Matematisk statistik V
Föreläsningsanteckningar
Tommy Norberg
10 mars 2004

F1-F2: Ch 1 Data and Distributions

Nyckelord: Diskreta och kontinuerliga variabler, histogram, diskreta och kontinuerliga fördelningar, percentil, kvantil, median, exponentialfördelningen, normalfördelningen, lognormalfördelningen, Binomialfördelningen, Poissonfördelningen.

Variabler (s 3-4)

Definition 1 (s 3-4) En *variabel* är något vi mäter eller observerar i ett (eller flera) försök. Devore & Farnum använder små bokstäver, t ex x, y, z, för att beteckna olika variabler.

Exempel 1 x skulle kunna vara en nyexaminerad civilingenjörs kön, y antalet garantireparationer för ett visst bilmärke, z hållfastheten hos en viss typ av vajrar, t en konstruktions livstid, etc.

Exempel 2 (s 4) En bilförsäljare säljer bilar med manuell (M) eller automatisk (A) växellåda. De 10 senast sålda bilarna var av typ:

$$M,\ \ A,\ \ A,\ \ A,\ \ M,\ \ A,\ \ A,\ \ M,\ \ A,\ \ A$$

Detta är ett exempel på en *kategorisk* datamängd. Variabeln som observeras kan bara anta de två kategoriska värdena M och A.

Exempel 3 (s 4) Man har mätt upp livstiden på 8 batterier av en viss typ och fått:

$$5.6,\ \ 5.1,\ \ 6.2,\ \ 6.0,\ \ 5.8,\ \ 6.5,\ \ 5.8,\ \ 5.5$$

Detta är ett exempel på en *numerisk* datamängd. Variabeln som studeras kan i princip anta alla icke-negativa reella värden.

I båda exemplena ovan är datamängderna *univariata*.

Bivariata datamängder erhålls när man samtidigt mäter eller observerar två variabler.

Multivariata datamängder erhålls när man samtidigt mäter eller observerar tre eller flera variabler.

Exempel 4 Man mäter längd x (i cm) och vikt y (i kg) på spelarna i sektionens innebandylag:

$$(165, 73), \quad (182, 79), \quad \ldots, \quad (172, 75)$$

Detta är en *bivariat* datamängd. Antag att man även observerar spelarens kön. Då skulle det kunna se ut så här:

$$(K, 165, 73), \quad (M, 182, 79), \quad \ldots, \quad (M, 172, 75)$$

Detta är en *multivariat* datamängd, som är blandat kategorisk och numerisk.

Definition 2 (s 11) En variabel är *kategorisk* om dess värdemängd är icke-reell och ändlig. Den är *diskret* om värdeförrådet är en ändlig eller uppräkneligt oändlig mängd av reella tal. Den är *kontinuerlig* om den kan anta alla värden i ett eller flera intervall av reella tal.

I den här kursen ska vi huvudsakligen syssla med reellvärda (numeriska) variabler. Teorin för diskreta och kontinuerliga variabler skiljer sig åt, så vi måste hela tiden hålla i minnet om en viss variabel är diskret eller kontinuerlig.

Diskreta variabler uppkommer nästan alltid i försök där något räknas.

Exempel 5a Under juni månad år 2003 kommer trestegshopparen NN att deltaga i 7 internationella tävlingar. Antalet gånger NN hoppar minst 17.00 meter är en diskret variabel, eftersom de möjliga utfallen är $0, 1, \ldots, 7$.

Exempel 5b Antalet cykelolyckor utmed Vasagatan en viss månad är en diskret variabel, eftersom dess möjliga utfall är $0, 1, 2, \ldots$.

Mätresultat är ofta kontinuerliga.

Exempel 6a En vätskas pH-värde är en kontinuerlig variabel, eftersom pH-värdet i princip kan vara vilket tal som helst i intervallet $(0, 14)$.

