Sannolikhetslära och Statistik
Föreläsning 5

Mattias Villani

Avdelningen för Statistik och Maskininlärning
Institutionen för datavetenskap
Linköpings universitet

Översikt

- Stora talen lag
- Centrala gränsvärdessatsen
- Simulering

Stora talens lag

- Medelvärde: $\bar{X}_n = \frac{X_1+X_2+...+X_n}{n}$
- Medelvärden av många oberoende slumpvariabler med samma fördelning kommer att ligga allt närmare variablernas väntevärde.
- **Stora talens lag**:

$$\lim_{n\to\infty} P\left(|\bar{X}_n - \mu| > \epsilon\right) = 0$$

- Bevis via Chebyshevs olikhet

$$P\left\{|X - \mu| > \varepsilon\right\} \leq \frac{\sigma^2}{\varepsilon^2}$$

eftersom σ^2 i detta fall är $Var(\bar{X}_n) = Var(X_i)/n \to 0$ när $n \to \infty$.

Centrala gränsvärdessatsen

- Hur är summan $S_n = X_1 + X_2 + ... + X_n$ utav $n$oberoende variabler fördelad?
- Demo av
  - S_n $\quad Var(S_n) = n\sigma^2 \to \infty$
  - S_n/n $\quad Var(S_n/n) = \sigma^2/n \to 0$
  - $S_n/\sqrt{n}$ $\quad Var(S_n/\sqrt{n}) = \sigma^2$.
- **CLT: Medelvärden** av n oberoende variabler med godtycklig fördelning **blir alltmer normalfördelade när** n **ökar**.
- $n > 30$ är en vanlig tumregel.

Centrala gränsvärdessatsen

Theorem

Låt $X_1, X_2, ..., X_n$ vara oberoende variabler med väntevärde $\mu = \mathbb{E}X_i$ och standardavvikelse $\sigma = \mathrm{Std}(X_i)$ och låt

$$S_n = \sum_{i=1}^{n} X_i = X_1 + X_2 + ... + X_n.$$

När $n \to \infty$ så kommer den standardiserade summan

$$Z_n = \frac{S_n - \mathbb{E}S_n}{\mathrm{Std}(S_n)}$$

*att **konvergera i fördelning** till en $N(0,1)$ variabel, dvs*

$$F_{Z_n}(z) = P\left\{\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z\right\} \longrightarrow \Phi(z)$$

SIMULERING

- **Pseudoslumptalsgenerator**: Datorer kan generera en lång sekvens tal som ser ut som $U(0,1)$ slumptal. Good enough.
- R: `runif(1)`. Matlab: `rand`. Python: `numpy.random.uniform()`.
- Från $U \sim U(0,1)$ kan vi skapa slumptal från andra fördelningar.
- Ex. **Bernoulli** med sannolikhet p att lyckas:

$$X = \begin{cases} 1 & \text{om } U < p \\ 0 & \text{om } U \geq p \end{cases}$$

  - R kod Bernoulli: `U=runif(1); X=(U<p)`
- Ex. **Binomial**. Summan av Bernoullis
  - R-kod för Binomial(n,p): `U=runif(n); X=sum(U<p`

