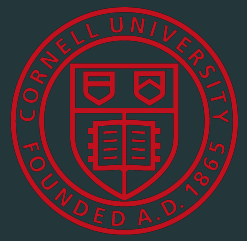


ORIE 4580/5580: Simulation Modeling and Analysis

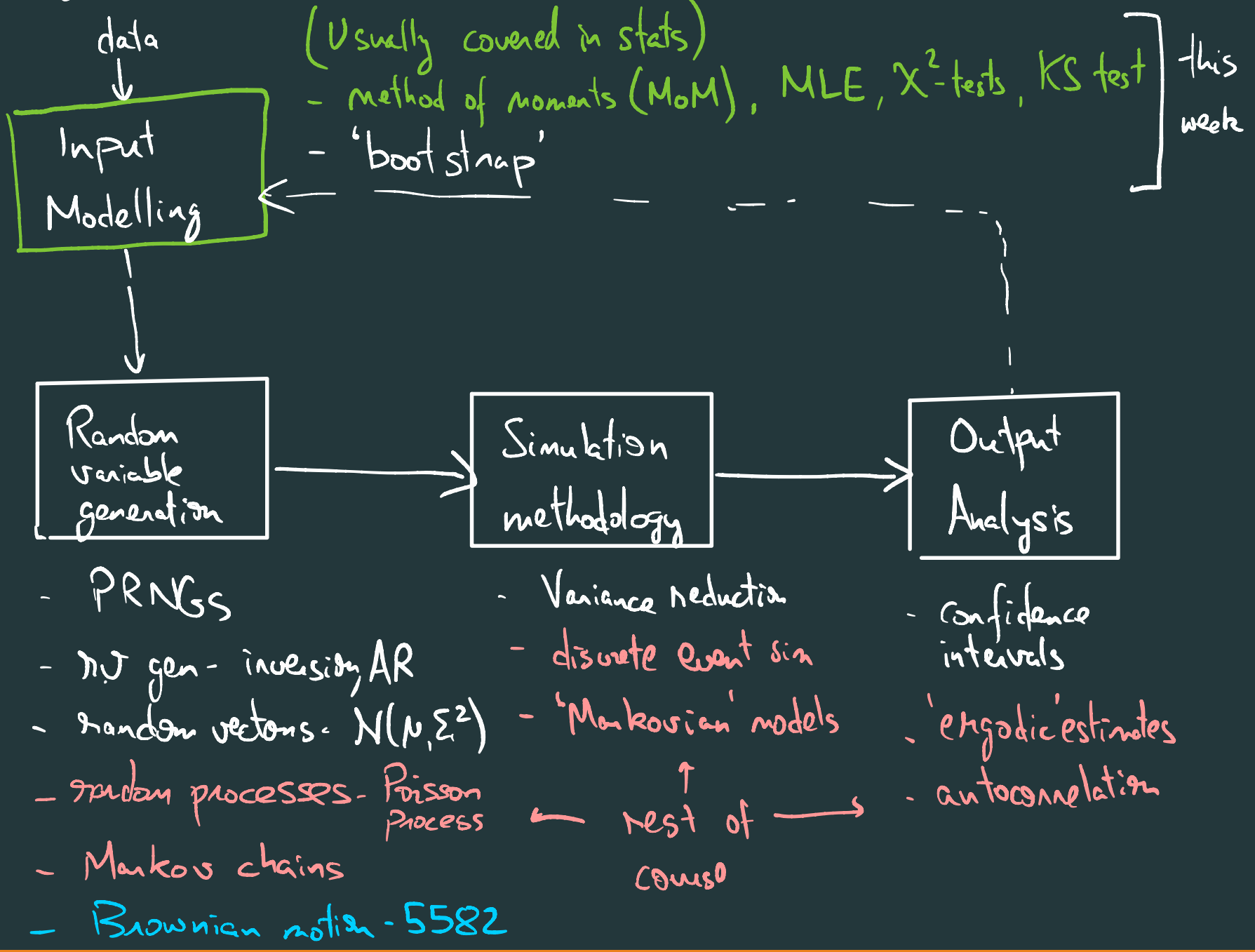
ORIE 5581: Monte Carlo Simulation

Unit 7: input modeling

Sid Banerjee

School of ORIE, Cornell University

Summary (till now + future)



input modeling

we want to answer two related questions:

- how can we use data to define the probability distributions of the 'input sequences' to a stochastic model? (real data)
- how can we determine the distribution of our simulation output? (sim data)

the basic question in both cases:

what distribution best models given data?

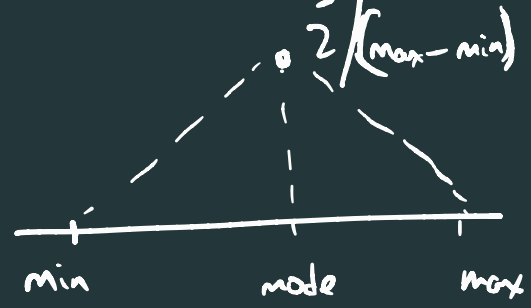
≈ 3 cases depending on how much data we have

- no / little data - use 'insights / inductive biases' | Eg. Sums of rvs ≡ Gaussian universality laws
- lots of data - bootstrap (ie, sample from data)
- medium data (relative to model size) - Parametric methods [MoM, MLE
  - non-parametric methods, Eg-kernel

case 1: no data

Eg - range (max, min) ⇒ Use Unif [min, max]

- min, max, mode
  most likely outcome



- occurs when
  - no previous records
  - introduction of new operating policy
- **approach**: use the triangular distribution
  3 parameters: minimum, mode (i.e., most likely) and maximum
  note: most likely value $\neq$ mean!
- other distributions (uniform/exponential/beta) can be used in this context
- be creative!
  - 'physics' of distributions

case 2: huge amount of data

- nowadays, many settings are in the big data regime
- if lots of 'clean' data available:
  approach: use data directly in a simulation model via bootstrapping (i.e., resampling data uniformly with replacement)

the bootstrap

we are given dataset $(X_1, X_2, \ldots, X_n)$ of n iid observations
to get new samples $(Y_1, Y_2, \ldots, Y_\ell)$, we
– generate ℓ iid indices $(I_1, I_2, \ldots, I_\ell)$ uniformly from $\{1, 2, \ldots, n\}$
– output $(Y_1, Y_2, \ldots, Y_\ell)$ where $Y_k = X_{I_k}$
the histogram of $(Y_1, \ldots, Y_\ell)$ is called the bootstrap distribution.

output ℓ datapoints uar, with replacement

warning: one should regard historical data with suspicion!