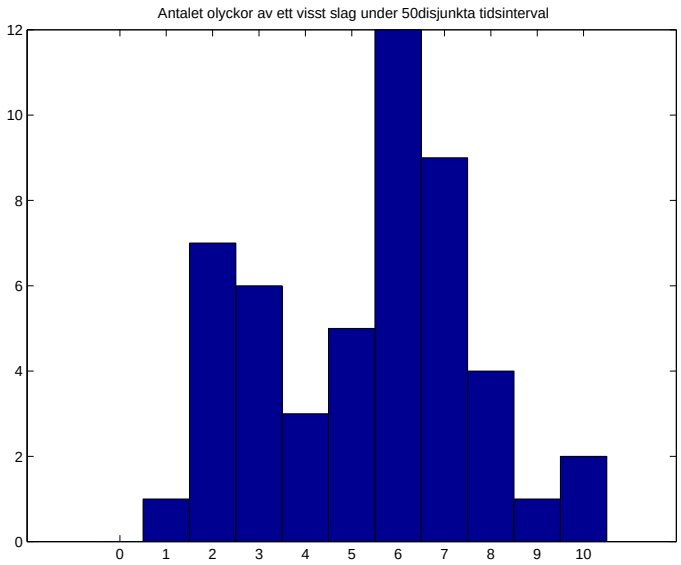
Exempel 6b Koncentrationen (i mg/m^3) av en viss tungmetall i ett jordprov är en kontinuerlig variabel, eftersom utfallsrummet i princip är hela $R_+ = [0, \infty)$.

Exempel 6c Differensen i blodets kolesterolhalt före och efter en viss behandling är en kontinuerlig stokastisk variabel. Det är naturligt att som utfallsrum för detta experiment välja hela $R = (-\infty, \infty)$, trots att det ju finns en naturligt undre gräns för differensen (kolesterolhalten kan ju inte sjunka längre än till noll).

Histogram (s 11-19)

Definition 3 (s 12) Histogram för kategoriska (eller diskreta) data. Låt kategorierna vara $a_1, \ldots, a_k$. Antag att du har f_1 observationer av kategori a_1, f_2 observationer av kategori a_2, etc, och att totala antalet observationer är n. Obs att $n = f_1 + f_2 + \ldots + f_k$. Ett *histogram* är ett stapeldiagram där höjden av stapeln ovanför a_1 är proportionell mot f_1/n, höjden av stapeln ovanför a_2 är proportionell mot f_2/n, etc.

I histogrammet nedan (producerat m.h.a. Matlab) är de respektive höjderna lika med f_1, f_2, etc.



Figur 1: Histogram över antalet olyckor av ett viss slag.

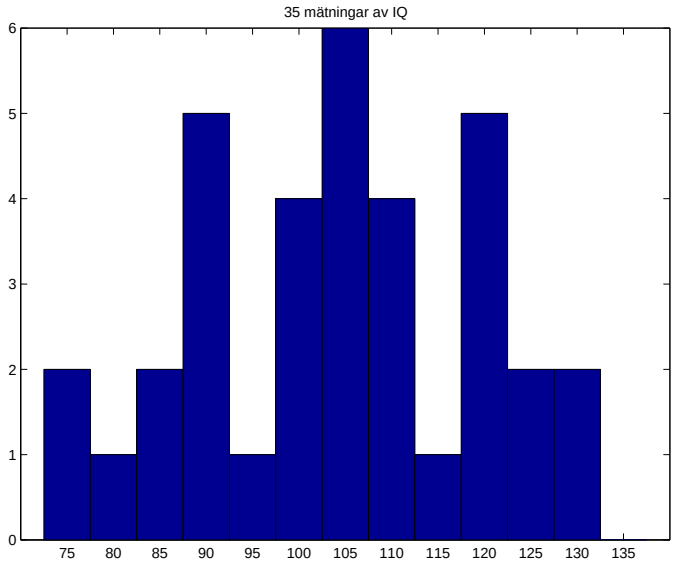
Definition 4 (s 14) Partitionera (dela in) utfallsrummet för variabeln som observeras:

$$x_0 < x_1 < x_2 < \ldots < x_k$$

Vi antar att klassbredderna är lika:

$$x_1 - x_0 = x_2 - x_1 = \ldots = x_k - x_{k-1}$$

Låt f_1 vara antalet observationer i $[x_0, x_1]$, f_2 vara antalet observationer $(x_1, x_2]$, etc, och låt $n = f_1 + f_2 + \ldots + f_k$ vara totala antalet observationer. Avsätt ovanför intervallet $[x_0, x_1]$ en stapel med höjden f_1/n, ovanför intervallet $[x_1, x_2]$ en stapel med höjden f_2/n, etc. Då får du ett *histogram för kontinuerliga data med lika klassbredd.*



Figur 2: Histogram över IQ i ett stickprov om 35 individer.