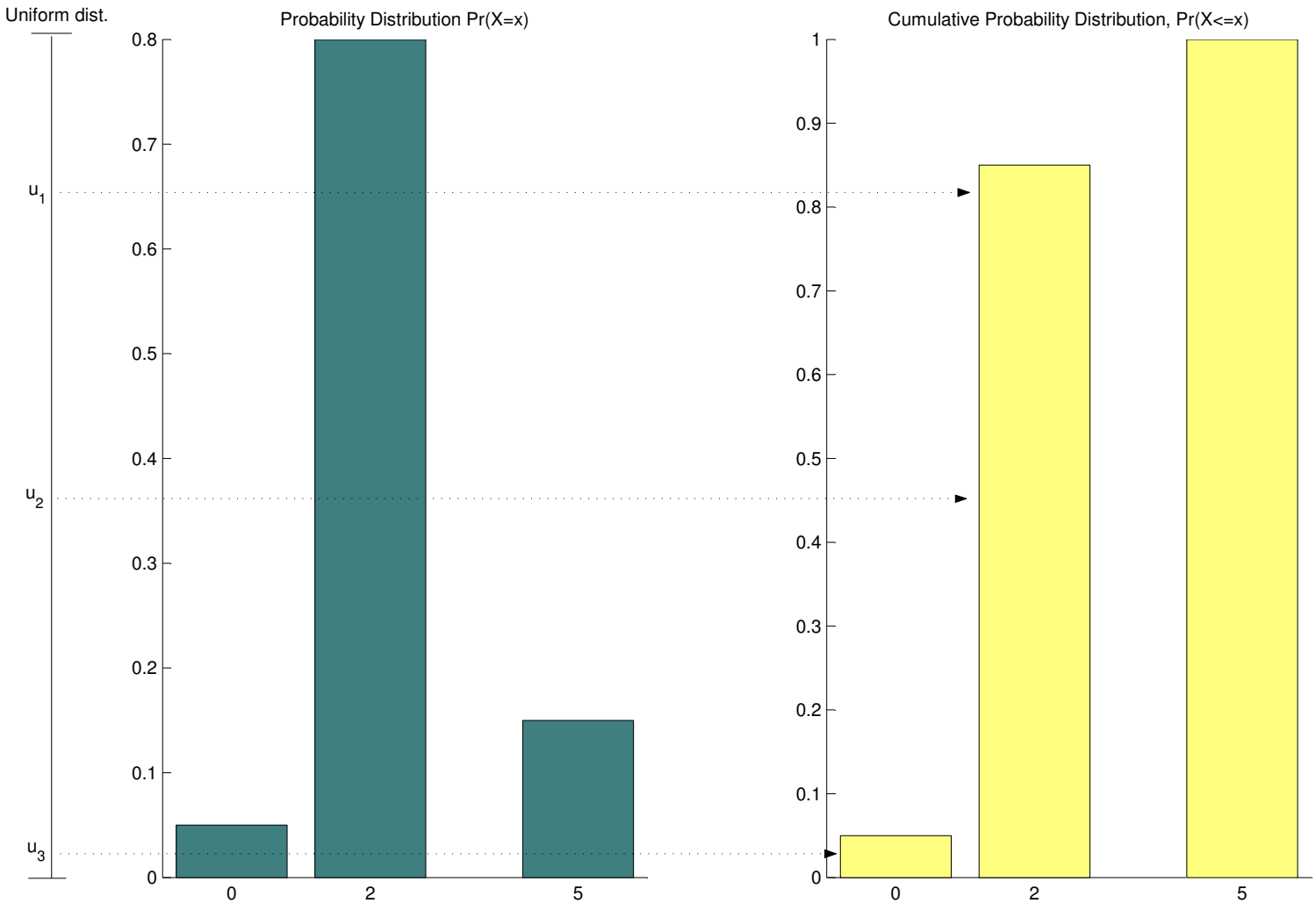
Simulering från diskret fördelning

- Simulering från allmän diskret fördelning:

$$p_i = \mathbf{P}\{X = x_i\}, \quad \sum_{i=1} p_i = 1$$

- Dela upp intervallet $[0, 1]$ i delintervall:
  - $A_1 = [0, p_1)$
  - $A_2 = [p_1, p_2)$
  - $\vdots$
  - $A_n = [p_{n-1}, 1)$
- Slumpa $U \sim U(0, 1)$
- Om $U \in A_i$ låt $X = x_i$

INVERSE CDF METHOD, DISCRETE CASE



Inversa transformationsmetoden

- Simulering från allmän **kontinuerlig** fördelning.

Theorem

Låt X vara en kontinuerlig variabel med cdf $F_X(x)$ och låt $U = F_X(X)$ vara en ny slumpvariabel. Då gäller att $U \sim U(0,1)$.

