Law of large numbers and central limit theorem

Suppose the rvs $X_1, \ldots X_n$ are **independently and identically** distributed (i.i.d.) as X where $\mu = E(X)$ and $\sigma^2 = \text{var}(X)$.

$$\boxed{\begin{array}{l}\text{Population: } X \sim f_\theta(x) \\ \mu = E[X],\ \sigma^2 = \text{var}[X]\end{array}}$$

$$\downarrow$$

$$X_1, X_2, \_\_\_, X_n$$

- $X_1, \ldots, X_n$ represent a **random sample** of size n from the population represented by X.

- Define **sample sum**, $T = \sum_{i=1}^{n} X_i$
- (verify) $E[T] = n\mu$, $\text{var}[T] = n\sigma^2$
- Define **sample average**, $\overline{X} = T/n$
- (verify) $E[\overline{X}] = \mu$, $\text{var}[\overline{X}] = \frac{\sigma^2}{n}$

always true.

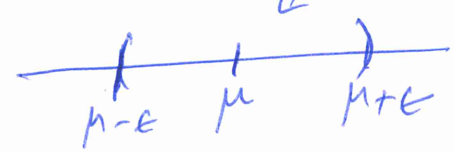
distribution of sample-based quantities is also called "sampling distribution".

Law of large numbers (LLN)

- As $n \to \infty$, $\overline{X} \to \mu$.
- If n is large, $\overline{X} \approx \mu$, i.e.,

$\overline{X}$ falls in this small interval with large prob.

$\epsilon > 0$



Central limit theorem (CLT)

Same set-up as LLN.

- If n is large, $T \overset{\cdot}{\sim} N[\underset{E(T)}{n\mu}, \underset{\mathrm{var}(T)}{n\sigma^2}]$, i.e.,

$$Z = \frac{T - n\mu}{\sqrt{n}\,\sigma} \overset{\cdot}{\sim} N(0,1)$$

- Accuracy of approx. depends on shape of $f(x)$.
- Often: $n \geq 30$ is good enough — Rule of thumb.
- Equivalently, if n is large, $\bar{X} \overset{\cdot}{\sim} N[\underset{E(\bar{X})}{\mu}, \underset{\mathrm{var}(\bar{X})}{\sigma^2/n}]$, i.e.,

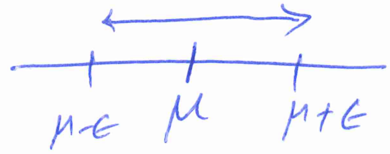
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{\cdot}{\sim} N(0,1).$$

Note: If X has normal distribution, then both T and $\overline{X}$ follow *exactly* normal distributions.

LLN vs. CLT:

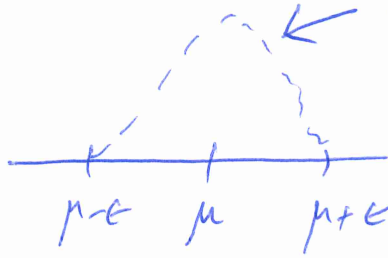
If n is large: **LLN**

$\bar{X}$ falls in this interval with large prob.



, $\epsilon > 0$

CLT:

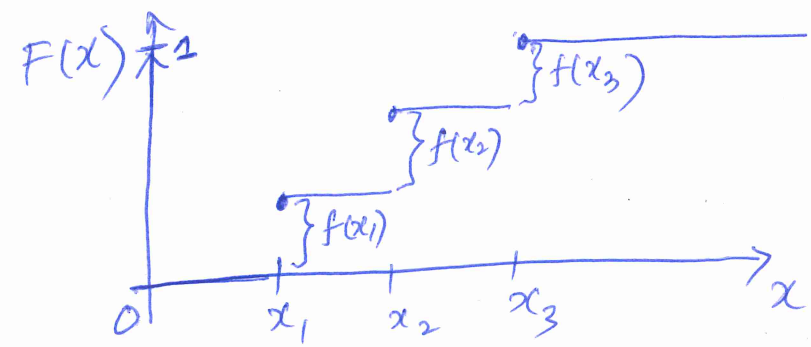


shape of dist. of $\bar{X}$.

