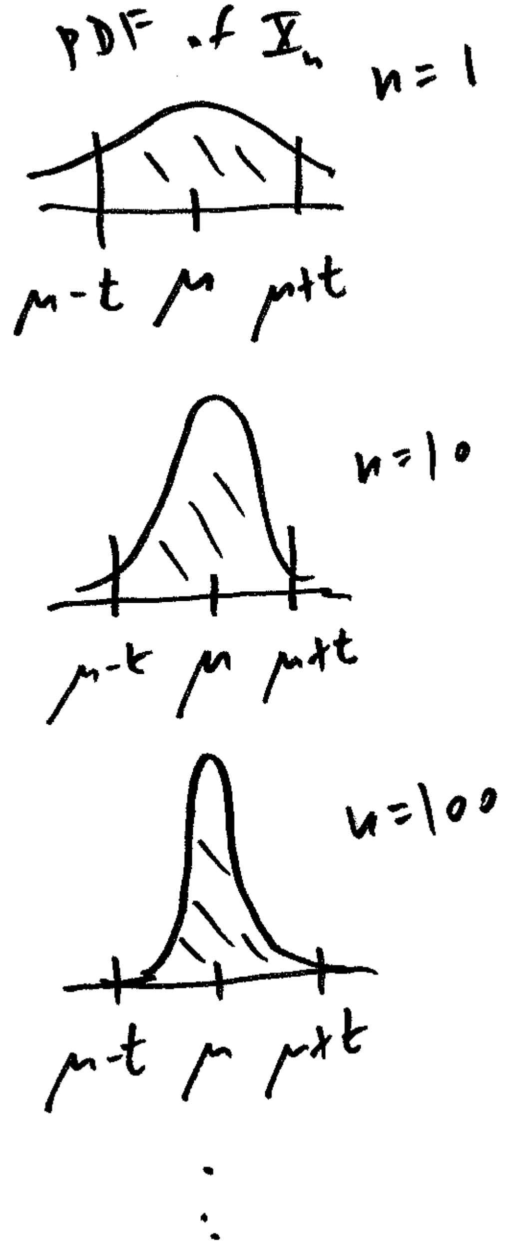


This suggests a way to quantify how close a r.v. like $\bar{X}_n$ is to a constant like μ:

Def. A sequence $Z_1, Z_2, \ldots$ of r.v. is said to converge in probability to a constant b if

for all $\epsilon > 0$, $\lim_{n \to \infty} P(|Z_n - b| < \epsilon) = 1$;

this is denoted $Z_n \xrightarrow{P} b$.

This definition is an immediate consequence of Chebyshev &

(Weak) Law of Large Numbers

$X_i \overset{IID}{\sim}$ a dist. with mean μ and variance $\sigma^2 < \infty$, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ $\rightarrow$ $\bar{X}_n \xrightarrow{P} \mu$

This result has a long history: the Italian mathematician Gerolamo Cardano (1501–1576) asserted it without proof; Jacob Bernoulli (1655–1705) proved it for $(X_i \mid \theta) \overset{IID}{\sim}$ Bernoulli(θ) (it took him 20 years to find a correct proof, published posthumously in 1713; Bernoulli thought that this theorem proved the existence of God); Siméon Denis Poisson named it the Law of Large Numbers in 1837.

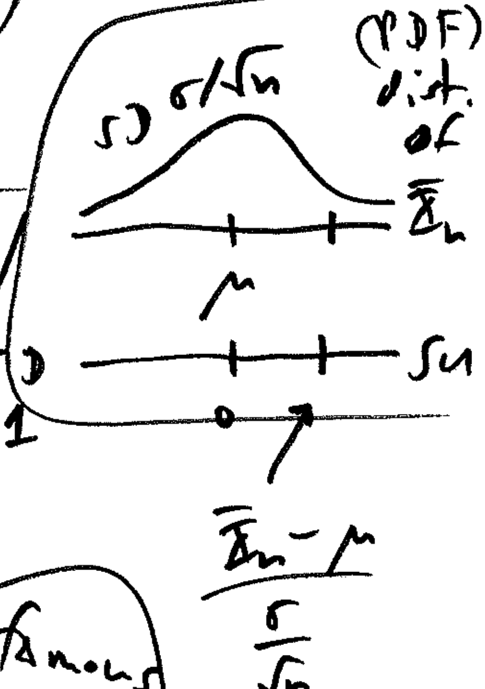
Corollary If $Z_n \xrightarrow{P} b$ and $g(z)$ is continuous at $z = b$ then $g(Z_n) \xrightarrow{P} g(b)$.

Central Limit Theorem (CLT)

Example $X_i \overset{IID}{\sim} N(\mu, \sigma^2)$, $\sigma < \infty$ $(i = 1, \dots, n)$

We know that $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ has mean μ, variance $\frac{\sigma^2}{n}$ and is normally distributed,

so that $$\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$ for all $n = 1, 2, \dots$



$\boxed{Q:}$ Does something like this work for other choices of $X_i \overset{IID}{\sim} \boxed{?}$?