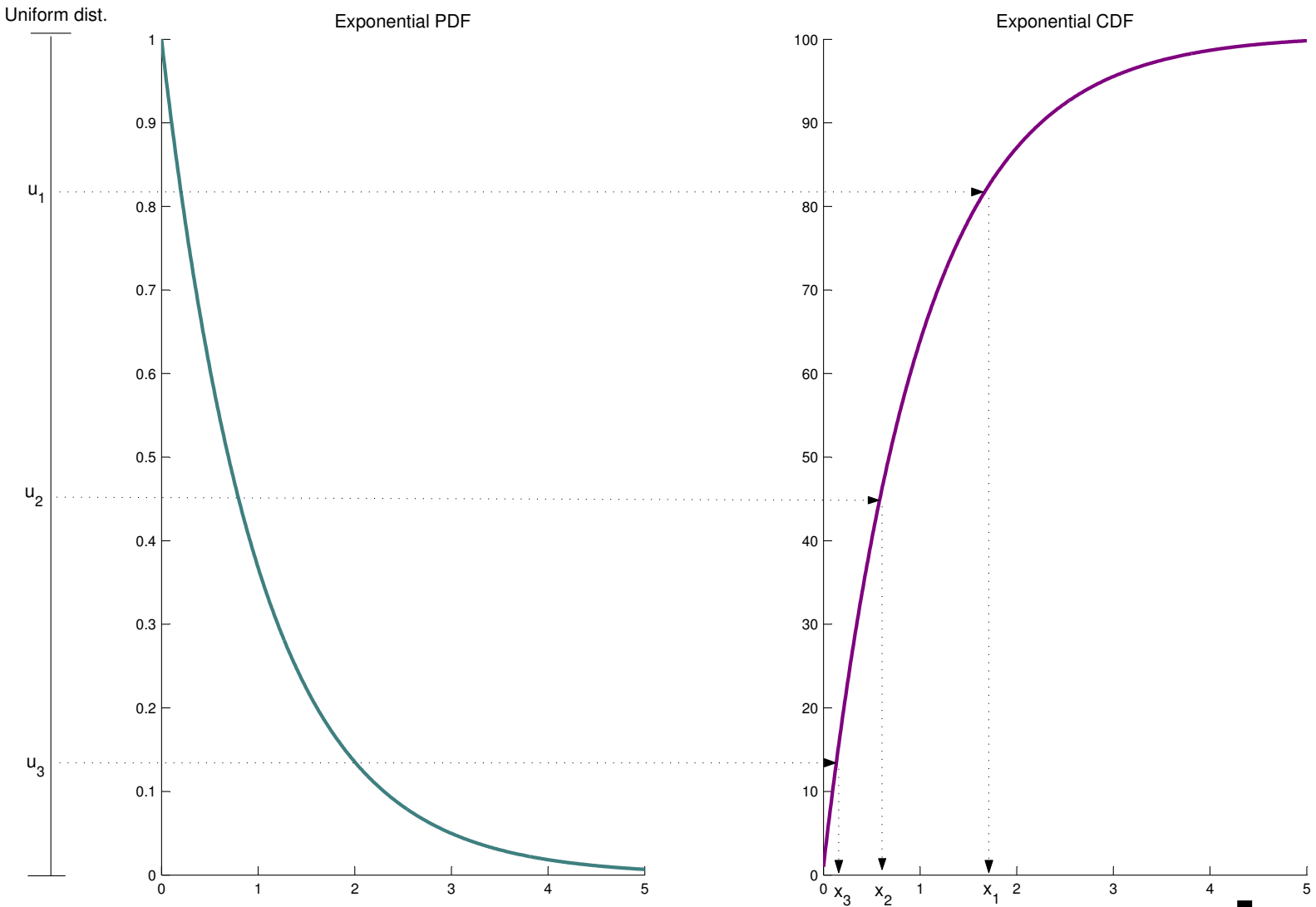
- ***Inversa transformationsmetoden**: Antag att X har cdf $F(X)$. X kan då simuleras med hjälp av en $U \sim U(0,1)$ variabel: $X = F^{-1}(U)$.*
- *Dvs lös ut X från ekvationen $U = F(X)$.*
- *Ex. $X \sim Exp(\lambda)$.*

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

vilket har lösningen

$$X = -\frac{1}{\lambda} \ln(1 - U)$$

Inverse CDF method, continuous case



Simulering i R

- n **slumptal** från $N(\mu = 2, \sigma^2 = 3^2)$ simuleras med `rnorm(n, mean = 2, sd = 3)`
- n **slumptal** från $Gamma(\alpha = 2, \lambda = 3)$ simuleras med `rgamma(n, shape = 2 , rate = 3)`

- Beräkna **pdf:en** i punkten $x = 1.5$ för $N(\mu = 2, \sigma^2 = 3^2)$: `dnorm(x=1.5, mean = 2, sd = 3)`
- Beräkna **cdf:en** i punkten $x = 1.5$ för $N(\mu = 2, \sigma^2 = 3^2)$: `pnorm(x=1.5, mean = 2, sd = 3)`