

SANNOLIKHETSLÄRA OCH STATISTIK

FÖRELÄSNING 4

Mattias Villani

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



ÖVERSIKT

- ▶ Täthetsfunktion
- ▶ Likformig fördelning
- ▶ Exponentialfördelningen
- ▶ Gammafördelningen
- ▶ Normalfördelningen
- ▶ Betafördelningen
- ▶ t-fördelningen

KONTINUERLIGA SLUMPVARIABLER

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

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

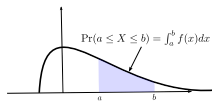
TÄTHETSFUNKTION

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

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



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: triangel fördelningen 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:

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

- ▶ För kontinuerliga slumpvariabler:

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

- ▶ Exempel: triangel fördelning.

SIMULTANFÖRDELNING FÖR KONTINUERLIGA VARIABLER

- ▶ **Simultan fördelningsfunktion**

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

- ▶ **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**

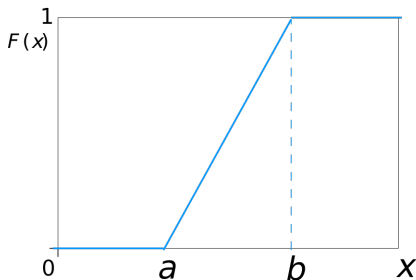
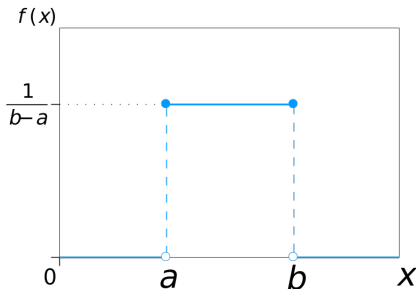
$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E} (X - \mu_X) (Y - \mu_Y) \\ &= \int \int (X - \mu_X) (Y - \mu_Y) f(x,y) dx dy \end{aligned}$$

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:

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

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

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

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

- 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 $(0, \infty)$

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

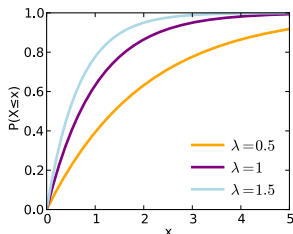
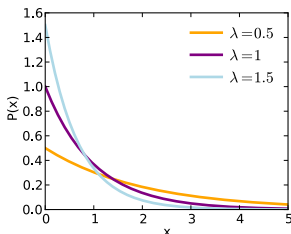
- Vi skriver: $X \sim \text{Exp}(\lambda)$.

- Väntevärde

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

- Varians

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



EXPONENTIALFÖRDELNINGEN

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

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

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

- ▶ Exponentialfördelade variabler är **minneslösa**:

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

GAMMAFÖRDELNINGEN

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

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

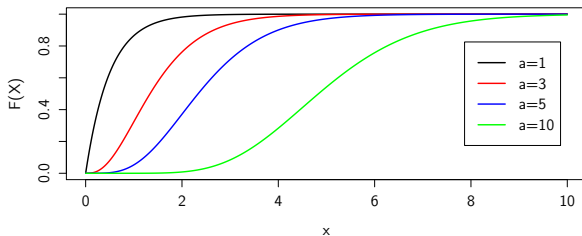
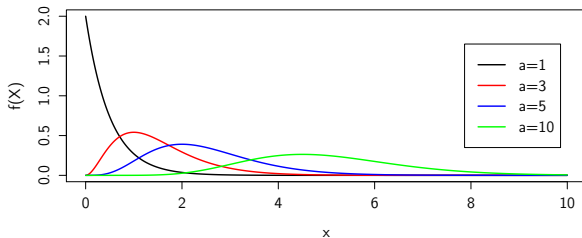
- ▶ α kallas för en **shape**parameter. λ är en **frekvens**parameter.
- ▶ Exponential är ett specialfall av Gamma: $\text{Gamma}(1, \lambda) = \text{Exp}(\lambda)$.
- ▶ Väntevärde

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

- ▶ Varians

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

GAMMAFÖRDELNINGEN



NORMALFÖRDELNINGEN

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

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

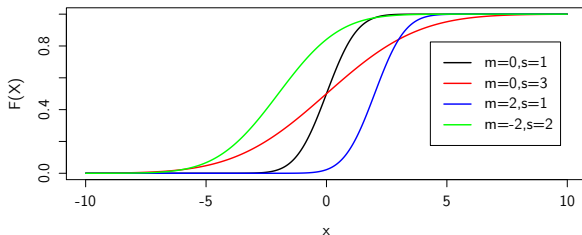
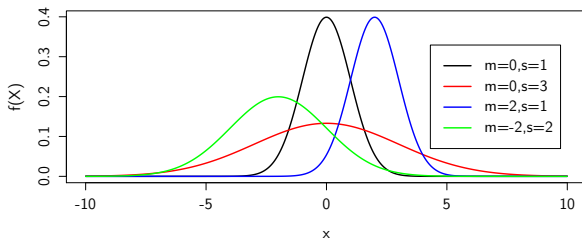
- **Väntevärde** och **varians**

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

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

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

- **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 $X \sim N(\mu = 900, \sigma = 200)$

$$\begin{aligned}\mathbf{P}\{600 < X < 1200\} &= \mathbf{P}\left\{\frac{600 - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{1200 - \mu}{\sigma}\right\} \\ &= \mathbf{P}\{-1.5 < Z < 1.5\} \\ &= \Phi(1.5) - \Phi(-1.5) = 0.9332 - 0.0668 = 0.8664\end{aligned}$$

BETAFÖRDELNINGEN

- ▶ Passar **kontinuerliga** variabler i intervallet $[0, 1]$, t ex andelar.
- ▶ Täthet för $X \sim \text{Beta}(\alpha, \beta)$

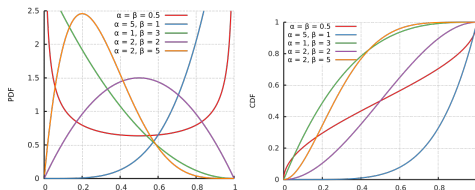
$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad \text{for } 0 \leq x \leq 1$$

- ▶ **Betafunktionen**

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

- ▶ **Väntevärde och varians**

$$\mathbb{E}X = \frac{\alpha}{\alpha + \beta} \quad \text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$



t -FÖRDELNINGEN

- ▶ Normalfördelningen har **tunna svansar**. Mycket osannolikt att observera extrema observationer.
- ▶ t -fördelningen är en generalisering av normalfördelningen med en parameter ν (**frihetsgrader**) som modellerar hur tunga svansarna är.
- ▶ En **t -fördelad** variabel $X \sim t_\nu(\mu, \sigma^2)$ har täthetsfunktionen

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \sqrt{\pi\nu\sigma^2}} \left(1 + \frac{1}{\nu} \left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}$$

- ▶ **Väntevärde och varians**

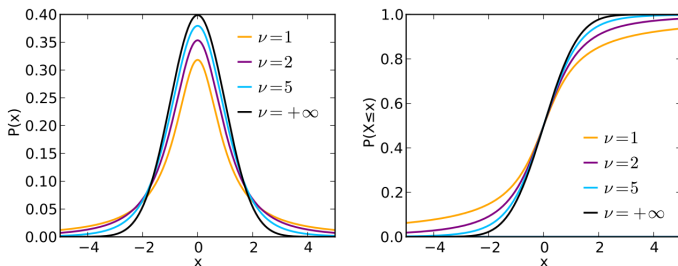
$$\mathbb{E}X = \mu \quad \text{om } \nu > 1$$

$$\text{Var}(X) = \sigma^2 \frac{\nu}{\nu-2} \quad \text{om } \nu > 2$$

- ▶ $t_\nu(0, 1)$ är standardmedlemmen.

t -FÖRDELNINGEN

- ▶ **Cauchy-fördelningen** är specialfallet med $\nu = 1$.
- ▶ Normalfördelningen fås när $\nu \rightarrow \infty$.



- ▶ Viktig koppling mellan t -fördelning och normalfördelning:
 - ▶ $X_1, \dots, X_n | \mu, \sigma^2 \stackrel{iid}{\sim} N(\mu, \sigma^2)$. σ^2 **känd**. $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$
 - ▶ $X_1, \dots, X_n | \mu, \sigma^2 \stackrel{iid}{\sim} N(\mu, \sigma^2)$. σ^2 **okänd**, skattas med s^2 .
 $T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}(0, 1)$.