Kapitel 5

Fordelinger på den reelle akse

5.1. Kontinuerte fordelinger.

Fra elementær mekanisk fysik kendes begrebet en *punktformet masse.* Man definerer f.eks. et “matematisk pendul” som en vægtløs stang af en given længde, ophængt frit i sit ene endepunkt, og i den anden ende forsynet med en sådan punktformet masse.

Begrebsmæssigt er der en oplagt analogi mellem følgende to objekter fra henholdsvis fysik og sandsynlighedsregning:

(1) En (uendeligt lang) vægtløs stang, hvortil er fæstnet et (evt. uendeligt) antal punktformede masser med sum 1 (= 1 gram, f.eks.)

(2) En diskret sandsynlighedsfordeling på $\mathbf{R}$.

Analogien behøver næppe yderligere forklaring. I forbifarten skal blot nævnes, at de fysiske begreber *tyngdepunkt* og *inertimoment* her svarer til de sandsynlighedsteoretiske begreber middelværdi og varians.

Når vi trækker analogien frem her, er det fordi den generalisering, vi nu skal til at foretage fra diskrete til kontinuerte fordelinger, i alt væsentligt svarer til den man foretager i fysikken, når man erstatter systemer af punktformede masser med faste legemer af given størrelse og form, hvis masse antages at være jævnt fordelt over hele legemets rumfang. Og bortset fra, at de fysiske begreber hører til i tre dimensioner, medens man i sandsynlighedsregningen beskedent starter med én (for til gengæld at ende med n), så er begrebet *tæthed*, som vi skal diskutere i næste paragraf, det sandsynlighedsteoretiske modstykke til det fysiske begreb *massefylde.*

Eksempel 5.1.1 (Ligefordelingen på enhedsintervallet). Betragt en stokastisk variabel X med værdier i enhedsintervallet $[0, 1]$ med den egenskab, at der for $0 \le a \le b \le 1$ gælder

$$(*) \qquad P(X \in [a, b]) = b - a.$$

Der findes ikke noget diskret sandsynlighedsmål, som beskriver denne situation. For $a = b$ skulle der jo så gælde $P(X = a) = 0$, svarende til at alle punktsandsynligheder var 0. Men vi kan konstruere diskrete fordelinger, som approksimerer denne situation. Hvis vi som X's fordeling tager ligefordelingen på $\{\frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1\}$ for et meget stort tal $N \in \mathbf{N}$, så vil $(*)$ være approksimativt opfyldt. For $N = 10^6$ vil ligefordelingen på $\{0.000001, 0.000002, \ldots, 0.999999, 1.000000\}$ f.eks. kunne opfattes som fordelingen af en “ægte” kontinuert ligefordelt variabel, som er afrundet (opad) til seks betydende cifre. For $N \to \infty$ fås i grænsen (hvad det så betyder) selve *ligefordelingen* eller *den rektangulære fordeling* på enhedsintervallet, som tilfredsstiller $(*)$. Denne fordeling siges at have *tæthed* 1 på enhedsintervallet, fordi forholdet mellem sandsynlighedsmassen i et interval og intervallets længde netop er 1.

EKSEMPEL 5.1.2 (Eksponentialfordelingen). I forbindelse med Poissonfordelingen og den geometriske fordeling var vi nær på at indføre en kontinuert fordeling. Faktisk måtte vi, i forbindelse med vores cyklist C (eksempel 3.6.2) og taxaeksemplet (opgave 3.6.1 og opgave 3.7.3) foretage en del krumspring for ikke at gøre det. Lad, for vores cyklist C, X betegne den præcise distance hun har tilbagelagt, når hendes første punktering indtræffer. Vi bemærkede, at antal punkteringer indenfor en distance på x kilometer ville følge en Poissonfordeling med parameter $\lambda = 0.003x$. Det har vi ganske vist kun benyttet for heltallige km–tal x, men det er klart at det gælder for vilkårlige $x > 0$. Heraf følger, at sandsynligheden for "0 punkteringer i intervallet fra 0 til x km" er $e^{-0.003x}$ (= Poissonfordelingens punktsandsynlighed i 0). Eller