- bootstrap data is NOT iid (but close to iid if $\ell \ll n$) — boot strap samples
- in practice, if n is really large $\Rightarrow$ hard to store dataset in memory
- need to convince yourself data is iid

case 3: moderate amount of data (amount of data ≈ number of parameters)

- reasonable amount of data, but not enough for bootstrap
- **approach**: (one approach)

– fit data to a parametric family of distributions - distrib with some 'input params'
(Normal, exponential, Weibull, binomial, Poisson) Unif[a,b], $N(\mu, \Sigma^2)$
– determine parameters of selected distribution from the data
– use the fitted distribution to generate samples for simulation

- important questions:

1. how to choose the family of distributions? (brief)
2. how to select the parameters of the distribution? - MoM, MLE
3. how to assess the fit of the distribution and the parameters?

- simplifying and major (!) assumption – i.i.d. samples

choice of distribution families

to pick the appropriate family of distributions:

– **capture the physics**

'story' behind different distributions

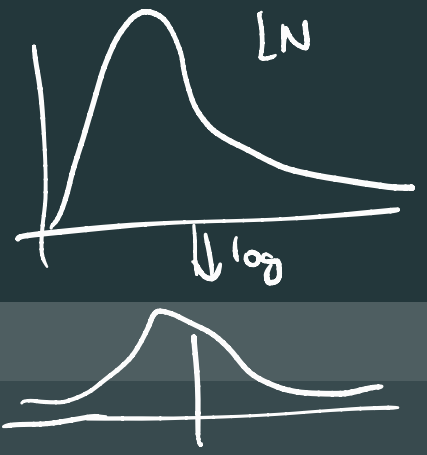
– sum of ~~iid~~ indep rvs $\implies$ Normal distribution (CLT)

– product of iid rvs $\implies$ Lognormal distribution (CLT $\Rightarrow$ $\log(\prod) \sim$ normal)

– max/min of iid rvs $\implies$ ~~Weibull~~ Weibull (Extreme-value) distribution

– superposition of independent arrivals $\implies$ Poisson process (later)

– use visual tests to guide distribution choice



physics of different distributions

normal distribution

- if X is the sum of a large number of other random quantities, i.e.

$$X = Y_1 + \ldots + Y_n$$

  then X can be approximately modeled as a normal random variable
- central limit theorem $\implies$ $X \approx \mathcal{N}(0, 1)$ for large n
- *example* – total value of claims received by insurance company in one day

lognormal distribution

- if $W \sim \mathcal{N}(0,1) \implies e^W$ is log-normally distributed.
- moral – if X is the product of a large number of other random quantities, i.e.

$$X = Z_1 Z_2 \ldots Z_n,$$

  or alternately, $\ln X = \ln Z_1 + \ln Z_2 + \ldots + \ln Z_n$); then X can be approximately modeled as a log-normal random variable
- *example* – many financial asset models:
  $G_n =$ proportional change in asset value during time period n
  $W_n =$ net worth of an asset at the beginning of time period n

Weibull distribution

- system that is made up of n components with lifetimes $Y_1, \ldots, Y_n$
  let L be the lifetime of the system:
  – components are connected in series:$L = \min[Y_1, \ldots, Y_n]$
  – components are connected parallel: $L = \max[Y_1, \ldots, Y_n]$
- extreme value theory: approximate distribution of L when n is large and $Y_1, \ldots, Y_n$ are i.i.d. random variables.
  – L has approximately Weibull distribution when n is large.
- *example* – lifetime of complicated system is approximately Weibull

others

- *Geometric*(p): number of coin tosses before heads/number of independent trials till success, with $p = \mathbb{P}[\text{success}]$
  (only memoryless discrete distn)
- *Binomial*(n, p): number of successes in n independent trials, where $p = \mathbb{P}[\text{success}]$
- *Poisson*(λ): number of outcomes in a very large number of independent trials ($n \to \infty$), where $\mathbb{P}[\text{outcome}] = \lambda/n$ is very small (for example, spontaneous radioactive emissions in a material over a day, positive COVID tests in a large population, etc.); mean number of successes λ
- *Exponential*(λ): good model for 'holding' times/inter-arrival times/delays, with mean $1/\lambda$ (only memoryless continuous distn)

Memorylessness - $X \sim F$ is said to be memoryless if $X | X \geq s$ has same dist^n as $X + s$

($X|X \geq s$: CDF; $\hat{X}$: iid with X)

$$\mathbb{P}[X \leq s+t \mid X \geq s] = \mathbb{P}[\hat{X} \leq t]$$

- **Claim** - only 2 distns are memoryless $\equiv$ discrete $\equiv$ Geometric, cont $\equiv$ exponential
- $X \sim \text{Exp}(\lambda) \equiv f_\lambda(x) = \lambda e^{-\lambda x}$, $F_\lambda(x) = 1 - e^{-\lambda x}$, $x \geq 0$

$$\mathbb{P}[X \leq s+t \mid X \geq s] = \frac{F(s+t) - F(s)}{1 - F(s)} \leftarrow \mathbb{P}[X \geq s]$$

($F(s+t) - F(s) = \mathbb{P}[s \leq X \leq t]$)

$$= \frac{-e^{-\lambda(s+t)} + e^{-\lambda s}}{e^{-\lambda s}} = 1 - e^{-\lambda t} = F_\lambda(t)$$

parameter fitting

parameter estimation

hypothesis: data $X_1, \ldots, X_n$ comes from parametric distribution family with cdf $F(\cdot)$

($X_1, \ldots, X_n$: iid samples)

how do we choose parameters of $F(\cdot)$?

Eg - $X_i \sim N(\mu, \sigma^2)$

two methods:

– method of moments

- 'equation-based' approach

set up system of equations for params θ

solve for θ exactly

– maximum likelihood estimation

'optimization-based' approach

Eg - given A, y, solve for x s.t. $y = Ax$

(can do if A invertible $\Rightarrow x = A^{-1}y$)

Eg - given A, y, find x s.t. $\min \|y - Ax\|_2$ ('least-squares')

$$\Rightarrow x = A^{\dagger} y = \underbrace{(A^T A)^{-1} A^T}_{\text{always exists!}} y$$

method of moments: definition

k^{th} moment of $X \triangleq \mathbb{E}[X^k]$

want to fit data to cdf $F(\cdot)$ with p unknown parameters

method of moments

1. using data $(X_1, X_2, \ldots, X_n)$, estimate the first p empirical moments. Let $m_1, \ldots, m_p$ be the estimated moments, where

$$m_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

2. compute the first p moments of the hypothesized p.d.f

let $\mu_1, \ldots, \mu_p$ be these exact moments, where

$$\mu_k = \mathbb{E}[X^k] = \int_{-\infty}^{\infty} x^k f(x)\,dx$$ (typically, look up Wiki)

3. set $\mu_k = m_k$ for $k = 1, \ldots, p$, and solve these p equations for the p unknown parameters

MOM for exponential rv

- given 5 interarrival times: 3, 1, 4, 3, 8
- want to model this as being from an exponential distribution
- recall: mean of $Exp(\lambda)$ rv is $1/\lambda$.

