TMS136

Föreläsning 4

Kontinuerliga stokastiska variabler

- Kontinuerliga stokastiska variabler är stokastiska variabler som tar värden i intervall av den reella axeln

- Det kan handla om längder, temperaturer, vikter, strömstyrkor osv…

Enkelt exempel

- Vi går, utan att kolla tidtabeller eller klockor, ut till Chalmershållplatsen för att ta vagnen eller bussen ner till stan

- Väntetiden tills ett lämpligt transportmedel kommer kan modelleras med en kontinuerlig stokastisk variabel

Allmänt

- För en kontinuerlig s.v. X finns en *täthetsfunktion (probability density function)* f sådan att

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $P(a < X \leq b) = \int_{a}^{b} f(x)dx$

Observation

- $P(S) = \int_{-\infty}^{\infty} f(x)dx = 1$

- $P(a < X < b) = P(a \leq X < b) =$

$$P(a < X \leq b) = P(a \leq X \leq b)$$

eftersom

$P(X = b) = P(X = a) = \int_{b}^{b} f(x)dx = \int_{a}^{a} f(x)dx = 0$

Fördelningsfunktionen för kontinuerlig s.v.

- För en kontinuerlig s.v. X ges fördelningsfunktionen av

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

- ”Arean under kurvan från lååååångt till vänster upp till x”
- Observera att $F'(x) = f(x)$

Egenskaper hos fördelningsfunktionen

- Om X är kontinuerlig gäller att

$$0 \leq F(x) \leq 1$$

och

$$\text{om } x < y \text{ så är } F(x) \leq F(y)$$

Observation

- Om X är kontinuerlig gäller att

$$P(X > x) = P(X \geq x) = 1 - P(X < x)$$
$$= 1 - P(X \leq x) = 1 - F(x)$$

- Vi har även att

$$P(a < X \leq b) = F(b) - F(a)$$

Väntevärdet för en kontinuerlig s.v.

- Väntevärdet för en kontinuerlig s.v. X ges av

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

- Om g är någon funktion har vi att

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) dx$$

Tolkning

- Vi tänker oss att vi kan generera slumptal från vår (kontinuerliga) s.v.
- Efter varje nytt genererat slumptal beräknar vi medelvärdet av de hittills genererade talen
- Detta medelvärde kommer konvergera mot $E(X)$ då antalet genererade slumptal växer

Variansen för en kontinuerlig s.v.

- Variansen för en kontinuerlig s.v. X med väntevärde μ ges av (verifiera sista likheten)

$$V(X) = E(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx =$$

$$= \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$$

Exempel

- En viss s.v. X kan ta värden i intervallet $(0, \pi)$ och har täthetsfunktionen

$$f(x) = Csin(x), \qquad 0 < x < \pi$$

- Vad är konstanten C?

- Det ska alltså gälla att

$$1 = \int_{-\infty}^{\infty} f(x)dx = C \int_{0}^{\pi} \sin(x)\, dx = 2C$$

- Alltså måste vi ha $C = \frac{1}{2}$

Exempel fortsättning

- Vad är $E(X)$?
- Vi har att

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \frac{1}{2}\int_{0}^{\pi} xsin(x)dx = \frac{\pi}{2}$$

- Detta hade man kunnat gissa eftersom täthetsfunktionen är symmetrisk runt $\frac{\pi}{2}$ …

Kontinuerlig likformig fördelning

- Om vi vet att vagnen till stan går var tionde minut och vi går ut till hållplatsen på måfå utan att kolla klockan är väntetiden tills vagnen kommer likformigt fördelad på intervallet $[0,10)$

- Lite mer precist är sannolikheten att vi väntar mellan två och tre minuter lika stor som att vi väntar mellan sju och åtta minuter.