Definition 5 (s 16) Gör som i definitionen ovan, fast släpp kravet att klasserna ska vara lika stora. Avsätt ovanför intervallet $[x_0, x_1]$ en stapel med bredden $x_1 - x_0$ och höjden

$$\frac{f_1/n}{x_1 - x_0}$$

avsätt ovanför intervallet $[x_1, x_2]$ en stapel med bredden $x_2 - x_1$ och höjden

$$\frac{f_2/n}{x_2 - x_1}$$

etc. Då får du ett *täthets-histogram* för variabeln.

Poängen med ett histogram av denna typ är att totala arean av staplarna blir ett, och att det mer och mer kommer att likna den teoretiska fördelningskurva vi strax ska införa om antalet observationer ökar.

Se **Exempel 1.9**, s 16-17 i D & F.

Kontinuerliga fördelningar (s 24-29)

Definition 6 (s 25) En *täthet* eller *täthetsfunktion* används för att beskriva en kontinuerlig variabels teoretiska fördelning. Följande egenskaper ska en täthetsfunktion uppfylla:

1. $f(x) \geq 0$ för alla $x \in R$
2. $\int_R f(x)\,dx = 1$

Observera att variabelns utfallsrum är

$$\Omega = \{x \in R : f(x) > 0\}$$

När man definierar en täthet $f(x)$ för en variabel x talar man bara om hur den ska beräknas för $x \in \Omega$.

Tanken med definitionen är att proportionen observationer av variabeln i intervallet $(a, b]$ ungefär ska ges av den teoretiska proportionen

$$\int_a^b f(x)\,dx$$

med likhet om antalet observationer vore oändligt.
Obs att det som D & F kallar teoretisk proportion ska vi (snart) kalla *sannolikhet* och till variabeln x ska vi koppla en *stokastisk variabel* X och sannolikheten att $a \leq X \leq b$ är $P(a \leq X \leq b)$, där

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

Figur 1.12 (s 25) Täthets-histogrammet beskriver den experimentella eller empiriska fördelningen och det gäller att den empiriska fördelningen konvergerar mot den teoretiska då antalet observationer går mot oändligheten.

Exempel 1.12 (s 27) Responstiden t (i s) tills något händer antas ha den teoretiska fördelningen

$$f(t) = \frac{1}{5}\, e^{-t/5}, \quad t > 0$$

Är detta en (kontinuerlig) täthet?

Vi kollar att definitionen är uppfylld.

1. $f(t) = 0$ då $t \leq 0$ (konvention) och $f(t) = e^{-t/5}/5 > 0$ då $t > 0$.
2. $\int_{-\infty}^{\infty} f(t)\, dt = \int_{0}^{\infty} f(t)\, dt = \int_{0}^{\infty} e^{-t/5}/5\, dt = \ldots = 1$

Lite längre fram i kursen ska vi införa begreppet stokastisk variabel, främst för att enkelt kunna uttrycka våra tankar i formler. Den stokastiska variabel som svarar mot t ska vi beteckna T, och vi skriver

$$P(a < T \leq b) = \int_{a}^{b} f(t)\, dt$$

för den teoretiska proportionen utfall i intervallet $(a, b]$. Talet

$$P(a < T \leq b)$$

ska vi kalla *sannolikheten att $a < T \leq b$ inträffar*, eller *sannolikheten för händelsen $a < T \leq b$.*

Vi räknar nu ut att

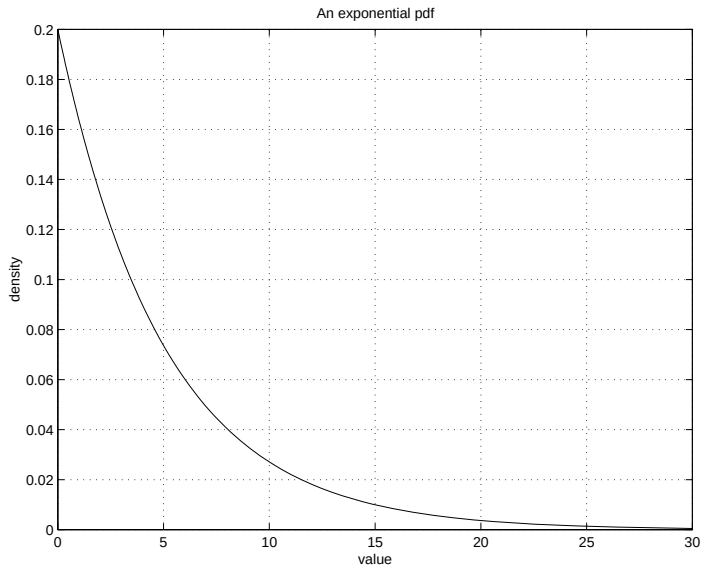
$$P(T \leq 5) = \int_{0}^{5} \frac{e^{-x/5}}{5}\, dx = \ldots \approx 0.632$$

Definition 7 (s 29) Om en variabels täthet är av formen

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

för något $\lambda > 0$, säger vi att den är *exponentialfördelad.*

Exponentialfördelningen betecknas $\mathrm{Exp}(\lambda)$.