$$\Rightarrow \quad \frac{1}{\lambda_{MoM}} = \frac{3+1+4+3+8}{5} \quad \Rightarrow \lambda_{MoM} = \frac{5}{19}$$

$$\mathbb{E}[X] \qquad \frac{1}{n}\sum_{i=1}^{n} x_i$$

MOM for Normal rv

example – hypothesis: $X_1, \ldots, X_n$ are i.i.d. samples from $\mathcal{N}(a, b^2)$

$$\text{Var}(X) = \mathbb{E}[X^2] - [\mathbb{E}X]^2$$

- $$a = \frac{1}{n}\sum_{i=1}^{n} X_i = M_1$$

$$b^2 + a^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = M_2$$

$$\Rightarrow \quad a = M_1, \; b = \sqrt{M_2 - M_1^2}$$

must be ≥ 0 for meaningful answer

Things may go wrong

- data - 1.1, 0.9, 1, 5, want to fit $U[0, a]$

$$\Rightarrow m_1 = \frac{8}{4} = 2, \quad \mathbb{E}[X] = \frac{a}{2}$$

$$\Rightarrow \frac{a}{2} = 2 \Rightarrow \underline{a = 4}$$

'valid' MoM estimator

Note - MoM does poorly for small n

MOM estimator: pros and cons

- advantage: easy to setup, and most of the time gives *some* answer
- con: answers not always very meaningful
- tl;dr: use MoM when more sophisticated procedures fail!

maximum likelihood estimation

fit i.i.d data $D = (X_1, X_2 \ldots, X_n)$ to cdf $F(\cdot)$ with unknown parameters Θ

likelihood function (objective fn for maximization)

$L(\Theta|D)$: measure of how well parameters Θ 'explain' given data D

– function of Θ (not a probability distribution)

– for discrete r.v., $L(\Theta|D) = \Pi_{i=1}^{n} p(X_i|\Theta)$

– for continuous r.v., $L(\Theta|D) = \Pi_{i=1}^{n} f(X_i|\Theta)$

NOT a probability

$L(\theta|D)$ = some fn of θ

maximum likelihood estimation

1. using data $D = (X_1, X_2, \ldots, X_n)$, define likelihood function $L(\Theta|D)$
2. find Θ which maximizes the log-likelihood, i.e.

$$\Theta^* = \arg\max \ell(\theta|D) = \sum_{i=1}^{n} \log\left(L(\theta|D)\right)$$

Clicker question: MoM for uniform

unknown

$U[-a, a]$

uniform random variable on $(-a, a)$ has mean 0 and variance $a^2/3$

given sample moments m_1 and m_2, an MOM estimator for a is

(a) $a = 0$

$\frac{\sum x_i}{n}$ $\frac{\sum x_i^2}{n}$

(b) $a = m_1$

(c) $a = \sqrt{3m_2}$

(d) No MoM estimator is possible

some function of θs — k^{th} empirical moment

$\mathbb{E}[X^k] = m_k$

MoM – enough eqns to solve for unknown params

- For this problem

$\mathbb{E}[X] = \boxed{0 = m_1}$ can not solve for a

$\mathbb{E}[X^2] = \underbrace{\mathrm{Var}(X)}_{a^2/3} + \underbrace{(\mathbb{E}[X])^2}_{0} = a^2/3$

$\Rightarrow a^2/3 = m_2 \Rightarrow a = \sqrt{3m_2}$

MLE: exponential rv

hypothesis: interarrival times $3, 1, 4, 1, 8$ are i.i.d. samples from $Exp(\lambda)$

- $L(\lambda | x_i) = \lambda e^{-\lambda x_i} \Rightarrow \ell(\lambda | x_i) = \ln(\lambda) - \lambda x_i$

$$\Rightarrow \log\left(L(\lambda | D)\right) = \sum_{i=1}^{n} \ell(\lambda | x_i) = n\ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

diff^{a}, we get

$$\frac{1}{\lambda_{MLE}} = \frac{\sum_{i=1}^{n} x_i}{n} = m_1 = \frac{1}{\lambda_{MoM}}$$

MLE for exponential

hypothesis: interarrival times $X_1, X_2, \ldots, X_n$ are i.i.d. samples from $Exp(\lambda)$

MLE: Geometric rv

$p \in [0,1]$

hypothesis: $X_1, \ldots, X_n$ are i.i.d. samples from $Geom(p)$ distribution

$$L(p \mid X_i) = p(1-p)^{X_i - 1} \qquad (\text{for } X_i \in \{1, 2, \ldots \infty\})$$

$$\Rightarrow \ell(p \mid D) = n \ln(p) + \left(\sum_{i=1}^{n} X_i - n\right) \ln(1-p)$$

Setting $\frac{\partial \ell}{\partial p} = 0 \Rightarrow \frac{n}{p} = \frac{\sum X_i - n}{1-p}$

$$\Rightarrow \frac{1}{p_{MLE}} = \frac{\sum_{i=1}^{n} X_i}{n} = m_1 = \frac{1}{p_{MoM}}$$

MLE: Normal rv

hypothesis: $X_1, \ldots, X_n$ are i.i.d. samples from $N(\mu, \sigma^2)$

- $\ell((\mu,\sigma)|X_i) = \ln\left(\frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma} \cdot e^{-(x_i-\mu)^2/2\sigma^2}\right)$

$$\Rightarrow \ell((\mu,\sigma)|D) = -n\ln(\sqrt{2\pi}) - n\ln(\sigma) - \underbrace{\sum_{i=1}^{n}(x_i-\mu)^2}_{A} \Big/ 2\sigma^2$$

Want $\frac{\partial \ell}{\partial \mu} = 0 : \sum_{i=1}^{n} \frac{2(x_i-\mu)}{2\sigma^2} = 0 \Rightarrow \mu_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}$

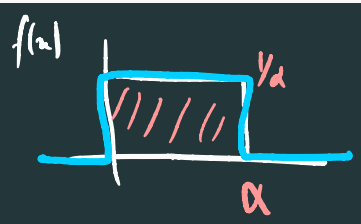
$\frac{\partial \ell}{\partial \sigma} = 0 : -\frac{n}{\sigma} + \frac{A}{\sigma^3} = 0 \Rightarrow \sigma^2_{MLE} = \frac{\sum_{i=1}^{n}(x_i - \mu_{MLE})^2}{n}$