Cumulative distribution function of X

Cumulative distribution function (cdf) of X:

- $F(x) = P[X \le x]$, $x \in \mathbb{R}$
- Nondecreasing function of x
- One to one correspondence with pdf/pmf
- Plays a key role in simulation of random variables

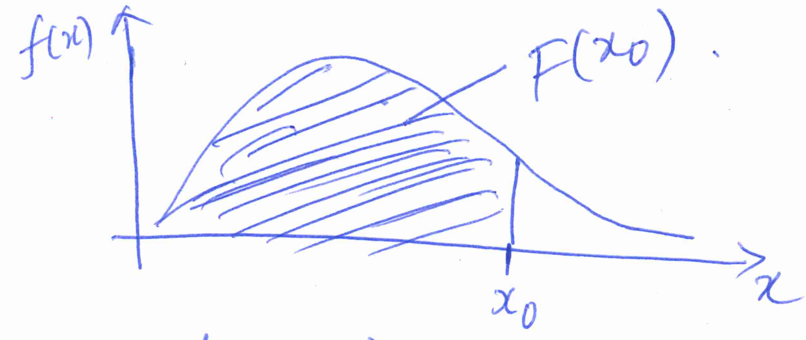


Discrete X with pmf $f(x)$

- $F(x) = \sum_{y \le x} f(y)$
- Jump (or step) function of x
  Right-cont. function
- *Getting pmf from cdf*: points of jump = possible values of X; sizes of jumps = probabilities

Continuous X with pdf $f(x)$

- $F(x) = \int_{-\infty}^{x} f(y)\,dy$
  - continuous.
- Increasing function of x
- *Getting pmf from cdf:* $f(x) = \frac{d}{dx} F(x)$.



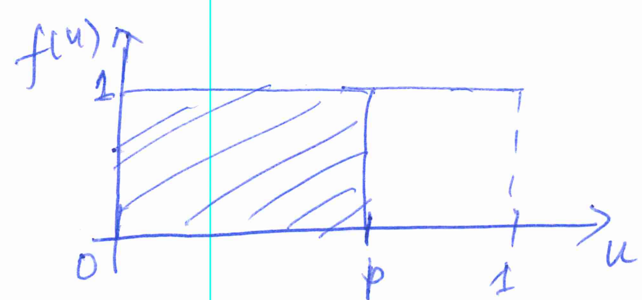
Computer simulations and Monte Carlo Methods (chapter 5)

HW: Go to `r-project.org` and download and install R. Start learning how to use it by reading the *Introduction to R* manual posted on eLearning or other resources on the web.

Assume that we can simulate $U \sim \text{Uniform}(0, 1)$. Recall that:

cont. dist.

$$f(u) = \begin{cases} 1, & 0 \le u \le 1 \\ 0, & \text{otherwise} \end{cases}$$



$F_U(p) = P[U \le p] = p.$

Every programming language has a *random number generator* that simulates a U. In R, this function is `runif()`. Subsequent calls to this function will give draws that are "independent" for all practical purposes. – "pseudo-random #'s".

Simulating from discrete distributions

Simulating $X \sim$ Bernoulli(p):

Recall: If $X \sim$ Bernoulli(p), $P(X = 1) = p$, $P(X = 0) = 1 - p$.

1. Generate U.
2. If $U \leq p$; set $X = 1$, else set $X = 0$.

Verification:

$$P[X=1] = P[U \leq p] = p.$$

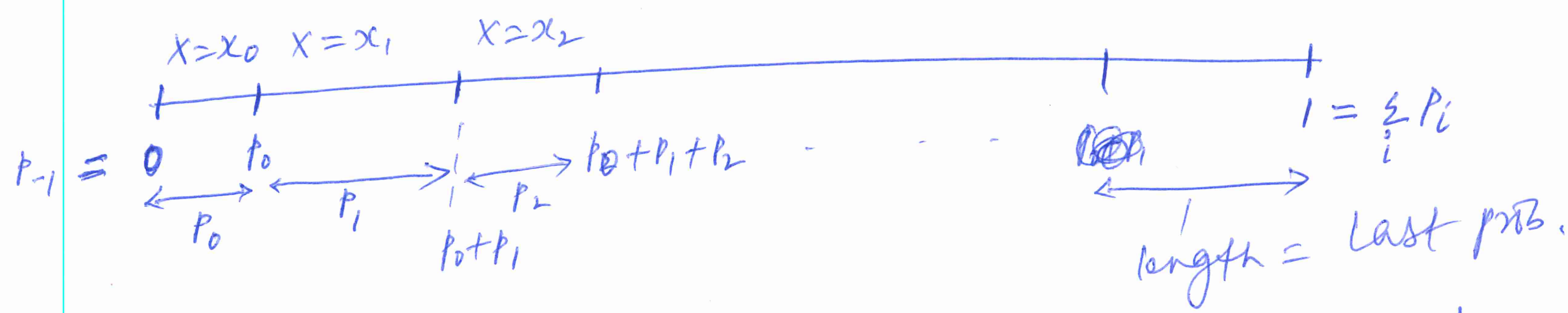
$$\Rightarrow P[X=0] = 1-p.$$

- We can now simulate draws from Binomial(n,p) and Geometric(p) distribution — just use their def. and connections with Bernoulli r.v's.