Figur 3: Exponentialtätheten i exempel 1.12.

Låt $T \sim \text{Exp}(1/5)$.

Det värde c som uppfyller

$$P(T \leq c) = 0.90$$

ska vi kalla för den 90:e *percentilen* eller 0.9-*kvantilen*. Observera att

$$P(T \leq c) = 0.9 \quad \Leftrightarrow \quad \int_0^c \frac{e^{-x/5}}{5}\, dx = 0.9 \;\Rightarrow \ldots \Rightarrow c \approx 11.5$$

Andra percentiler definieras analogt. De beräknas genom att man löser en ekvation av typen

$$P(T \leq c) = p \quad \Leftrightarrow \quad \int_0^c f(x)\, dx = p$$

där $0 < p < 1$ är givet.

Den 50:e percentilen kallas för *medianen*. Den fås genom att lösa ekvationen

$$P(T \leq c) = 0.5 \quad \Leftrightarrow \quad \int_0^c \frac{e^{-x/5}}{5}\, dx = 0.5$$

Hemövning 1 Visa att medianen i exponentialfördelningen är

$$m = \frac{\ln 2}{\lambda}$$

Diskreta fördelningar (s 29-30)

Definition 8 (s 29) En diskret variabels teoretiska fördelning definieras av en *mass-* eller *frekvensfunktion* $p(x)$ uppfyllande

1. $p(x) \geq 0$ för alla $x \in R$
2. $p(x) > 0$ för högst uppräkneligt många x
3. $\sum_x p(x) = 1$

där summationen sker över de högst uppräkneligt många möjliga variabelvärdena.

Om de möjliga utfallen är $x_1 < x_2 < \ldots$, så ska alltså massfunktionen uppfylla

1. $p(x_i) > 0$ för alla x_i, och $p(x) = 0$ för alla andra x
2. $\sum_i p(x_i) = 1$

Den teoretiska proportionen observationer i "intervallet" $\{x_k, x_{k+1}, \ldots, x_l\}$ ges av

$$p(x_k) + p(x_{k+1}) + \ldots + p(x_l) = \sum_{i=k}^{l} p(x_i)$$

Detta kallar vi *sannolikheten för ett utfall mellan x_k och x_l*. Om motsvarande stokastiska variabel kallas X skriver vi

$$P(X \in \{x_k, \ldots, x_l\}) = P(x_k \leq X \leq x_l) = \sum_{i=k}^{l} p(x_i)$$

Exempel 1.13 (s 30) Antag att x är diskret och har massfunktionen

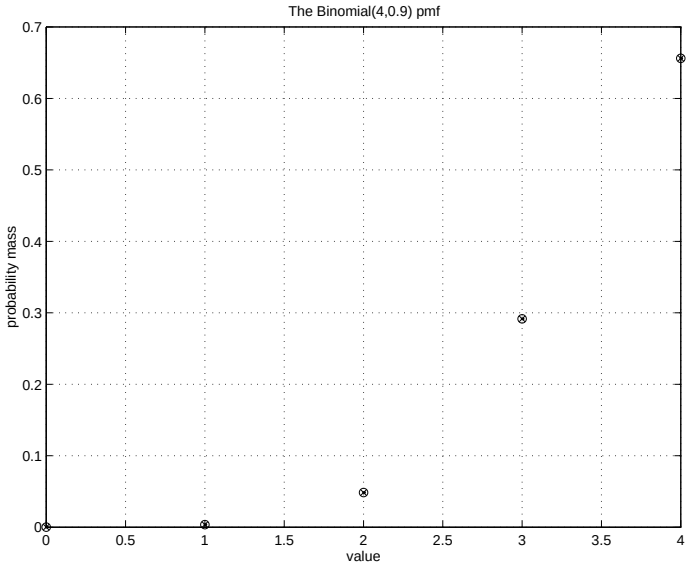
$$p(k) = \underbrace{\frac{4!}{k!\,(4-k)!}}_{= \binom{4}{k}} 0.9^k\, 0.1^{4-k}, \quad k = 0, 1, 2, 3, 4$$

Att detta är en massfunktion följer av Binomialteoremet: För reella a, b och positiva heltal n gäller

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k \, b^{n-k}$$

Obs

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$



Figur 4: Binomialtätheten i exempel 1.13

Hemövning 2 Beräkna $P(1 \leq X \leq 3)$.