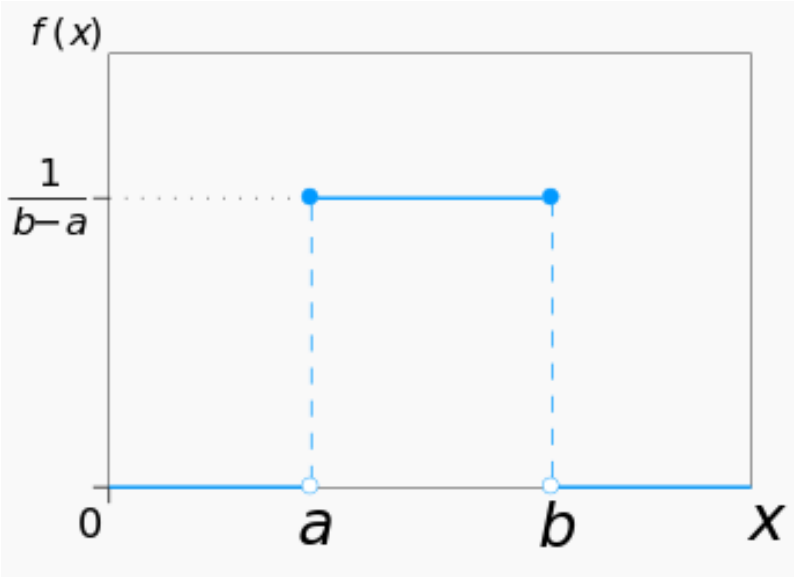
- Observera att när vi tänker oss väntetiden som en kontinuerlig s.v. blir det med nödvändighet så att sannolikheten att vi väntar prick åtta minuter (eller prick vilken tid som helst mellan noll och tio) är noll!

Kontinuerlig likformig fördelning

- Generellt säger vi att X är *kontinuerligt likformigt fördelad* på intervallet (a, b) om täthetsfunktionen för X är

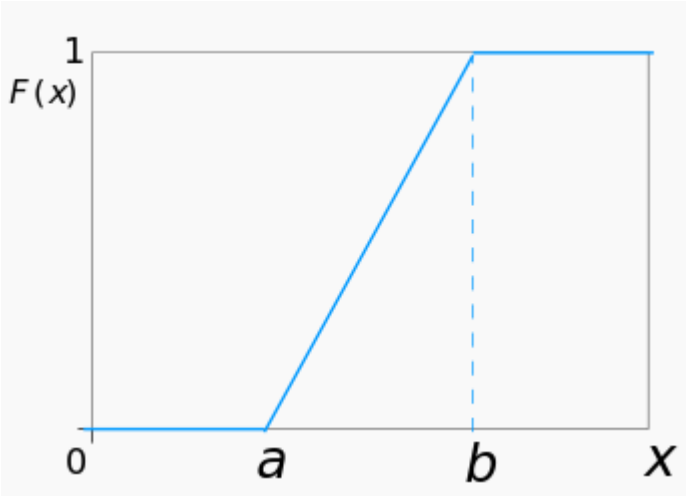
$$f(x) = \frac{1}{b-a}, a < x < b$$

$$f(x) = 0, \quad annars$$



Fördelningsfunktionen

- Vi ser att för X kontinuerligt likformigt fördelad på intervallet (a, b) så har vi att
- $F(x) = 0$ om $x \leq a$
- $F(x) = \frac{1}{b-a}\int_a^x dy = \frac{x-a}{b-a}$ om $a < x < b$
- $F(x) = 1$ om $x \geq b$



Väntevärde och varians

- Om X är kontinuerligt likformigt fördelad på (a, b) gäller att

$$E(X) = \frac{b + a}{2}$$

och

$$V(X) = \frac{(b - a)^2}{12}$$

Exempel forts

- Den förväntade tiden vi väntar tills vagnen kommer om vi antar att det är tio minuter mellan avgångarna (och inte vet vad klockan är då vi ställer oss på hållplatsen) är alltså fem minuter