$\boxed{A:}$ Yes: it's the most famous result in all of probability:

Central Limit Theorem

$X_i \overset{IID}{\sim}$ any dist. with mean μ and finite variance $0 < \sigma^2 < \infty$,

$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \rightarrow$ for large n $$\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \,\dot{\sim}\, N(0, 1)$$

Careful statement | **Def.** $X_1, X_2, \ldots$ a sequence (307)

of r.v.; let F_n be the CDF of X_n

& if there exists a CDF F^* such

that $\lim_{n\to\infty} F_n(x) = F^*(x)$ for all x at

which $F^*(x)$ is continuous, then

people say that $X_n \xrightarrow{D} F^*$ ("X_n converges in distribution to F^*")

CLT | $X_i \overset{IID}{\sim}$ any dist. with mean μ and variance $0<\sigma^2<\infty$, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

& $\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1)$. The CLT

also has a long history: it was

first demonstrated for $X_i \overset{IID}{\sim}$ Bernoulli(θ)
by the French/British mathematician
Abraham de Moivre (1667 - 1754) in
1733; almost forgotten until revived by
the French mathematician Pierre-Simon de
Laplace (1749 - 1827) in 1812; almost
forgotten again until 1901, when the
Russian mathematician Aleksandr Lyapunov
gave a more general proof; even more general
proof provided by JW Lindeberg (Finnish
mathematician (1876 - 1932)) and independently
by Paul Lévy (French mathematician (1886-
1971)) in the early 1920s. CLT name due to
Hungarian-American mathematician (1887-1985) George Pólya in 1920

$\boxed{\text{Example}}$ Contaminated water supply:

X = arsenic concentration

Y = lead concentration

(same units) $\binom{\text{both}}{\geq 0}$

Interest focuses on $R = \dfrac{Y}{X+Y}$

(proportion of contamination due to lead)

$$E(R) = E\left(\frac{Y}{X+Y}\right)$$ difficult to calculate.

$\boxed{\text{Simulation approach}}$ Randomly sample n pairs (X_i, Y_i) from the joint PDF of (X, Y), calculate $R_i = \dfrac{X_i}{X_i + Y_i}$ and

$$\boxed{\bar{R}_n = \frac{1}{n}\sum_{i=1}^{n} R_i}$$ ← good Monte Carlo (simulation) estimate of $E(R)$.

310

$\boxed{Q:}$ How big does n need to be to achieve a desired accuracy target?

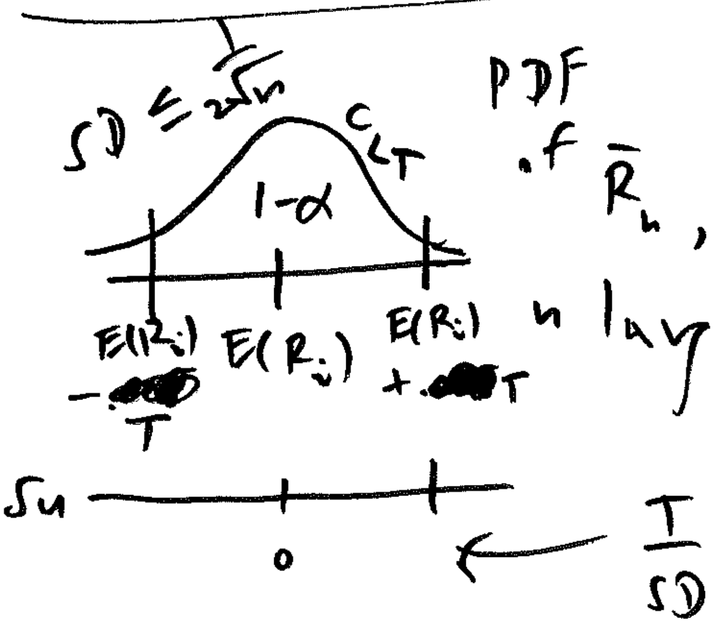
By definition

$$|R_i| = \left| \frac{I_i}{\Sigma_i + I_i} \right| \le 1 \text{ ; can show that}$$

as a result $V(R_i) \le \frac{1}{4}$.

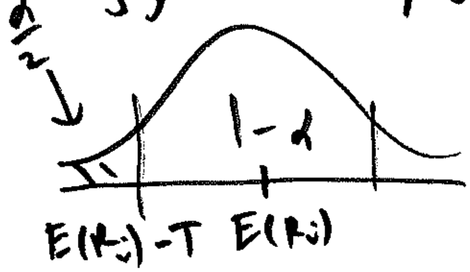
CLT says that dist. of $\bar{R}_n$ will be close to Normal for large n, with mean $E(R_i)$ and variance $\frac{V(R_i)}{n} \le \frac{1}{4n}$

Suppose we want $\bar{R}_n$ to differ from $E(R_i)$ by no more than some tolerance T with probability at least $(1-\alpha)$ ...



PDF of $\bar{R}_n$, n large

$SD \leq \frac{1}{2\sqrt{n}}$, so $\frac{1}{SD} \geq 2\sqrt{n}$ and



$$\frac{-T}{SD} \leq 2T\sqrt{n}$$

$$\Phi^{-1}\left(\frac{\alpha}{2}\right) = \frac{[E(R_i)-T] - E(R_i)}{SD} = \frac{-T}{SD} \leq 2T\sqrt{n}$$

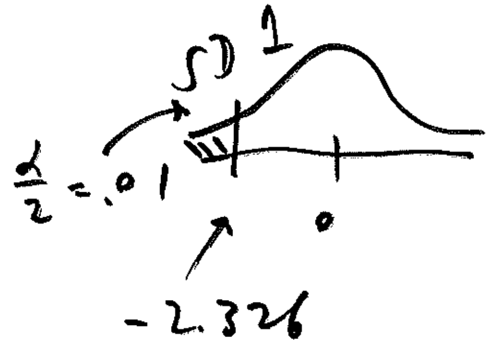
from which $n \geq \left[\frac{\Phi^{-1}\left(\frac{\alpha}{2}\right)}{2T}\right]^2$

For instance, set $T = 0.005$ ($\frac{1}{2}$ of 1%)

and $\alpha = .02$ to get

$$n \geq \left[\frac{-2.326}{2(.005)}\right]^2 \doteq 54{,}119$$

simulation replications needed



Case study: Escalators in the London Underground