1.4 Normalfördelningen (s 32-40)

Definition 9 (s 32) En kontinuerlig variabel x (eller stokastisk variabel X) säges vara *normalfördelad* om dess täthetsfunktion är av typen

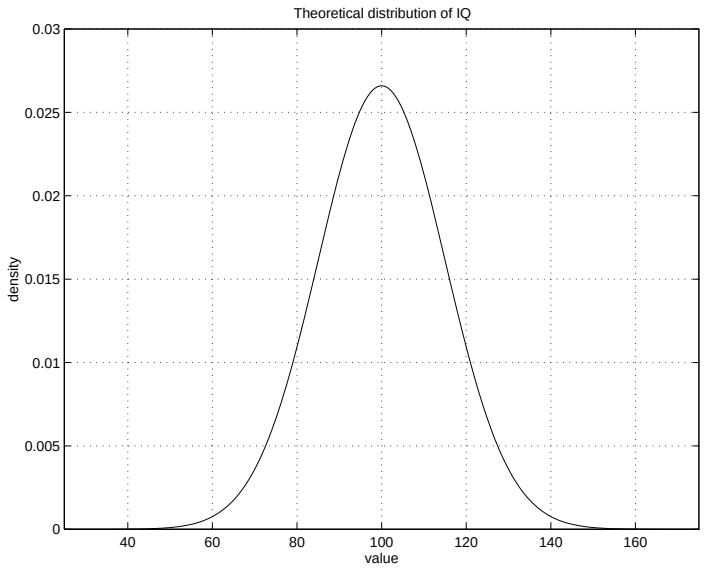
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \, e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty$$

där $-\infty < \mu < \infty$ och $\sigma > 0$. Talen μ, σ är fördelningens parametrar. Av uttrycket för tätheten framgår att μ bestämmer fördelningens läge och σ hur utspridd den är, samt att utfallsrummet är hela R.

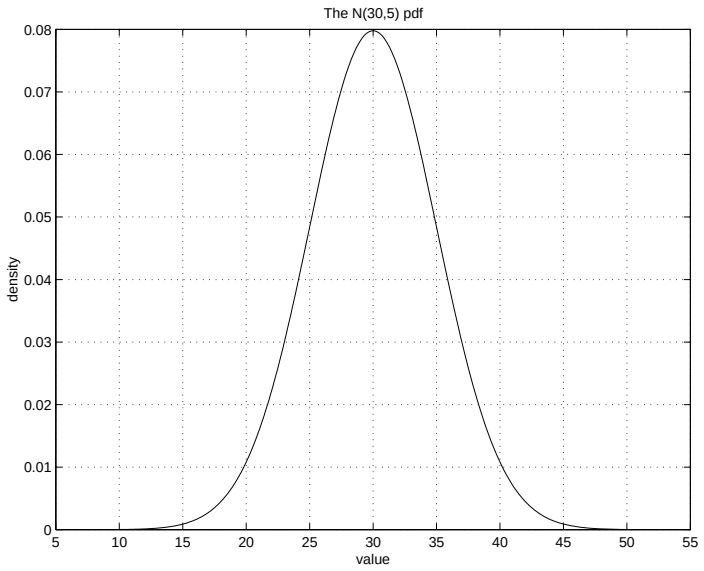
Normalfördelningen betecknas $\mathrm{N}(\mu, \sigma)$.

Så här beräknar man sannolikheter för normalfördelade variabler:

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$



Figur 5: Teoretisk IQ är $\mathrm{N}(100, 15)$.