Note: not unbiased

(:) However, it is the MoM estimator

MLE: uniform rv

hypothesis: $X_1, \ldots, X_n$ are i.i.d. samples from $U[0, \alpha]$

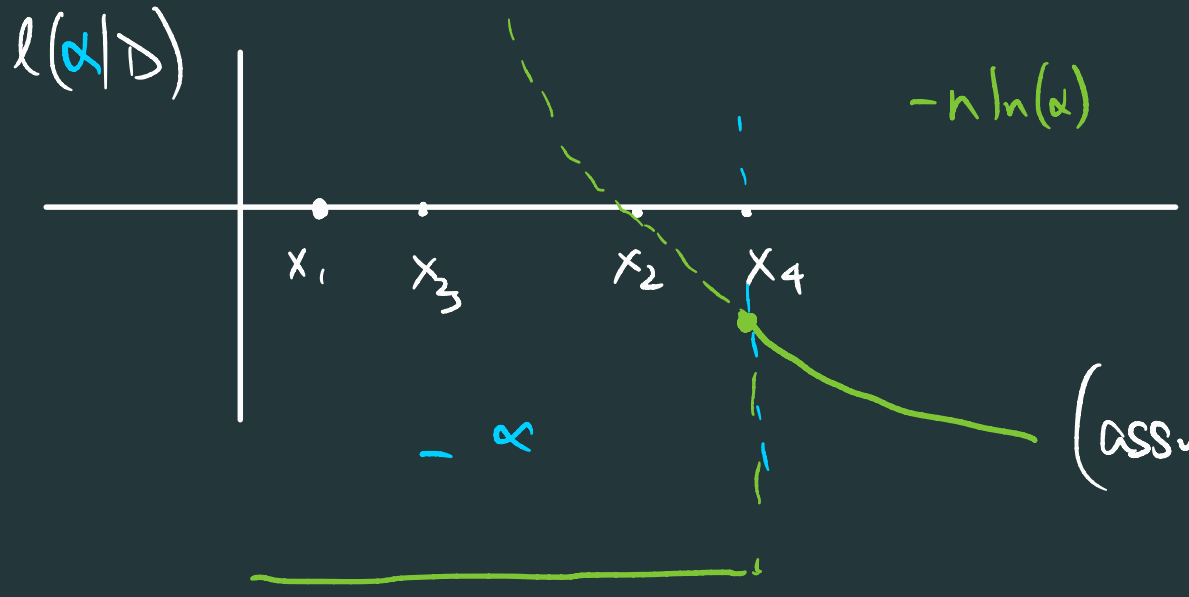


$$L(\alpha | x_i) = \begin{cases} 1/\alpha & ; \quad x_i \in [0, \alpha] \\ 0 & ; \quad x_i \notin [0, \alpha] \end{cases}$$

$$\Rightarrow \ell(\alpha | x_i) = \begin{cases} \ln(1/\alpha) = -\ln(\alpha) & ; x_i \in [0, \alpha] \\ -\infty & ; x_i \notin [0, \alpha] \end{cases}$$

$$\ell(\alpha | \mathcal{D}) = \sum_{i=1}^{n} \ell(\alpha | x_i) = \begin{cases} -n \ln(\alpha) & ; \text{ all } x_i \in [0, \alpha] \\ -\infty & ; \text{ some } x_i \notin [0, \alpha] \end{cases}$$

MLE: uniform rv



(assuming all $x_i \geq 0$)

$$\Rightarrow \quad \arg\max_{\alpha \geq 0} \ell(\alpha|D) = \max_i \{x_i\}$$

- In general, if $x_i \in \text{Unif}[\alpha, \beta]$

  then $\alpha_{MLE} = \min\{x_i\}$, $\beta_{MLE} = \max\{x_i\}$

MLE: notes

- sometimes (rarely) MLE can be computed in closed-form
- usually: compute MLE via numerical optimization
- why use MLE's?
  – they contain *all* the available statistical information about parameters in the data
  – they (asymptotically) have the smallest variance of any possible parameter estimator

Imp - In practice, often - easy to find numerically
- usually quite meaningful

goodness of fit - How do we convince ourselves/ others that our distributions are a 'good fit' for the data?

the methods I will show

Caveat - This is what statisticians say

1) Do not take the 'numbers' too seriously

2) Better approach - do all the 'tests', and present all the 'evidence' to justify choice

visualizing fit: histograms and Q-Q plots (test fit 'visually')

hypothesis: data comes from a distribution with cdf $F(\cdot)$

method 1: (visually) compare empirical histogram to hypothesized pdf

note: scale appropriately

– Δ: bin width, n: # of data points $\implies$ area under histogram $= n\Delta$

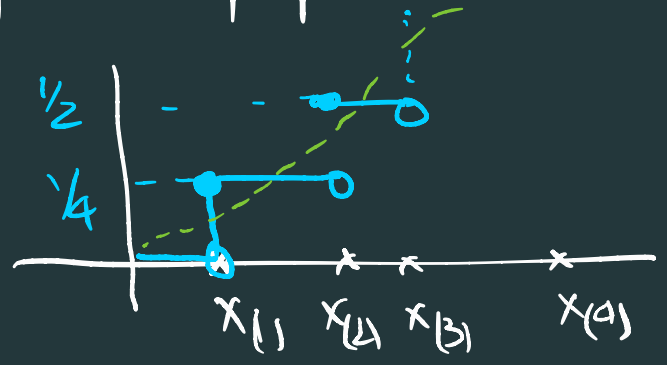
– must scale pdf by $n\Delta$ to compare

– discrete data: $\hat{p}(i)$ = fraction of times observe outcome i in data set

method 2: compare cdfs (how?) - 'empirical cdf' of $X_1, X_2, \ldots X_n$

$$\hat{F}_n(x) = \frac{\sum_{i=1}^{n} \mathbb{1}\{x_i \le x\}}{n} = \text{'fraction of pts below } x\text{'}$$

$\neq 0 \rightarrow$ Yaxis values $\rightarrow \{1/n, 2/n, 3/n, \ldots, n/n\}$ ≤ 1



Q-Q Plots

Main idea - $F_X(X) \sim U(0,1)$

- more informative visual tool
- helps understand **tails** of the distribution

QQ plot

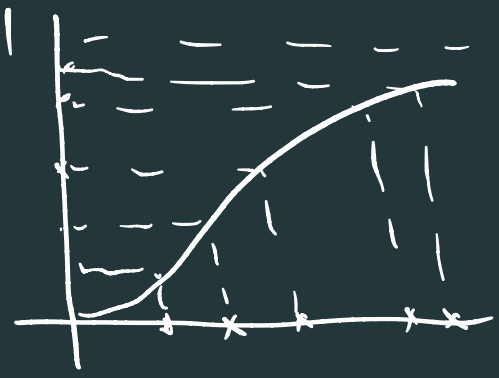
- order data in increasing order $Y_1 \leq Y_2 \leq \ldots \leq Y_n$
  – fraction of observations $\leq Y_j$ is j/n

Compute emp cdf

- empirical cdf can be defined as $\hat{F}(Y_j) = \left(\frac{j-0.5}{n}\right)$ 'correction' *
- for test cdf $F(\cdot)$, compute 'quantiles' $Z_j = F^{-1}\left(\frac{j-0.5}{n}\right)$
- Q-Q plot $\implies [(Y_j, Z_j) : j = 1, \ldots, n]$

empirical CDF

F^{-1}(quantiles of Unif)