$$P(X > x) = e^{-0.003x}.$$

Ligesom i foregående eksempel har vi åbenbart at gøre med en kontinuert fordeling, idet halesandsynligheden $P(X > x)$ afhænger kontinuert af x. I en diskret fordeling vil halesandsynligheden jo aftage i spring, som finder sted hver gang x passerer et punkt af støtten. Specielt kan fordelingen af X ikke beskrives ved en sandsynlighedsfunktion, fordi der gælder $P(X = x) = 0$ for ethvert x. Men vi kan i stedet, for ethvert punkt x, angive fordelingens tæthed, som er forholdet mellem sandsynlighedsmasse og intervallængde for et lille interval omkring (eller nær) x:

$$\begin{aligned}\frac{P(X \in [x, x+h])}{h} &= \frac{P(X \geq x) - P(X > x+h)}{h} \\ &= \frac{e^{-0.003x} - e^{-0.003(x+h)}}{h} \\ &= e^{-0.003x}\frac{1 - e^{-0.003h}}{h},\end{aligned}$$

som for $h \to 0$ konvergerer mod

$$p(x) = 0.003e^{-0.003x}.$$

Fortolkningen af tætheden $p(x)$ går således ud på, at der for et lille interval $[x, x+h]$ gælder

$$P(X \in [x, x+h]) \approx p(x)h.$$

Fordelingen med tæthed $p(x) = 0.003e^{-0.003x}$ kaldes *eksponentialfordelingen* med *skalaparameter* $1/0.003 = 333.33$. Den normerede variabel $X_0 = X/333.33$ (som er distancen til første punktering, målt i "middel punkterings distancer") vil være *normeret eksponentialfordelt*, dvs.

$$P(X_0 > x_0) = e^{-x_0}.$$

Tætheden for denne fordeling bliver åbenbart $p_0(x_0) = e^{-x_0}$, ved et tilsvarende argument.

OPGAVE 5.1.1. Lad X være ventetiden til første punktering i eksempel 5.1.2. Gør rede for, at

$$R = e^{-X/333.33} = e^{-X_0}$$

er rektangulært fordelt på enhedsintervallet, jvf. eksempel 5.1.1.

5.2. Tæthed.

§5.1 var kun opvarmning. Nu går vi mere brutalt til værks:

DEFINITION. Ved en *sandsynlighedstæthed* p på et interval $E \subseteq \mathbf{R}$ forstås en funktion $p\colon E \to \mathbf{R}$, som opfylder følgende to betingelser:

$$\text{(p1)} \qquad p(x) \geq 0,$$

$$\text{(p2)} \qquad \int_E p(x)\,dx = 1.$$

EKSEMPEL 5.2.1. For $E = [a, b]$ vil den konstante funktion $p(x) = \frac{1}{b-a}$ være en sandsynlighedstæthed. Den tilsvarende fordeling kaldes den *rektangulære fordeling* eller *ligefordelingen* på $[a, b]$.

For at præcisere definitionen ovenfor er vi nødt til at beskæftige os lidt med integration. Når vi taler om integralet $\int_a^b f(x)\,dx$ eller $\int_{[a,b]} f(x)\,dx$ af en funktion f på $[a, b]$ vil vi i første omgang antage, at f er en "pæn" funktion, hvormed vi mener en begrænset, kontinuert – eller i det mindste stykkevis kontinuert – funktion. Integralet kan så defineres på sædvanlig måde, f.eks. ved

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \frac{b-a}{n} \sum_{i=1}^{n} f\left(a + (i - \frac{1}{2}) \times \frac{b-a}{n}\right).$$

Her er udtrykket på højre side (efter limes–tegnet) en såkaldt *middelsum*, i dette tilfælde svarende til inddelingen

$$a < a + \frac{b-a}{n} < a + 2\frac{b-a}{n} < \cdots < b$$

af intervallet $[a,b]$ i n lige store intervaller. Værdierne $f\left(a+(i-\frac{1}{2})\frac{b-a}{n}\right)$ er her valgt som funktionens værdier i de tilsvarende intervalmidtpunkter, men de kan erstattes med $f(x_i)$ for vilkårlige $x_i \in [a+(i-1)\frac{b-a}{n}, a+i\frac{b-a}{n}]$. I tilfælde af en kontinuert funktion f er konvergensen for $n \to \infty$ en velkendt konsekvens af en hovedsætning om kontinuerte funktioner. Hvis f kun er stykkevis kontinuert, følger konvergensen af, at bidragene fra de intervaller, der indeholder eller støder op til et af de endeligt mange diskontinuitetspunkter, er uden betydning i grænsen. Alternativt kan man definere integralet af en stykkevis kontinuert funktion f som summen af integralerne over de intervaller, hvor f er kontinuert, og dette giver samme resultat.