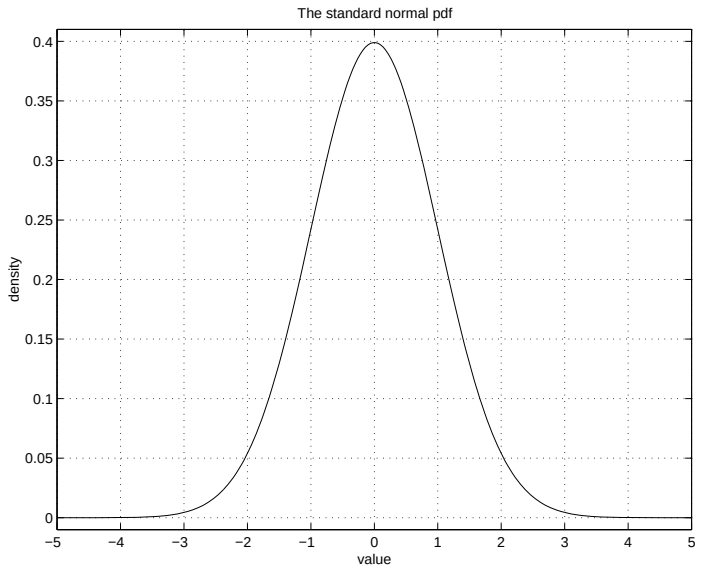


Figur 6: N(30, 5)-tätheten.

Definition 10 (s 34) Då $\mu = 0$ och $\sigma = 1$ brukar man säga att normalfördelningen är *standardiserad*. Motsvarande täthet är

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty$$

Skälet till att den oberoende variabeln här benämns z är att man brukar använda just z eller Z som beteckning för en normalfördelad variabel med $\mu = 0$ och $\sigma = 1$.



Figur 7: N(0, 1)-tätheten.

Operationen som avbildar x på $(x-\mu)/\sigma$ brukar benämnas *standardisering*.

Den standardiserade normalfördelningsfunktionen är

$$\Phi(z) = \int_{-\infty}^{z} \varphi(z)\, dz$$

Exempel 7 Antag att vi vet att en variabel x är normalfördelad med parametrar $\mu = 5$ och $\sigma = 2$, och att vi vill veta den teoretiska proportionen observationer i intervallet $[3, 7]$. Då ska vi beräkna

$$\begin{aligned}\int_3^7 f(x)\, dx &= \int_3^7 \frac{1}{\sqrt{2\pi}\cdot 2}\, e^{-(x-5)^2/2\cdot 2^2}\, dx \\ &= \ldots \\ &= \int_{-1}^{1} \varphi(z)\, dz\end{aligned}$$

(Gör substitutionen $z = (x-5)/2$. Då $dz = dx/2$ samt $x = 3 \Rightarrow z = -1$ och $x = 7 \Rightarrow z = 1$.)

Ur tabellen över normalfördelningen s 534-535 (eller insidan av pärmen) fås nu att

$$\begin{aligned}\int_{-1}^{1}\varphi(x)\,dx &= \int_{-\infty}^{1}\varphi(x)\,dx - \int_{-\infty}^{-1}\varphi(x)\,dx \\ &= \Phi(1) - \Phi(-1) \\ &= 0.8413 - 0.1587 \\ &= 0.6826\end{aligned}$$

Ca 68% av observationerna kommer alltså att hamna i intervallet $[3, 7]$. Denna approximation blir bättre ju fler observationer som görs.

Man kan visa att om x är normalfördelad med parametrar μ och σ, så är $z = (x - \mu)/\sigma$ normalfördelad med $\mu = 0$ och $\sigma = 1$. Operationen

$$x \mapsto \frac{x - \mu}{\sigma}$$

brukar därför benämnas *standardisering*.

Ett annat sätt att formulera detta är att om den stokastiska variabeln X är $\mathrm{N}(\mu, \sigma)$-fördelad, vilket brukar skrivas $X \sim \mathrm{N}(\mu, \sigma)$, så är $(X - \mu)/\sigma$ $\mathrm{N}(0, 1)$-fördelad. M.a.o,

$$X \sim \mathrm{N}(\mu, \sigma) \quad \Rightarrow \quad \frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1)$$

Exempel 1.15 (s 36) Vilket z-värde c är sådant att arean under z-kurvan till vänster om c är 0.67?

Alltså, om variabeln z är $\mathrm{N}(0, 1)$-fördelad, för vilket c gäller att den teoretiska proportionen observationer $\leq c$ är lika med 0.67?

Alltså, om den stokastiska variabeln $Z \sim \mathrm{N}(0, 1)$, för vilket c gäller att

$$P(Z \leq c) = 0.67$$

Vi ska alltså bestämma den 67:e percentilen i den standardiserade normalfördelningen. Vi löser därför ut c ur

$$\int_{-\infty}^{c}\varphi(z)\,dz = 0.67$$

Rätt svar är $c = 0.43991$. Ett approximativt svar erhålles genom att man i normalfördelningstabellen letar upp sannolikheten $p = 0.67$ och i marginalerna läser av motsvarande z-värde.