(*) to make quantiles symmetric
- take $\left\{\frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}\right\}$ or $\left\{\frac{0.5}{n}, \frac{1.5}{n}, \ldots, \frac{n-0.5}{n}\right\}$

QQ plots: notes

- observed values will never fall exactly on the straight line
- ordered values are not independent, because we ordered them
  if one point lies above the line, the next is likely to do the same. . .
- values at extremes have much higher variance than those in the middle

goodness of fit tests

more general approach

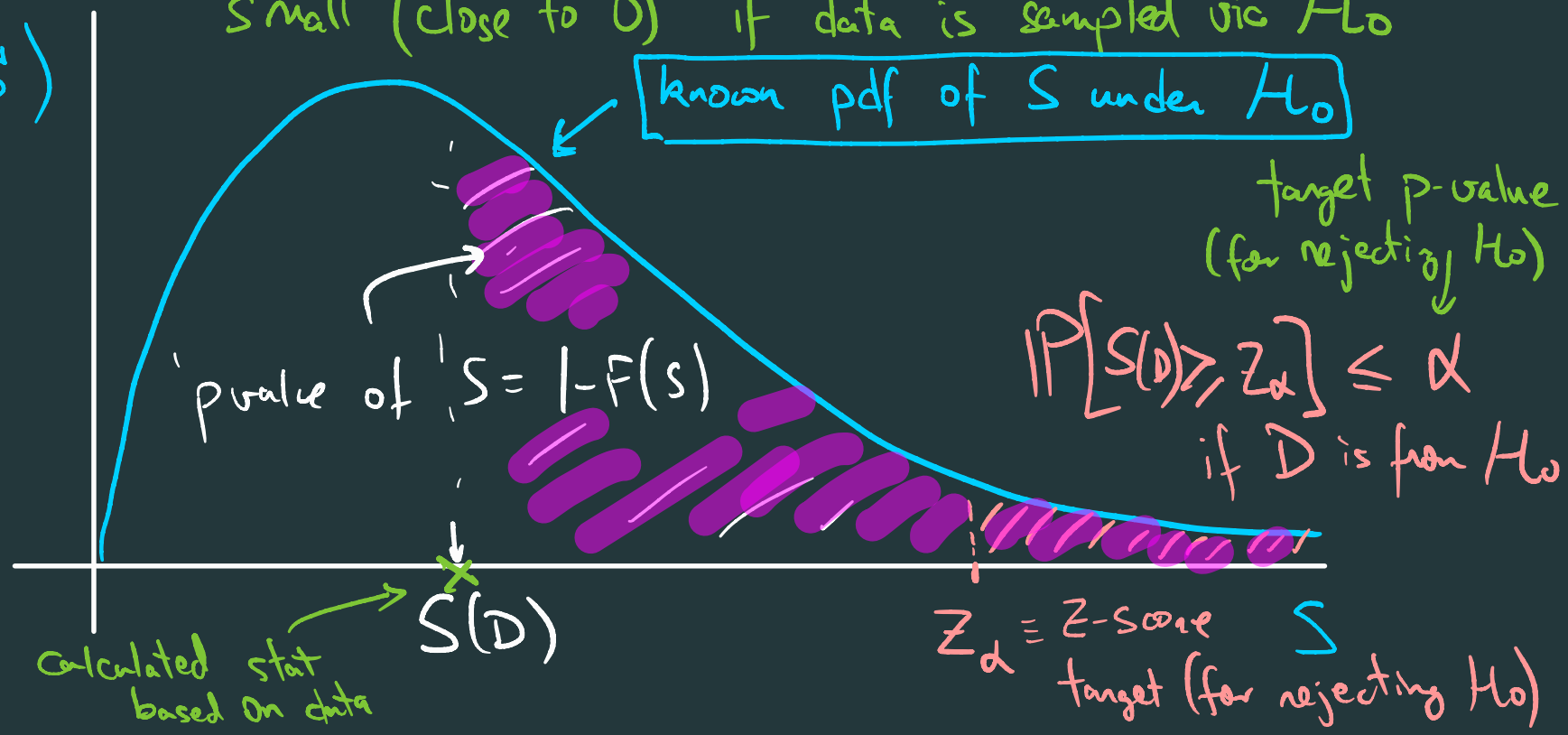
- fitting distributions = hypothesis testing

$S = g(X_1, X_2, \ldots, X_N)$

NULL H_0 : data come from the hypothesized distribution (null hypothesis)

ALT H_1 : data do not come from the hypothesized distribution.

Idea - Compute a 'statistic' S (non-negative fn of the data) that is small (close to 0) if data is sampled via H_0



Clicker question: MLE for uniform

given data $(X_1, X_2, \ldots, X_n)$ with sample moments m_1 and m_2, the MLE for α assuming the data comes from a uniform distribution over $(-\alpha, \alpha)$ is

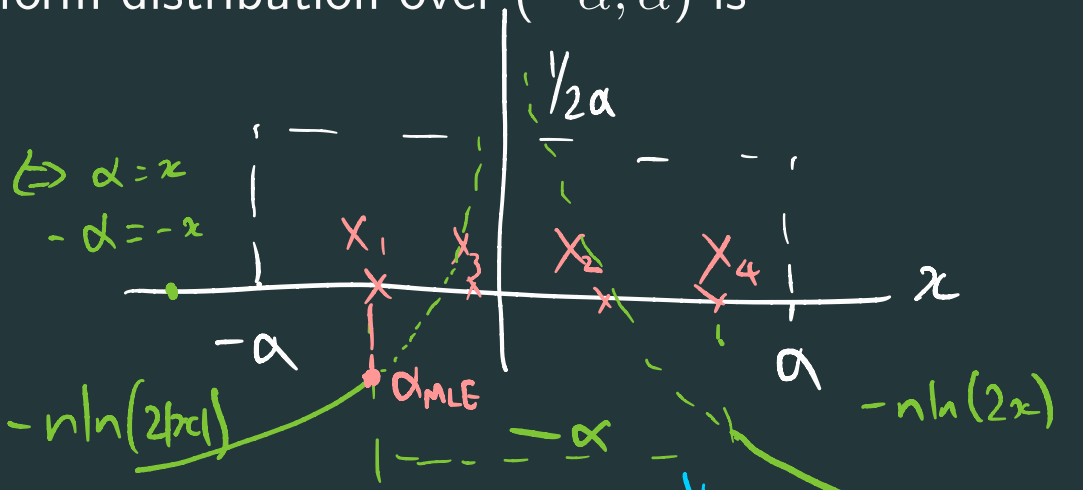
(a) $\alpha = \sqrt{3m_2}$

(b) $\alpha = \max_i X_i$] correct if $X \sim U[0, \alpha]$

(c) $\alpha = \min_i X_i$] correct if $X \sim U[-\alpha, 0]$

(d) $\alpha = \max_i |X_i|$ ✓

(e) No MoM estimator is possible



maximized at α = "furthest pt from origin"

$$\ell(x_i|\alpha) = \begin{cases} -\infty & \text{if } x_i < -\alpha \text{ or } x_i > \alpha \\ -\ln 2|\alpha| & \text{if } -\alpha \le x_i \le \alpha \end{cases}$$

$$\ell(D|\alpha) = \prod_{i=1}^{n} \ell(x_i|\alpha)$$

Input Modelling - **Aim**: Given (iid) data, try to find 'best distribution' that models the data

