Sannolikhetslära och Statistik
Föreläsning 4

Mattias Villani

Avdelningen för Statistik och Maskininlärning
Institutionen för datavetenskap
Linköpings universitet

Översikt

- **Täthetsfunktion**
- **Likformig fördelning**
- **Exponentialfördelningen**
- **Gammafördelningen**
- **Normalfördelningen**

Kontinuerliga slumpvariabler

- Kontinuerliga slumpvariabler kan anta alla reela värden på ett inteval (a, b), speciellt $(-\infty, \infty)$.
- X kontinuerlig $\Rightarrow P(x) = 0$ för alla x. Pmf inte användbar.
- Fördelningsfunktionen funkar dock: $F(x) = \mathbf{P}\left\{X \leq x\right\}$.
- Eftersom $P(x) = 0$ för alla x så gäller $\mathbf{P}\left\{X \leq x\right\} = \mathbf{P}\left\{X < x\right\}$.
- Om X kontinuerlig slumpvariabel: $F(x)$ **kontinuerlig**. Inga hopp. **Icke-avtagande**.

$$\lim_{x\to\infty} F(x) = 1 \quad \lim_{x\to-\infty} F(x) = 0.$$

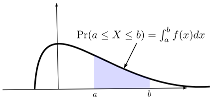
TÄTHETSFUNKTION

Definition. **Täthetsfunktionen** $f(x)$ för en kontinuerlig slumpvariabel X är derivatan av CDF:n

$$f(x) = F'(x).$$

- Fördelningen är kontinuerlig om den har en täthetsfunktion.
- Täthetsfunktion heter **probability density function**, **pdf** på engelska.
- cdf:n $F(x)$ är antiderivatan av pdf:n.
- Sannolikheter för intervall ges av ytor under pdf:n

$$\mathbf{P}\{a < X < b\} = \int_a^b f(x)dx$$



Täthetsfunktion

- $f(x) = F'(x)$ så
$$\int_{-\infty}^{b} f(x)dx = F(b) - F(-\infty) = F(b) - 0 = F(b).$$
- Täthetsfunktioner integrerar till ett:
$$\int_{-\infty}^{\infty} f(x)dx = F(\infty) - F(-\infty) = 1 - 0 = 1.$$
- Täthetsfunktionens värden, t ex $f(2)$, är inte en sannolikhet. $f(2) > 1$ helt ok. Men $f(x) \geq 0$ måste gälla.
- För litet ϵ: $\Pr\left(a - \frac{\epsilon}{2} \leq X \leq a + \frac{\epsilon}{2}\right) \approx \epsilon \cdot f(a)$.
- `Exempel`: triangelfördelningen över support $[0, a]$.
Normaliseringskonstant. Fördelningsfunktion. $P\{X > a/2\}$.
Se också Example 4.1 i Baron.
- Se Table 4.1 i Baron för en jämförelse av diskreta och kontinuerliga fördelningar.

Väntevärde och varians

- Repetition: för diskreta slumpvariabler:

$$\mathbb{E}X = \sum_x x \cdot P(x) \quad Var(X) = \mathbb{E}\left(X-\mu\right)^2 = \sum_x \left(x-\mu\right)^2 P(x)$$

- För kontinuerliga slumpvariabler:

$$\mathbb{E}X = \int x \cdot f(x) dx \qquad Var(X) = \mathbb{E}\left(X-\mu\right)^2 = \int (x-\mu)^2 f(x) dx$$

- Exempel: triangelfördelning.

Simultanfördelning för kontinuerliga variabler

- **Simultan fördelningsfunktion**

$$F_{(X,Y)}(x,y) = \mathbf{P}\left\{X \leq x \cap Y \leq y\right\}$$

- **Simultan täthetsfunktion**

$$f_{(X,Y)}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{(X,Y)}(x,y)$$

- Ofta skriver vi bara $f(x,y)$ istället för $f_{(X,Y)}(x,y)$.
- **Kovarians**

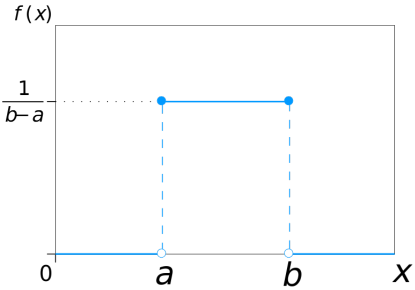
$$\begin{aligned} Cov(X,Y) &= \mathbb{E}\left(X-\mu_X\right)\left(Y-\mu_Y\right) \\ &= \int\int\left(X-\mu_X\right)\left(Y-\mu_Y\right) f(x,y)dxdy \end{aligned}$$

