

SANNOLIKHETSLÄRA OCH STATISTIK

FÖRELÄSNING 5

Mattias Villani

**Avdelningen för Statistik och Maskininlärning
Institutionen för datavetenskap
Linköpings universitet**



ÖVERSIKT

- ▶ Stora talen lag
- ▶ Centrala gränsvärdessatsen
- ▶ Simulering
- ▶ Monte Carlo metoder

STORA TALENS LAG

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

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

- ▶ Bevis via Chebyshevs olikhet

$$P\{|X - \mu| > \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}$$

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

CENTRALA GRÄNSVÄRDESSATSEN

- ▶ Hur är summan $S_n = X_1 + X_2 + \dots + X_n$ utav n oberoende variabler fördelad?
- ▶ Demo av
 - ▶ S_n $\text{Var}(S_n) = n\sigma^2 \rightarrow \infty$
 - ▶ S_n/n $\text{Var}(S_n/n) = \sigma^2/n \rightarrow 0$
 - ▶ S_n/\sqrt{n} $\text{Var}(S_n/\sqrt{n}) = \sigma^2.$
- ▶ **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, \dots, X_n vara oberoende variabler med väntevärde $\mu = \mathbb{E}X_i$ och standardavvikelse $\sigma = \text{Std}(X_i)$ och låt

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

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

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

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

- ▶ **Pseudoslumtalsgenerator:** 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:

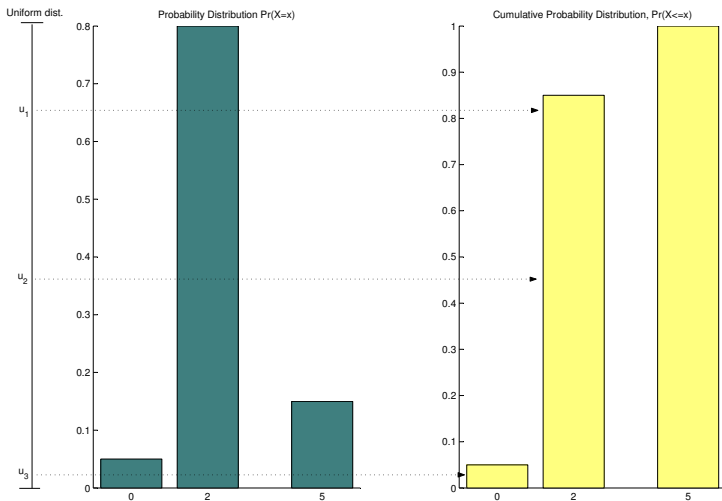
$$p_i = \mathbf{P} \{X = x_i\}, \quad \sum_{i=1} p_i = 1$$

- ▶ 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 $X = x_i$

INVERSA CDF METODEN - DISKRETA FALLET



INVERSA CDF METODEN - KONTINUERLIG

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:

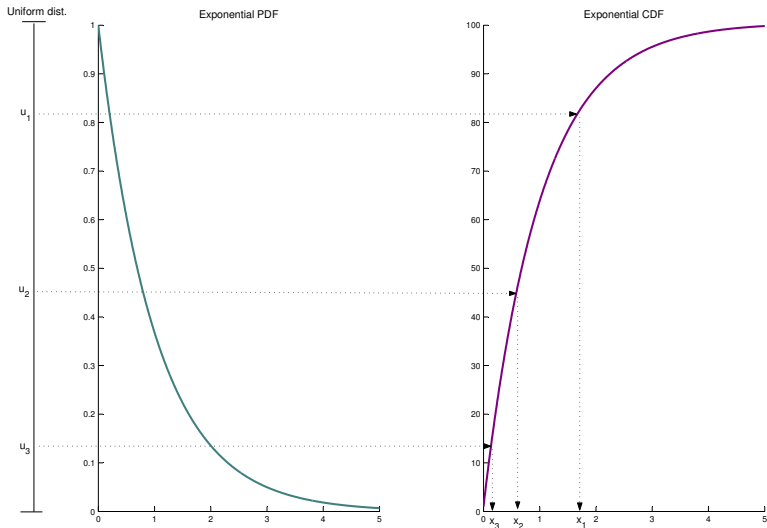
$$X = F^{-1}(U)$$

- ▶ Dvs lös ut X från ekvationen $U = F(X)$.
- ▶ Ex. $X \sim \text{Exp}(\lambda)$.

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

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

INVERSA CDF METODEN - KONTINUERLIG

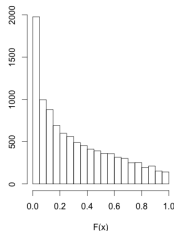
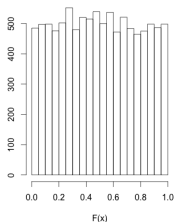


SIMULERING I R

- ▶ n slumpstal från $N(\mu = 2, \sigma^2 = 3^2)$ simuleras med
`rnorm(n, mean = 2, sd = 3)`
- ▶ n slumpstal från $\text{Gamma}(\alpha = 2, \lambda = 3)$ 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)`

TESTAR INVERSA CDF METODEN

- ▶ Följande funkar (dvs F_x blir likformigt fördelad):
 - ▶ `x = rgamma(10000, shape = 2, rate = 3)`
 - ▶ `Fx = pgamma(x, shape = 2, rate = 3)`
 - ▶ `hist(Fx,30)`