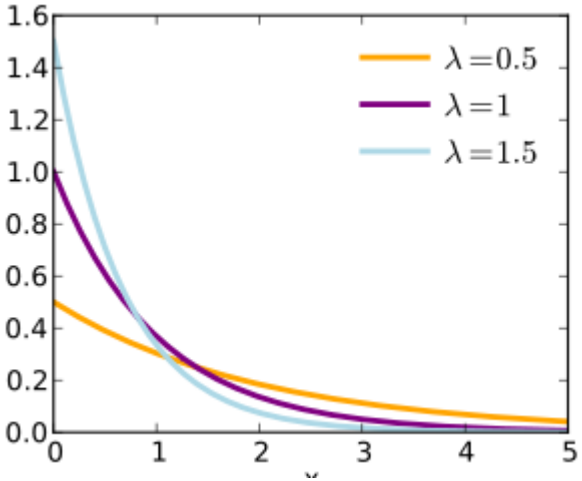
- Alltså, om vi skulle upprepa ”experimentet” många gånger skulle medelväntetiden närma sig fem minuter

Exponentialfördelning

- En annan fördelning som är viktig då man talar om väntetider eller livslängder är *exponentialfördelningen*

- Vi säger att X är exponentialfördelad med parameter $\lambda > 0$ om täthetsfunktionen för X ges av

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

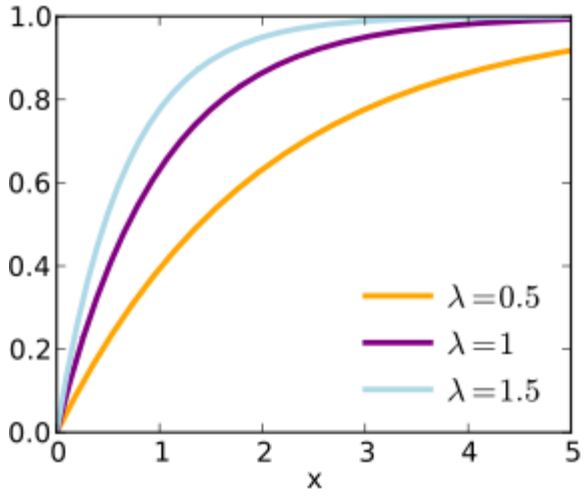


Fördelningsfunktionen

- Om X är exponentialfördelad med parameter λ ges fördelningsfunktionen av

- $F(x) = 0$ om $x \leq 0$

- $F(x) = 1 - e^{-\lambda x}$ om $x > 0$



Exponentialfördelade s.v. är dementa…

- Om man modellerar livslängden för exempelvis en elektronikkomponent med en exponentialfördelad s.v. får man en speciell egenskap på köpet…

- Man kanske undrar ”vad är slh att komponenten lever i ytterligare y tidsenheter om den redan levt i x tidsenheter?”

Exponetialfördelade s.v. är dementa…

- Observervera att om vi låter X vara den exponentialfördelade livslängden så har vi att

$$P(X > x + y | X > x) = \frac{P(X > x + y, X > x)}{P(X > x)}$$

$$= \frac{P(X > x + y)}{P(X > x)} = \frac{e^{-\lambda(x+y)}}{e^{-\lambda x}} = e^{-\lambda y} = P(X > y)$$

Exponentialfördelade s.v. är dementa…

- Om vi nu ska modellera en livslängd med exponentialfördelning finns alltså inget ”åldrande” i modellen

- Man säger att exponentialfördelningen är *minneslös*

Väntevärde och varians

- Om X är exponentialfördelad med parameter λ har vi att (verifiera detta)

$$E(X) = \frac{1}{\lambda} \quad \text{och} \quad V(X) = \frac{1}{\lambda^2}$$

Kopplingen mellan exponential och Poisson

- Kom ihåg vi säger att (den diskreta) s.v.:n X är Poissonfördelad med parameter $\lambda > 0$ om massfunktionen ges av

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \ , x = 0,1,2 \ldots$$

- Poissonfördelningen används ofta till att räkna antalet händelser i ett visst tidsintervall, exempelvis antal utstrålade alfapartiklar från ett radioaktivt material eller antalet kunder som kommer till en butik…