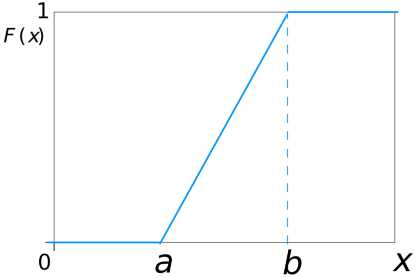
Likformig fördelning

- **Täthetsfunktion** för likformig fördelad slumpvariabel över $[a, b]$

$$f(x) = \frac{1}{b-a} \quad \text{för } a \leq x \leq b, \text{ och } f(x) = 0 \text{ annars.}$$

- Man skriver of $X \sim U(a, b)$ för att säga:
'Slumpvariabel X följer en likformig fördelning på intervallet (a, b).
Likformig = **U**niform på engelska.





Likformig fördelning

- **Väntevärde**:

$$\mathbb{E}X = \int x \cdot f(x)dx = \frac{1}{b-a}\int xdx = \frac{1}{b-a}\left[\frac{1}{2}x^2\right]_a^b$$

$$= \frac{1}{2(b-a)}\left(b^2 - a^2\right) = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}$$

- **Varians**: $Var(X) = \mathbb{E}X^2 - \mu^2$

$$\mathbb{E}X^2 = \int x^2 \cdot f(x)dx = \frac{1}{b-a}\int x^2 dx = \frac{a^2+b^2+ab}{3}$$

$$Var(X) = \mathbb{E}X^2 - \mu^2 = \frac{a^2+b^2+ab}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$$

- Alt. härledning, se Baron s. 81. Alla likformiga variabler kan genereras från **standardmedlemmen**: $Y \sim U(0,1)$ genom följande resultat:

$$X = a + (b-a)Y \text{ där } Y \sim U(0,1) \implies X \sim U(a,b).$$

Exponentialfördelningen

- **Täthetsfunktion** för exponentialfördelad slumpvariabel över $(0, \infty)$

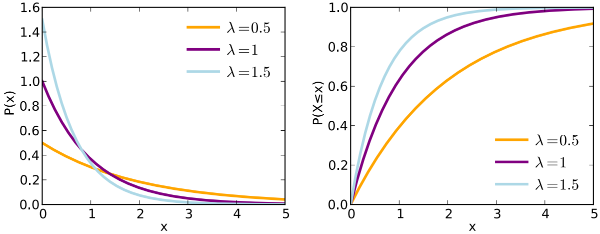
$$f(x) = \lambda e^{-\lambda x}, \text{ för } x > 0.$$

- Vi skriver: $\mathrm{X} \sim \mathrm{Exp}(\lambda)$.
- Vänteväarde

$$\mathbb{E}X = \frac{1}{\lambda}$$

- Varians

$$Var(X) = \frac{1}{\lambda^2}$$



Exponentialfördelningen

- Tiden mellan Poissonhändelser är exponentialfördelad.
- Låt $t \sim Po(\lambda t)$ räkna antalet händelser i tidsintervallet $[0, t)$.

$$
\begin{aligned}
P\left\{\text{nästa händelse innan } t\right\} &= 1 - P\left\{\text{nästa händelse efter } t\right\} \\
&= 1 - P\left\{\text{inga händelser i intervallet } [0, t)\right\} \\
&= 1 - \frac{e^{-\lambda t}(\lambda t)^0}{0!} = 1 - e^{-\lambda t}
\end{aligned}
$$

vilket är cdf:en för en $\mathrm{Exp}(\lambda)$ variabel.

- Exponentialfördelade variabler är **minneslösa**:

$$
\mathbf{P}\left\{T > t + x \middle| T > t\right\} = \mathbf{P}\left\{T > t + x\right\}
$$