Till Now - 3 Regimes

- small data: use prior knowledge / inductive biases

  Eg - If min, max, mode known $\Rightarrow$ Δ distn

- very large data: use the bootstrap (# of bootstrap samples $\ll$ # of data samples)

- amount of data $\approx$ # of params $\Rightarrow$ Use Parametric models

  Choose params via $\begin{cases} \text{MoM} \leftarrow \text{solve eqns} \\ \text{MLE} \leftarrow \text{optimization} \end{cases}$

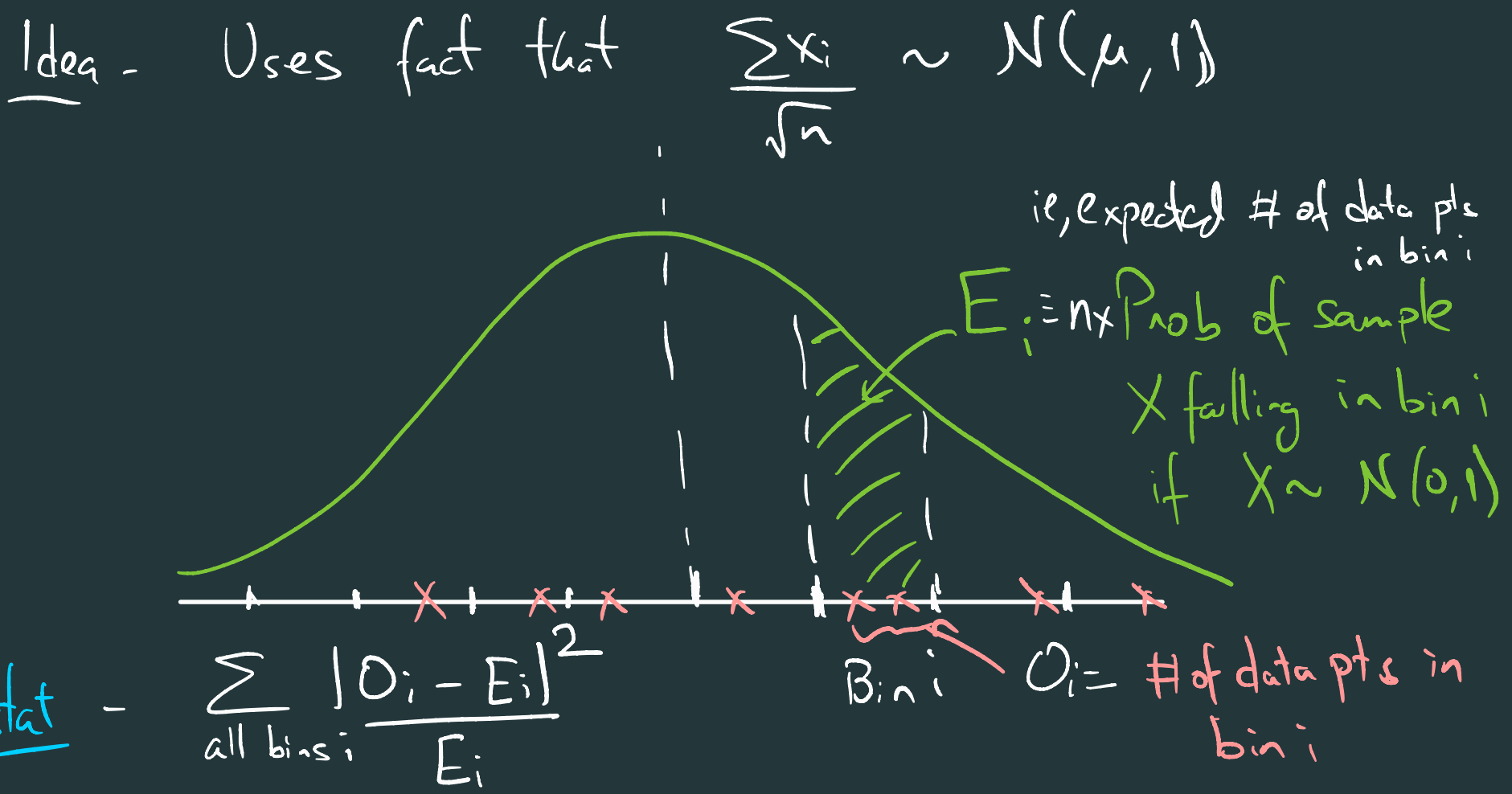
- Test 'Goodness of Fit'
  - Visual methods - plot empirical histogram vs proposed his
  - Q-Q plots

Today

- Quantitative Goodness of fit tests $\equiv$ 'hypothesis testing'
- 'Perturbation analysis' - analysing downstream effects

chi-square goodness of fit test

- can be used for discrete or continuous distributions
- compare histogram of data with expected frequencies under hypothesized distribution



chi-square goodness of fit test

chi-square test

1. choose k: number of bins

$[b_{i-1}, b_i)$: i-th bin

$[b_0, b_k]$ should cover the whole range. (leftmost is $(-\infty, b_0]$, rightmost is $[b_k, \infty)$

2. compute O_i = observed number in bin i

   E_i = expected number in bin i (under hypothesis)

3. compute the test statistic $D^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$.

4. under the null hypothesis, D^2 has (approximately) a chi-squared distribution with $df = k - s - 1$ degrees of freedom

   (s is the number of parameters estimated from the data)

4. compute $\chi^2_{df,1-\alpha} = F^{-1}_{\chi^2_{df}}[1-\alpha]$

chi-square test: example

Fit: $X_i \sim \text{Exp}(\lambda)$

example – chi-sqared test for car interarrival times.

	data	data	hypothesis	statistic
bin	cumulative	observed	expected	$(O-E)^2/E$
0, 0.05	33	33	27.674	1.024
0.05, 0.1	58	+ 25	24.330	0.018
0.1, 0.15	80	22	21.389	0.017
0.15, 0.2	90	10	18.804	4.122
0.2, 0.3	121	31	31.066	0.000
0.3, 0.5	165	44	42.570	0.048
0.5, ∞	229	64	63.163	0.011

$s = 1, \quad D^2 = 5.242, \quad \text{d.f.} = 5, \quad \chi^2_{5,1-0.05} = 11.070.$

$7 - 1 - 1$ ← initializing −1

↑ # of params

↗ # of bins

chi-square test:notes

how many bins

- range of a continuous distribution can be divided into any number of bins
- too many $\implies$ expected frequencies become small
  too few $\implies$ test has little power of discrimination
- desirable to divide the continuous range bins with equal probabilities.
  $E_i = E_j$ for all $i, j = 1, \ldots, k$.
  then, k is the only decision.