Hemövning 3 Bestäm den 95:e percentilen i den standardiserade normalfördelningen.

Hemövning 4 Låt $X \sim \mathrm{N}(100, 15)$. Bestäm den 95:e percentilen i X:s fördelning. Tolka resultatet.

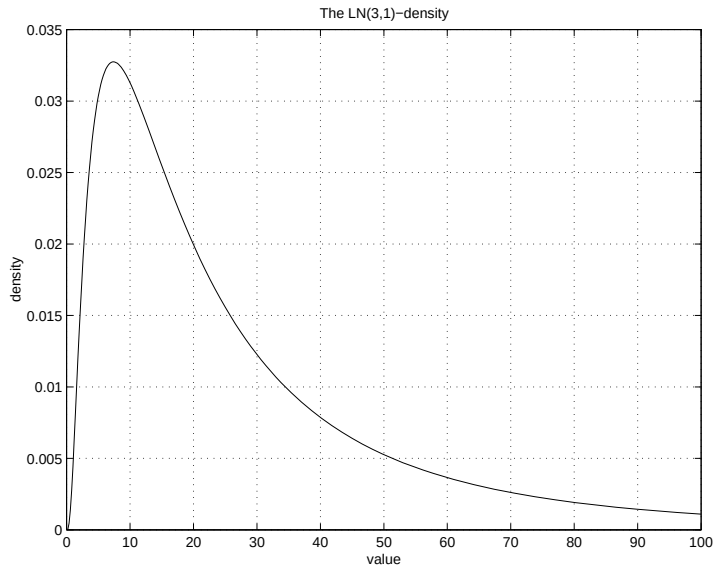
Hemövning 5 För vilket z-värde gäller

$$P(-z \leq Z \leq z) = 0.95$$

Vad innebär resultatet för en $\mathrm{N}(\mu, \sigma)$-variabel?

1.5 Fler kontinuerliga fördelningar (s 42-46)

Definition 11 (s 43) En icke-negativ kontinuerlig variabel x säges vara lognormalfördelad om $y = \ln x$ är normalfördelad.



Figur 8: Ett exempel på en lognormal-täthet.

Obs att $y = \ln x \Leftrightarrow x = e^y$.

Läs själva i D & F om lognormalfördelningen. Glöm inte att studera Figur 1.29, s 43.

Läs själva i D & F om Weibullfördelningen, som många gånger är en bra modell för livslängd i utmattningssammanhang.

1.6 Ett par viktiga diskreta fördelningar (s 47-52)

Definition 12 (s 48) En variabel sägs vara Binomialfördelad om den är diskret och dess massfunktion är

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$

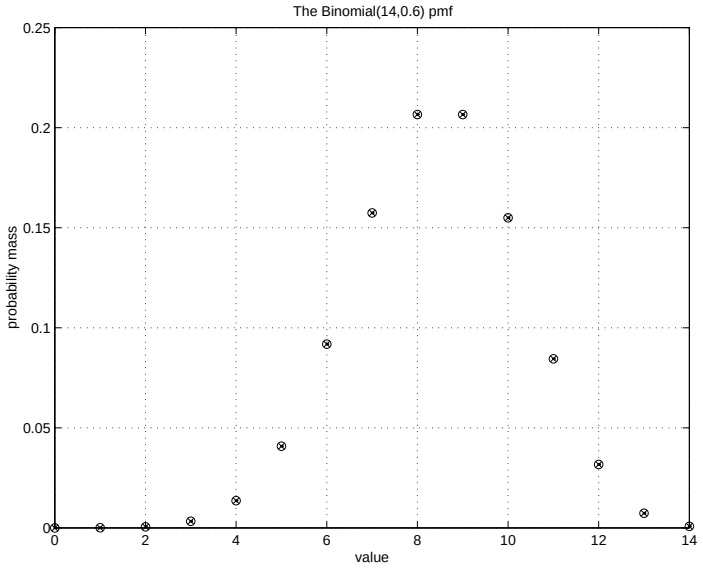
för något heltal $n \geq 1$ och någon sannolikhet $p \in (0, 1)$.

Binomialfördelningen betecknas $\text{Bin}(n, p)$.