Kopplingen mellan exponential och Poisson

- Om man antar att tiderna mellan händelserna är oberoende och exponentialfördelade med paramater λ kommer antalet händelser att vara Poissonfördelade med samma parameter λ

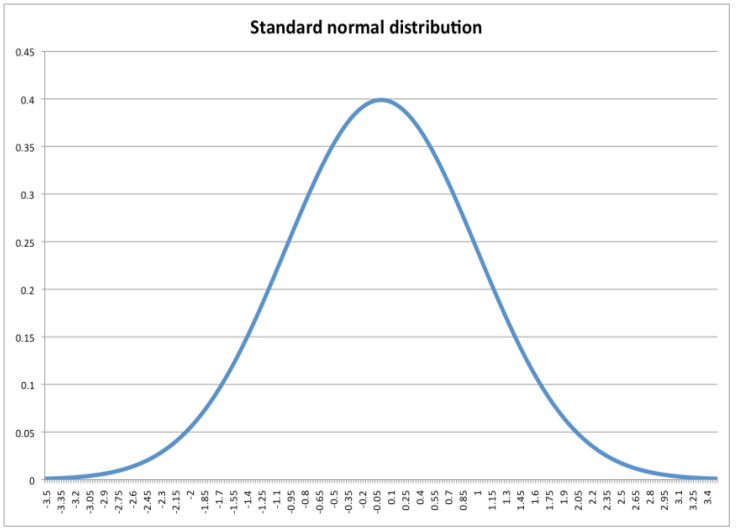
- Om vi vet att det i medelväntetiden mellan att två kunder kommer till en affär är en halv minut och antar att tiderna mellan kundernas ankomster är oberoende och exponentialfördelade betyder detta att sannolikheten att det kommer tre kunder under en minut är

$$\frac{e^{-2}2^3}{3!} \approx 0.18$$

Världens viktigaste fördelning

- Den mest förekommande fördelningen i världen är *normalfördelningen*
- Speciellt viktig är den fördelning vi kallar *standard normal*
- Vi säger att X är standard normalfördelad om täthetsfunktionen för X ges av

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$



Världens viktigaste fördelning

- Att man säger standard normal har att göra med att väntevärdet är noll och standardavvikelsen är ett

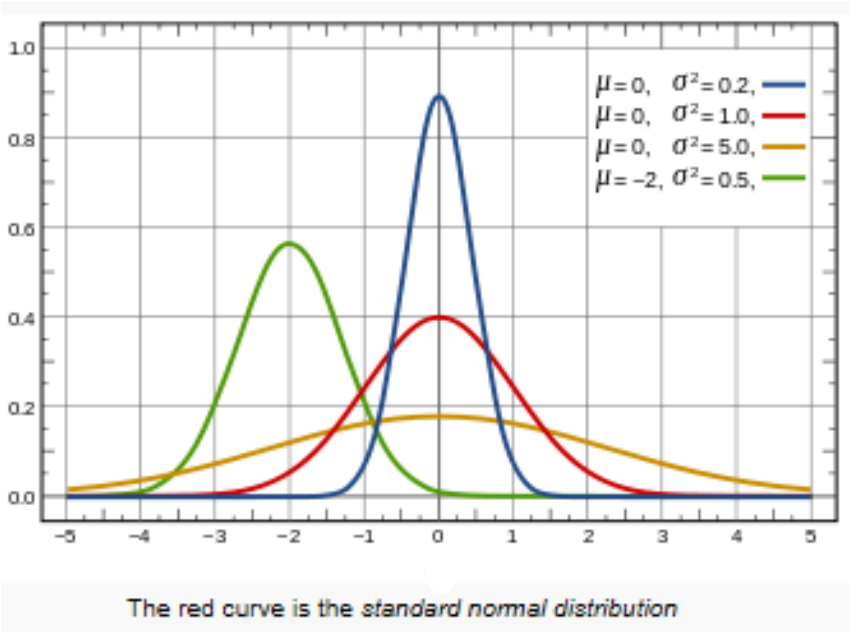
- Utifrån en standard normalfördelad s.v. X kan man skapa en normalfördelad s.v. Y med väntevärde μ och standardavvikelse σ genom att låta