  Choose $b_i \equiv F^{-1}(i/k+1)$
  ie, quantiles
- the size of the bins should be such that $E_i \geq 5$.

p-value: $\mathbb{P}\left[X > D^2\right]$, where X is a chi-squared distributed random variable with $k - s - 1$ d.f., and D^2 is the test statistic

Kolmogorov-Smirnov (KS) test

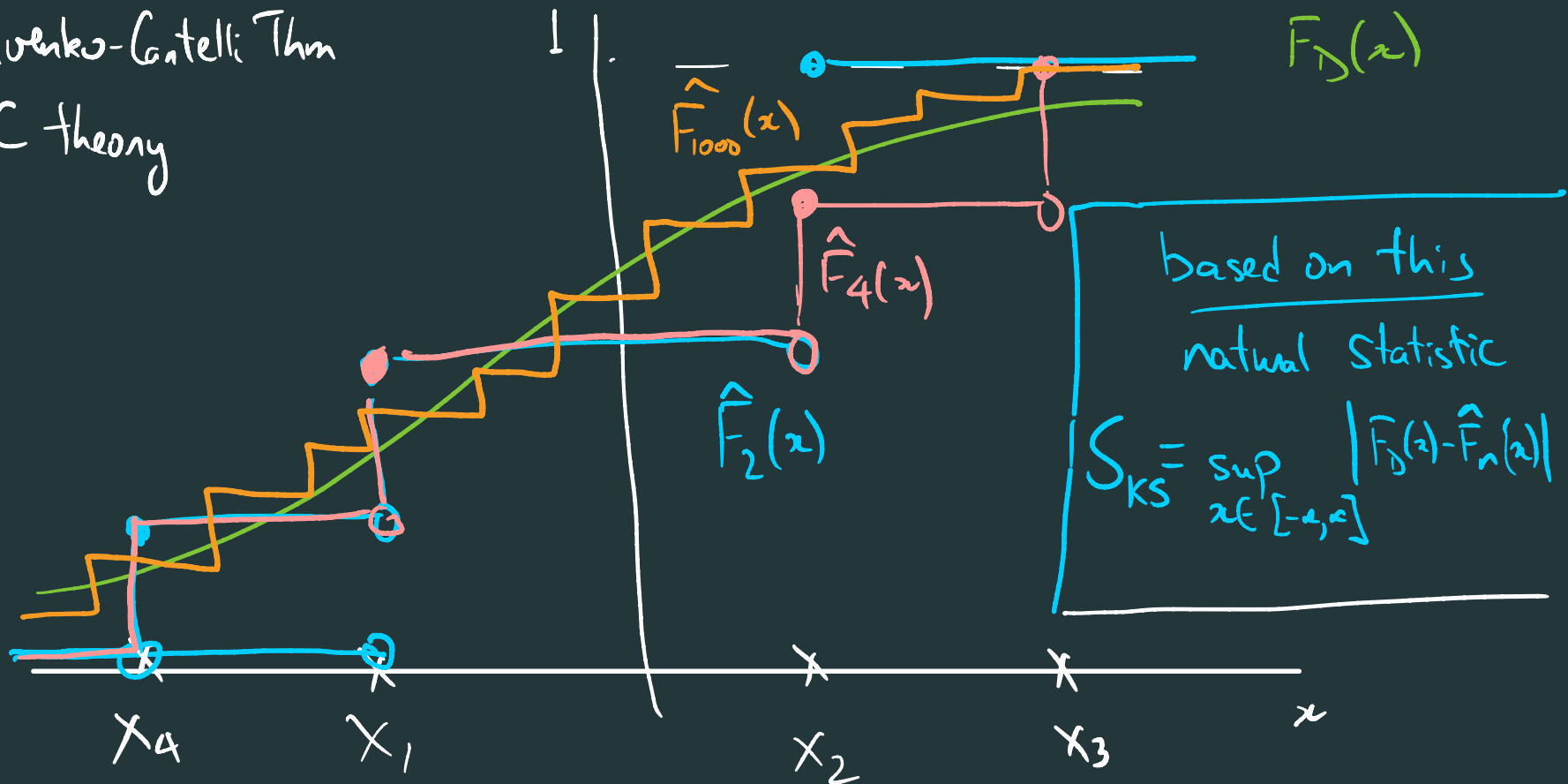
- chi-square test:
  histogram of the data $\Longleftrightarrow$ pdf of the hypothesis
- KS:
  empirical cdf of the data $\Longleftrightarrow$ cdf of the hypothesis
- advantages:
  – more discriminating power than Chi-square
  – does not require grouping the data into bins

KS Test – Tries to capture the fact that for iid $X_i \sim D$

empirical histograms $\left(\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\} \right)$

FACT → converge to the true histogram $F_D(x)$ FOR ALL x simultaneously

- Glivenko-Cantelli Thm
- VC theory



based on this natural statistic

$$S_{KS} = \sup_{x \in [-a,a]} \left| F_D(x) - \hat{F}_n(x) \right|$$

Kolmogorov-Smirnov test

KS goodness-of-fit test

- data: $(X_1, \ldots, X_n)$, hypothesis distribution with cdf $F(\cdot)$.
- construct the empirical cdf function from the data (Without continuity correction)
- KS test statistic is $D = \max_x |F(x) - \hat{F}(x)|$
- reject the null hypothesis if $D > D_{n,\alpha}$
  - $\lim_{n\to\infty} \sqrt{n}D \sim$ Kolmogorov distribution
  - $D_{n,\alpha}$: confidence level of Kolmogorov distribution
  - $n =$ sample size, $\alpha =$ is the level of significance
  - values of $D_{n,\alpha}$ are tabulated