Vi får brug for to udvidelser af det elementære integralbegreb. Den vigtigste af dem går ud på følgende:

Definition. Lad $f\colon \mathbf{R} \to \mathbf{R}$ være en ikke–negativ funktion. Antag, at den voksende følge af integraler

$$I_n = \int_{-n}^{n} f(x)\,dx$$

er begrænset. Vi siger så, at f er *integrabel* med integralet

$$\int_{-\infty}^{+\infty} f(x)\,dx = \lim_{n\to\infty} I_n.$$

Hvis $I_n \to +\infty$ skriver man naturligt $\int_{-\infty}^{+\infty} f(x)\,dx = +\infty$ (men f kaldes ikke integrabel i dette tilfælde).

Integralet af en ikke–negativ funktion på $[0,+\infty[$, $]-\infty,0]$ eller et andet "halv–uendeligt" interval defineres tilsvarende.

Eksempel 5.2.2. I eksempel 5.1.2 angav vi tætheden for den normerede eksponentialfordeling som funktionen $p\colon [0,+\infty[\to \mathbf{R}$ givet ved $p(x) = e^{-x}$. Eftersom integralet

$$I_n = \int_0^n e^{-x}\,dx = 1 - e^{-n}$$

konvergerer mod 1 for $n \to \infty$, kan vi skrive

$$\int_0^{+\infty} e^{-x}\,dx = 1,$$

så funktionen p er en sandsynlighedstæthed på $[0,+\infty[$, efter de definitioner vi har givet.

Vi får senere brug for en tilsvarende definition af integraler af funktioner på hele aksen, som kan antage både positive og negative værdier. Dette problem løser vi ved opdeling i positiv og negativ del. En vilkårlig funktion $f\colon \mathbf{R} \to \mathbf{R}$ kan skrives som differens

$$f = f_+ - f_-$$

mellem de to ikke-negative funktioner $f_+ = f \vee 0$ og $f_- = (-f) \vee 0$, som kaldes henholdsvis f's positive og negative del. For at kunne tale om $\int_{-\infty}^{+\infty} f(x)\,dx$ kræver vi, at begge de ikke–negative funktioner f_+ og f_- er integrable, og vi definerer så

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_{-\infty}^{+\infty} f_+(x)\,dx - \int_{-\infty}^{+\infty} f_-(x)\,dx.$$

Betingelsen ses i øvrigt let at være ækvivalent med, at funktionen $|f| = f_+ + f_-$ er integrabel. Hvis betingelsen ikke er opfyldt, tillægger vi ikke f noget integral.

Den anden udvidelse i forhold til det elementære integralbegreb, som vi får brug for, går ud på følgende. Lad $f\colon\,]a,b] \to \mathbf{R}$ være en ikke–negativ funktion, som ikke er defineret i punktet a. Den typiske situation, som vi tænker på, er den hvor $f(x) \to +\infty$ for $x \to a$, således at man naturligt ville skrive $f(a) = +\infty$. Her kan vi betragte den voksende følge af integraler

$$I_n = \int_{a+\frac{1}{n}}^{b} f(x)\,dx.$$

Hvis denne følge er begrænset, sætter vi

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} I_n,$$

og funktionen siges da at være integrabel. I modsat fald skriver vi evt. $\int_a^b f(x)\,dx = +\infty$ (og kalder ikke funktionen integrabel). Situationen er helt analog til den vi betragtede for funktioner defineret på hele aksen (eller en halvakse), og udvidelsen til funktioner, som kan antage både positive og negative værdier, kan foretages på præcis samme måde.