$$Y = \mu + \sigma X$$

Världens viktigaste fördelning

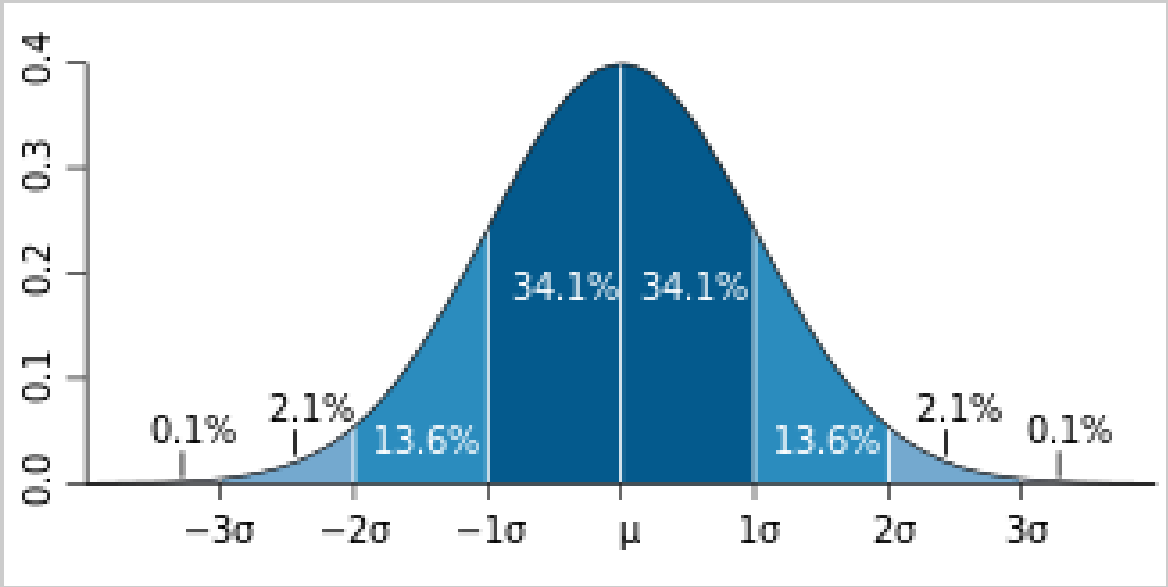
- Om Y är normalfördelad med väntevärde μ och standardavvikelse σ ges tätheten av

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$



The red curve is the *standard normal distribution*

Världens viktigaste fördelning



Dark blue is less than one standard deviation away from the mean. For the normal distribution, this accounts for about 68% of the set, while two standard deviations from the mean (medium and dark blue) account for about 95%, and three standard deviations (light, medium, and dark blue) account for about 99.7%.

Världens viktigaste fördelning

- Så om Y är normalfördelad med väntevärde μ och standardavvikelse σ gäller att $\frac{Y-\mu}{\sigma}$ är standard normal.

- Detta är viktigt att förstå eftersom om man vill finna sannolikheter för normalfördelade s.v. måste man ibland transformera till standard normal (man kan inte skriva upp fördelningsfunktionen och måste leta i tabeller eller använda numerik)

Världens viktigaste fördelning

- Eftersom standard normal förekommer så ofta låter man i många böcker bokstaven Z vara reserverad för s.v. med standard normalfördelning