Emp CDF

hypothesized CDF

known distn

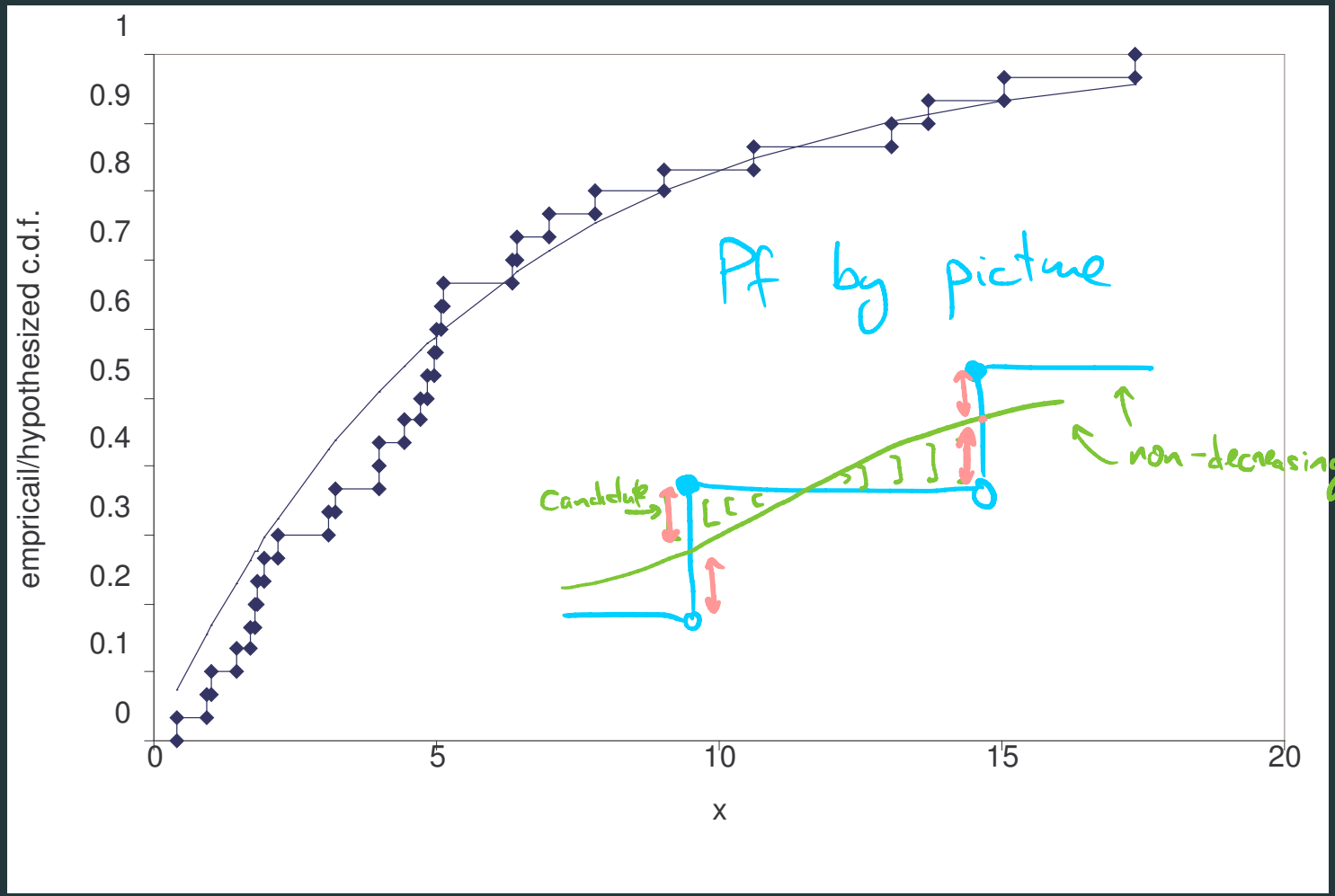
hypothesis testing

- Issue - How to compute D efficiently

Kolmogorov-Smirnov test

- $\hat{F}(\cdot)$ is a step function.
- to compute test statistic $D = \max_x |F(x) - \hat{F}(x)|$:
  enough to evaluate $|F(x) - \hat{F}(x)|$ only at the "jump" points

Can be computed in O(n)



Kolmogorov-Smirnov Test

- the test needs to be adjusted if:
  - used for a discrete rv
  - if one uses the data to estimate any parameters
- with adjustments, distribution of D depends on the particular distribution that is hypothesized
- tables of percentiles are available for many common distributions

goodness-of-fit tests: final remarks

- little data $\implies$
  all goodness of fit tests will have trouble rejecting any distribution
- enormous data $\implies$
  theoretical families of distributions may not be broad enough to accurately reflect the data
- should not have blind faith in goodness of fit tests
- software fits all distributions and ranks them based on p-values
  don't trust those rankings completely

parameter estimation error

parameter estimation error

we have now seen how to
– choose a **distribution family**
– fit **parameters**
– visualize/measure **goodness-of-fit**

even if we do everything right, $\exists$ errors in parameter estimates

- how can we estimate the magnitude of this error?
- can this error affect our simulations?

parameter estimation error: example

given $n = 200$ inter-arrival times of people arriving to a COVID testing site

- distribution guess:
- MLE estimate:
- parameter estimation error: $\hat{\lambda} - \lambda =$

parameter estimation error: example

given $n = 200$ inter-arrival times of people arriving to a COVID testing site

- distribution guess: Exponential(λ)
- MLE estimate: $\hat{\lambda} = 200/\sum_{i=1}^{2} 00A_i$
- parameter estimation error: $|\hat{\lambda} - \lambda|$

suppose there is one tester, and each test time is iid, with mean $\mu = 20s$, standard deviation $\sigma = 20s$

- service rate $= \mu = 3$ per minute
- set $\rho = \lambda/\mu = \lambda/3$
- average number of people in test center?
  Pollaczek-Khintchine formula:

$$L = \rho + \frac{\rho^2 + \lambda^2\sigma^2}{2(1-\rho)} = \frac{\lambda}{3} + \frac{\lambda^2}{9 - 3\lambda},$$

parameter estimation error: example

suppose our estimate $\hat{\lambda} =$ is accurate to ± 0.25.

- we can use PK formula to compare the expected number of cars when $\mu = 3$ and $\mu = 6$.
- Much less variability in L as μ increases

bootstrapping

- pretend we knew the **true** cdf F
- mimic the sampling process:
  generate many samples, each of size n, from F
- get an estimate of λ, $\hat{\lambda}_i$ say, from the ith sample
- plot a histogram of the $\hat{\lambda}_i$s
- **problem:** don't know F
- **solution:** replace it with a good guess

parametric bootstrap

1. given data $X_1, X_2, \ldots, X_n$ and family of distributions with one or more parameters θ
2. compute parameter estimate $\hat{\theta}_0$ using MoM/MLE
   let $\hat{F}$ be the cdf with this parameter
3. for $i = 1, 2, \ldots, m$
   - 3.1 generate sample $Y_1(i), Y_2(i), \ldots, Y_n(i)$ from $\hat{F}$
     (note: sample size same as in the original data)
   - 3.2 use same estimation procedure as in step 2 to get new estimate $\hat{\lambda}_i$ from the generated sample
4. plot a histogram of the m estimates $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_m$ to get a sense of the distribution of the estimation error in $\hat{\lambda}_0$