Notera att $p(k) > 0$ för $k = 0, 1, \ldots, n$ och att

$$\sum_k p(k) = (p + (1-p))^n = 1$$

enligt Binomialteoremet (som vi redan har haft uppe till diskussion).



Figur 9: Massfunktionen Bin$(14, 0.6)$.

Antag att man gör försök. I varje försök är man intresserad av huruvida en viss händelse A inträffar eller ej. Låt p vara den teoretiska proportionen som A inträffar i (d v s proportionen för A om antalet försök $\to \infty$. Vi ska senare kalla detta tal för sannolikheten för A och beteckna det $P(A)$.)

Tänk dig nu att man utför försöken så att de blir oberoende av varandra. D v s inget av det som händer i försök nr 1 får lov att påverka vad som händer i de andra försöken, inget av det som händer i försök nr 2 får lov att påverka vad som händer i de andra försöken, etc. Tänk dig också att vi gör exakt n försök. Låt f vara frekvensen för A, d.v.s. antalet gånger A inträffar. Då är f Binomialfördelad med parametrar n och p.

Vi skriver $f \sim \text{Bin}(n, p)$.

Om vi nu låter $n \to \infty$ och $p \to 0$ på ett sådant sätt att $np \to \lambda > 0$, så kommer

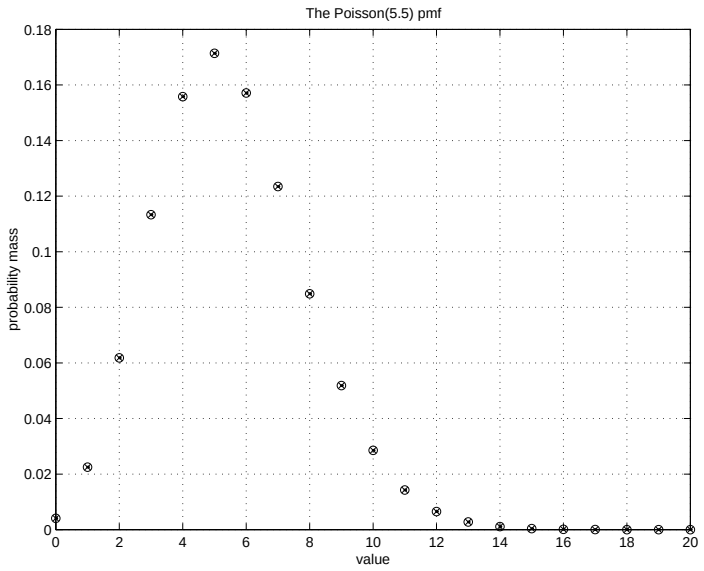
$$\binom{n}{k} p^k (1-p)^{n-k} \to e^{-\lambda} \frac{\lambda^k}{k!}$$

Till höger ovan har vi massfunktionen för Poissonfördelningen, som är nästa fördelningsklass som vi ska känna till.

Definition 13 (s 50) *Poissonfördelningens* massfunktion är

$$p(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

Att detta är en massfunktion ser man genom att Taylorutveckla e^{λ}. Glöm aldrig att Poissonfördelningen är diskret.



Figur 10: Massfunktionen Poi(5.5).

Hemövning 6 Låt variabeln x vara Poissonfördelad med parameter $\lambda = 5.5$. Beräkna den teoretiska proportionen observationer skilda från noll.

Förr approximerade man ofta Binomialfördelningen med Poissonfördelningen om proportionen p var liten. Det behövs inte nu när vi har så bra beräkningsverktyg.

Men approximationen är viktig av principiella skäl, ty den säger oss att Poissonfördelningen är en bra modell för inträffandet av ovanliga händelser.

Exempel 8 Vi har tidigare noterat att antalet cykelolyckor utmed Vasagatan en månad, vilken som helst, är en diskret variabel eftersom värdemängden består av talen $0, 1, 2, \ldots$. För varje cyklist som cyklar utmed gatan tänker vi oss ett slumpmässigt försök där vi observerar ifall cyklisten råkar ut för en olycka eller ej.

Av det som redan sagts förstår vi att frekvensen olyckor f är Binomialfördelad med parametrar n och p, där n är totala antalet cyklister och p är den teoretiska proportionen olyckor. Så frekvensen olyckor under en månad är approximativt Poissonfördelad med parameter $\lambda = np$.

Varken n eller p är lätta att mäta och kanske inte heller så intressanta. Däremot kan man m h a olycksstatistik skaffa sig kunskap om parametern λ. Hur det går till ska vi se senare i kursens sista del.