- ▶ Följande funkar inte (dvs F_x blir **inte** likformigt fördelad):
 - ▶ `x = rgamma(10000, shape = 2, rate = 3)`
 - ▶ `Fx = pgamma(x, shape = 1, rate = 3)`
 - ▶ `hist(Fx,30)`

MONTE CARLO METODER

- ▶ Simulering från fördelningar kan användas för att approximera t ex olika sannolikheter.
- ▶ Låt X_1, X_2, \dots, X_N vara oberoende dragningar från en sannolikhetsfördelning. Vi kan t ex approximera sannolikheten $p = \mathbf{P}\{X < 2\}$ med

$$\hat{p} = \hat{\mathbf{P}}\{X < 2\} = \frac{\text{antal av } X_1, X_2, \dots, X_N \text{ som är mindre än } 2}{N}$$

- ▶ $\hat{\theta}$ (t ex \hat{p}) är en **estimator** (uppskattning) av kvantiteten θ (t ex p).
- ▶

```
x = rnorm(10000, mean = 1, sd = 2);  
pHat = sum(x<2)/10000
```

MONTE CARLO METODER, FORTS.

- ▶ Men \hat{p} är bara en **skattning** av p . Varierar från stickprov till stickprov.
- ▶ Om vi upprepar hela receptet flera ggr, varje gång med ett nytt stickprov av storleken N , kommer vi då att ha rätt i genomsnitt?
Är $\mathbb{E}(\hat{p}) = p$?
- ▶ Hur mycket kommer \hat{p} att variera från stickprov till stickprov?
Hur stor är $\text{Var}(\hat{p})$?
- ▶ $Y = \text{Antal } X_1, \dots, X_N \text{ som är mindre än } 2$. $Y \sim \text{Bin}(N, p)$. Så

$$\mathbb{E}(\hat{p}) = \mathbb{E}\left(\frac{Y}{N}\right) = \frac{1}{N}N \cdot p = p$$

så \hat{p} är en **väntevärdesriktig (unbiased)** estimator av p .

$$\text{Var}(\hat{p}) = \text{Var}\left(\frac{Y}{N}\right) = \frac{1}{N^2}Np(1-p) = \frac{p(1-p)}{N}.$$

- ▶ Se Baron s. 115-116 om hur man kan välja N för att nå given exakthet $\mathbf{P}\{|\hat{p} - p| > \varepsilon\} \leq \alpha$.

Monte Carlo Integration

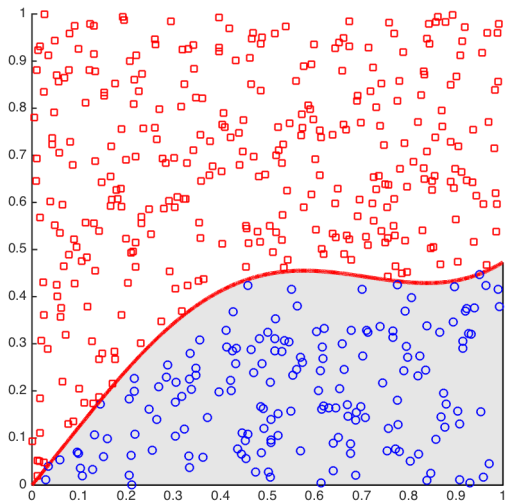
- ▶ Mål: $\mathcal{I} = \int_0^1 g(x) dx$ där $0 \leq x \leq 1$ och $0 \leq g(x) \leq 1$.
- ▶ Simulera likformigt fördelade tal U_1, \dots, U_N och V_1, \dots, V_N .
- ▶ Monte Carlo skattning

$$\hat{\mathcal{I}} = \frac{\text{Antal dragningar där } V_i < g(U_i)}{N}$$

- ▶

```
u = runif(10000);  
v = runif(10000);  
IHat = mean(v < g(u))
```

Monte Carlo Integration



IMPORTANCE SAMPLING

- Räkna **integraler** som väntevärden

$$\mathcal{I} = \int_a^b g(x) dx = \frac{1}{b-a} \int_a^b (b-a)g(x) dx = \mathbb{E} \{ (b-a)g(X) \}$$

- Stora talens lag: Om X_1, \dots, X_N dras från $U(a, b)$ så kommer

$$\frac{(b-a)g(X_1) + \dots + (b-a)g(X_N)}{N}$$

vara nära $\mathbb{E} \{ (b-a)g(X) \}$ när N är stort (konvergerar i sannolikhet).

- **Importance sampling**. Samma trick, med godtycklig pdf $f(x)$

$$\mathcal{I} = \int_a^b g(x) dx = \int_a^b \frac{g(x)}{f(x)} f(x) dx = \mathbb{E} \left(\frac{g(X)}{f(X)} \right)$$

där väntevärdet beräknas med avseende på $f(x)$.

- **Importance sampling estimatoren**: X_1, \dots, X_N oberoende från $f(X)$:

$$\hat{\mathcal{I}} = \frac{1}{N} \sum_{i=1}^N \frac{g(X_i)}{f(X_i)}$$