Simulating $X \sim f(x)$, arbitrary discrete distribution:

Suppose X takes values $x_0, x_1, \ldots$, with probabilities $p_0, p_1, \ldots$, where $p_i = P(X = x_i)$ and $\sum_i p_i = 1$.

1. Divide the interval $[0, 1]$ into subintervals as shown below.



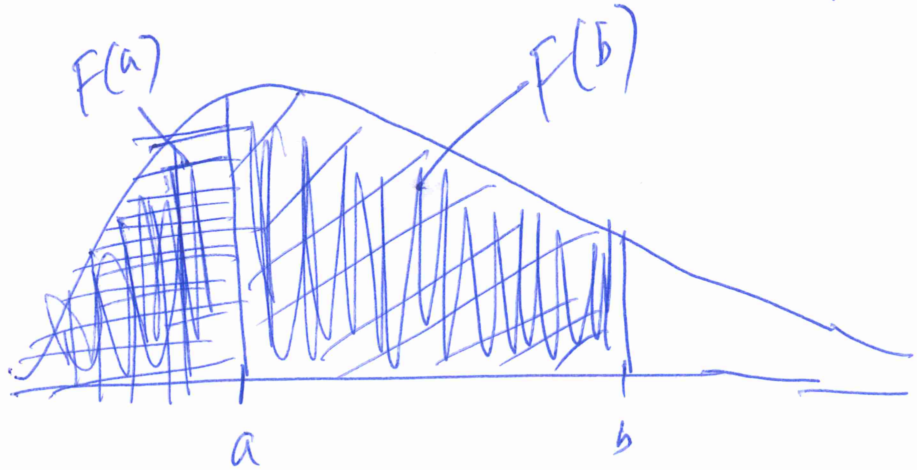
2. Generate U.
3. If U falls in subinterval i, take $X = x_i$.

- General purpose algorithm
- Possible to do better than this.

Verification:

$$P[X = x_i] = P[U \text{ falls in } i\text{-th subinterval}]$$
$$= P[p_0 + p_1 + \cdots + p_{i-1} < U \le p_0 + p_1 + \cdots + p_i]$$
$$= P[U \le p_0 + p_1 + \cdots + p_i] - P[U \le p_0 + p_1 + \cdots + p_{i-1}]$$
$$= (p_0 + \cdots + p_i) - (p_0 + p_1 + \cdots + p_{i-1}) = p_i$$

$$P[a < X \le b] = P[X \le b] - P[X \le a]$$
$$= F(b) - F(a)$$



Simulating from continuous distributions

Result: If X is a continuous rv with cdf $F(x)$, then $U = F(X)$ follows Uniform(0, 1) distribution.

<u>Inverse transform method</u>: To simulate a X,

1. Generate U.
2. Set $U = F(X)$
3. Solve for X (i.e., invert the cdf), i.e., $X = F^{-1}(U)$.

Often the equation cannot be solved explicitly or efficiently. Alternatives are available.

Simulating from Exponential(λ) distribution:

Recall: If $X \sim \text{Exponential}(\lambda)$, $F(x) = 1 - \exp(-\lambda x)$, $x > 0$.

Solve: $U = F(X) = 1 - e^{-\lambda X}$

$$\Rightarrow e^{-\lambda X} = 1 - U$$

$$\Rightarrow \log[e^{-\lambda X}] = \log[1-U]$$

"natural log

$$\Rightarrow -\lambda X = \log[1-U]$$

$$\Rightarrow X = -\frac{1}{\lambda}\log[1-U].$$

→ Doesn't always work. But other algorithms can be designed. (See book).

Solving problems by Monte Carlo methods

Estimating $\mu = E(X)$ and $\sigma^2 = var(X) = E(X-\mu)^2$:

Simulate a large number (N) of draws from the distribution of X, say, $X_1, X_2, \ldots, X_N$

MC estimator of μ: $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$

- LLN: $\bar{X} \approx \mu$ because N is large.
- often $N = 1000$ is good enough.

MC estimator of $E[g(X)]$ *where* g *is a given function:*

MC estimator of σ^2: