# Lecture 2 – Statistics for Data Science (Dennis Fok, Erasmus University Rotterdam)

## 1. Distributies en basisbegrippen

Een distributie beschrijft de waarschijnlijkheid van uitkomsten van een willekeurige variabele.  
• Discrete variabelen: aannemen van specifieke waarden (zoals dobbelsteen).  
 - Probability mass function (pmf): f(x) = P(X = x)  
 - Cumulative distribution function (cdf): F(x) = P(X ≤ x)  
• Continue variabelen: waarden in een interval.  
 - Probability density function (pdf): f(x) zodat P(a < X ≤ b) = ∫\_a^b f(x)dx  
 - f(x) = F'(x)  
• Bij continue variabelen is P(X=x)=0; enkel de oppervlakte onder f(x) heeft betekenis.

## 2. Grafische illustraties

Discrete voorbeeld: dobbelsteen (kans 1/6 per uitkomst).  
Continue voorbeeld: f(x) = (2/9)(x−1) voor 1 ≤ x ≤ 4.  
Controle: ∫\_1^4 f(x)dx = 1.

## 3. Python implementatie

Gebruik scipy.stats voor standaardverdelingen:  
from scipy import stats  
stats.norm.cdf(1, loc=0, scale=1) # standaardnormaal  
stats.binom.pmf(15, n=30, p=0.5) # binomiale kans  
stats.expon.ppf(0.95, scale=10) # 95e percentiel exponentieel

## 4. Belangrijke standaardverdelingen

Vier centrale verdelingen:  
1. Normaal: X ~ N(μ, σ²), f(x)=1/(σ√(2π)) exp(−½((x−μ)/σ)²)  
2. t-verdeling (tₙ)  
3. Chi-kwadraat (χ²ₖ)  
4. F-verdeling (F(d₁,d₂))

## 5. Schatting van distributies

Histogram = benadering van echte verdeling.  
Gebruik scipy.stats.ecdf(data) en fit():  
fit = stats.fit(stats.norm, data, bounds={'loc':(-4,4),'scale':(0,1)})  
fit.plot(plottype='qq')

## 6. Schattingsonzekerheid

We schatten μ en σ² uit steekproeven.  
Var(X̄) = σ²/n → onzekerheid daalt bij grotere n.  
Standaardfout: SE(X̄) = √(s²/n).

## 7. Betrouwbaarheidsintervallen

Als σ² bekend: X̄ ± 1.96√(σ²/n)  
Als σ² onbekend: X̄ ± t₀.₉₇₅,ₙ₋₁ √(s²/n)  
Voorbeeld (huisprijsdata): [4967.998, 5332.533].

## 8. Central Limit Theorem (CLT)

Voor grote n geldt: √n( X̄ − μ ) → N(0, σ²)  
Zelfs bij niet-normale data benadert X̄ een normale verdeling.

## 9. Samenvattend overzicht

pmf = P(X=x)  
pdf = f(x)  
cdf = ∫ f(x)dx  
SE = √(s²/n)  
CI(σ bekend) = X̄ ± 1.96σ/√n  
CI(σ onbekend)= X̄ ± t₀.₉₇₅,ₙ₋₁ s/√n  
CLT: X̄ ~ N(μ, σ²/n) bij grote n.