- Även fördelningsfunktionen reserveras ofta en egen symbol

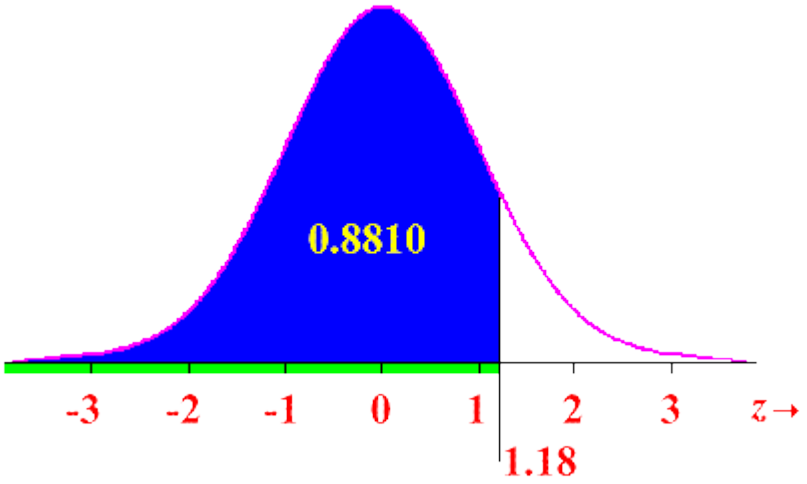
$$\Phi(z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}} \, dx$$

- Värden på Φ för olika z brukar vara tabellerade i statistikböckers appendix

Världens viktigaste fördelning

- Sannolikheten att $Z \leq z$ är arean under ”klockan” från ”lååååångt till vänster” till z

- Till exempel har vi $\Phi(1.18) = \mathrm{P}(Z \leq 1.18) = 0.8810$



Symmetrier

- Eftersom ”klockan” är symmetrisk följer att

$$\Phi(-z) = 1 - \Phi(z)$$

- Verifiera detta genom att rita några ”fyllda klockor” för olika värden på z

Och nu undrar ni varför den är så viktig

- Normalfördelningen är viktig dels för att många statistiska metoder bygger på att man har normalfördelade data

- Men det mest remarkabla av alla resultat i sannolikhetsteorin har också med normalfördelningen att göra

- Det är nämligen så att om man har ett stickprov från vilken som helst fördelning (som har väldefinierat väntevärde och varians) så kommer stickprovsmedelvärdets fördelning konvergera mot normalfördelning då stickprovsstorleken växer. Mer om detta senare…

Exempel

- Låt X vara normal $\mu = 2$ och $\sigma = 0.5$. Vad är sannolikheten att $X \geq 1$?

- Lösning (verifiera likheterna)

$$P(X \geq 1) = P\left(\frac{X-2}{0.5} \geq \frac{1-2}{0.5}\right) = P(Z \geq -2)$$

$$= P(Z \leq 2) = \Phi(2) = 0.977$$

Approximationer

- Som en följd av resultatet som ”hintades” om ovan har man att både binomialfördelning och Poissonfördelning kan approximeras med normalfördelning

- Man kan alltså approximera (vissa) diskreta fördelningar med en kontinuerlig (!)

Approximation av binomial med normal

- Om X är binomialfördelad med parametrar n och p gäller att

$$\frac{X - np}{\sqrt{np(1-p)}}$$

är approximativt standard normalfördelad.

- Approximationen blir bättre ju större n är men kan anses duglig om $np > 5$ och $n(1-p) > 5$

Kontinuitetskorrigering

- Man kan bättre på normalapproximationen av binomial genom *kontinuitetskorrigering*

$$P(X \leq x) = P(X \leq x + 0.5) \approx P\left(Z \leq \frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

$$P(X \geq x) = P(X \geq x - 0.5) \approx P\left(Z \geq \frac{x - 0.5 - np}{\sqrt{np(1-p)}}\right)$$

Approximation av Poisson med normal

- Om X är Poissonfördelad med parameter λ gäller att

$$\frac{X-\lambda}{\sqrt{\lambda}}$$

är approximativt standard normalfördelad.

- Approximationen blir bättre ju större λ är men kan anses duglig då $\lambda > 5$

- Precis som för normalapproximation av binomial kan man även här göra (på precis samma sätt) kontinuitetskorrigering för att få en bättre approximation