Gammafördelningen

- Antag att tiden för att ladda ner en fil är $\mathrm{Exp}(\lambda)$ fördelad. Tiden för att ladda ner α filer följer en $\mathrm{Gamma}(\alpha, \lambda)$ fördelning om nedladdningstiderna är oberoende.
- Alltså: Om $X_1, X_2, ..., X_\alpha$ är α stycken **oberoende** $\mathrm{Exp}(\lambda)$ variabler:

$$Y = X_1 + X_2 + ... + X_\alpha \sim \mathrm{Gamma}(\alpha, \lambda)$$

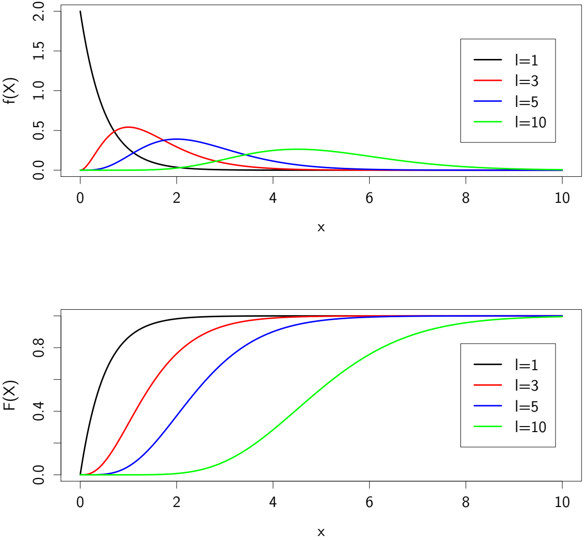
- α kallas för en **shape**parameter. λ är en **frekvens**parameter.
- Exponential är ett specialfall av Gamma: $\mathrm{Gamma}(1, \lambda) = \mathrm{Exp}(\lambda)$.
- Väntevärde

$$\mathbb{E}X = \frac{\alpha}{\lambda}$$

- Varians

$$Var(X) = \frac{\alpha}{\lambda^2}$$

Gammafördelningen



Normalfördelningen

- **Täthetsfunktion** för $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad \text{för} \quad -\infty < x < \infty$$

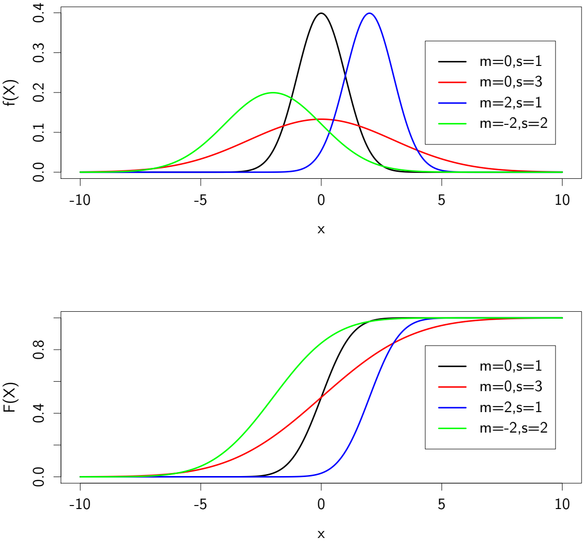
- **Väntevärde** och **varians**

$$\mathbb{E}X = \mu, \qquad Var(X) = \sigma^2$$

- **CDF** finns inte i sluten form. Om $Z \sim N(0, 1)$ så är CDFn

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

Normalfördelningen



Normalfördelningen

- **Standardmedlem**: $Z \sim N(0,1)$.

$$X = \mu + \sigma Z \text{ där } Z \sim N(0,1) \Longrightarrow X \sim N(\mu, \sigma^2).$$

- **Standardisering**

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

- $\mathbf{P}\{Z < 1.35\} = \Phi(1.35) = 0.9115$ och
  $\mathbf{P}\{Z > 1.35\} = 1 - \Phi(1.35) = 0.0885$
- Standardisering är praktiskt. Låt $X \sim N(\mu = 900, \sigma = 200)$

$$\begin{aligned}\mathbf{P}\{600 < X < 1200\} &= \mathbf{P}\left\{\frac{600-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{1200-\mu}{\sigma}\right\} \\ &= \mathbf{P}\{-1.5 < Z < 1.5\} \\ &= \Phi(1.5) - \Phi(-1.5) = 0.9332 - 0.0668 = 0.866\end{aligned}$$

Centrala gränsvärdessatsen

- Hur är summan $S_n = X_1 + X_2 + ... + X_n$ utav $n$oberoende variabler fördelad?
- Demo av
  - S_n $\quad Var(S_n) = n\sigma^2 \rightarrow \infty$
  - S_n/n $\quad Var(S_n/n) = \sigma^2/n \rightarrow 0$
  - $S_n/\sqrt{n}$ $\quad Var(S_n/\sqrt{n}) = \sigma^2.$
- **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*

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