Eksempel 5.2.3. Betragt funktionen

$$f(x) = \frac{1}{\sqrt{x}}.$$

Denne funktion er defineret for alle $x > 0$, men vi opfatter den her som en funktion på $]0,1]$. Da integralet

$$I_n = \int_{\frac{1}{n}}^{1} \frac{1}{\sqrt{x}}\,dx = 2 - \frac{2}{\sqrt{n}}$$

vokser mod 2 for $n \to \infty$, er denne funktion integrabel med integralet 2. Heraf følger i øvrigt, at funktionen

$$p(x) = \frac{1}{2} f(x) = \frac{1}{2\sqrt{x}}$$

er en sandsynlighedstæthed på enhedsintervallet. Den tilsvarende fordeling kan fortolkes som fordelingen af R^2, hvor R er en stokastisk variabel som er ligefordelt på enhedsintervallet. Dette følger af, at der for en sådan variabel gælder

$$P(R^2 \in [x, x+h]) = P\left(R \in \left[\sqrt{x}, \sqrt{x+h}\,\right]\right) = \sqrt{x+h} - \sqrt{x} \approx \frac{1}{2\sqrt{x}} h,$$

fordi funktionen $\sqrt{x}$ er differentiabel med differentialkvotient $\frac{1}{2\sqrt{x}}$. Bemærk, at fordelingen af $X = R^2$ i en vis forstand har tæthed $+\infty$ i punktet 0.

Tilsvarende udvidelser af det elementære integralbegreb kan anvendes i en lang række tilfælde, som vi ikke direkte har omtalt. For eksempel kan en ikke–negativ funktion $f :]0, +\infty[\to \mathbf{R}$, som er udefineret i punktet 0, naturligt tillægges integralet

$$\int_0^{+\infty} f(x)\, dx = \lim_{n \to \infty} \int_{\frac{1}{n}}^{n} f(x)\, dx.$$

Vi kan ligeledes på naturlig måde tillægge en ikke–negativ funktion på et interval $[a, b]$ et integral, selv om den er udefineret eller $+\infty$ i et indre punkt x_0 af intervallet, ved at sætte

$$\int_a^b f(x)\, dx = \lim_{n \to \infty} \left(\int_a^{x_0 - \frac{1}{n}} f(x)\, dx + \int_{x_0 + \frac{1}{n}}^{b} f(x)\, dx \right).$$

Adskillige varianter er mulige, og alle kan generaliseres til tilfældet, hvor $f(x)$ kan antage både positive og negative værdier, præcis som vi gjorde det for funktioner på hele aksen.

Til slut skal blot nævnes, at vi ofte vil udelade grænserne ved opskrivning af integraler, når dette ikke kan føre til misforståelser. Formelt kan man tænke på ethvert integral som et integral fra $-\infty$ til $+\infty$, hvor man har sat integranden til 0 udenfor det interval, der i virkeligheden skal integreres over. Ligefordelingen på $[a, b]$ har for eksempel den konstante tæthed $\frac{1}{b-a}$ på intervallet $[a, b]$, men vi kan også tænke på denne fordeling som en fordeling på hele aksen med den stykkevis konstante tæthed

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{ellers.} \end{cases}$$

Vi kan så, uden fare for misforståelser, skrive $\int p(x)\, dx = 1$. Derfor, når vi i det følgende taler om "en sandsynlighedstæthed på $\mathbf{R}$" kan det for det meste lige så godt læses "en sandsynlighedstæthed på $\mathbf{R}$ eller et interval $E \subseteq \mathbf{R}$".

OPGAVE 5.2.1. Lad $f\colon [a,b] \to \mathbf{R}$ være en kontinuert funktion. Lad X_n være en diskret stokastisk variabel, ligefordelt på mængden $\{a + i\frac{b-a}{n} \mid i = 1, 2, \ldots, n\}$. Vis at $\mathrm{E}f(X_n) \to \frac{1}{b-a}\int_a^b f(x)\,dx$ for $n \to \infty$ (jvf. den i eksempel 5.1.1 omtalte approksimation af ligefordelingen på enhedsintervallet med diskrete fordelinger). Giv et forslag til definition af $\mathrm{E}f(X)$ når X er (kontinuert) ligefordelt på $[a,b]$. Hvad sker der, hvis den kontinuerte funktion f erstattes med indikatoren 1_A for et interval $A \subseteq [a,b]$?

5.3. Sandsynlighedsfordeling og fordelingsfunktion.

Vi har flere gange i dette kapitel omtalt den til en sandsynlighedstæthed knyttede fordeling, uden at give en præcis definition. Det vil vi heller ikke gøre her, men vi vil i det mindste give en upræcis:

DEFINITION. Lad p være en sandsynlighedstæthed på $\mathbf{R}$. Den til p hørende sandsynlighedsfordeling P er den afbildning, som til enhver pæn delmængde A af $\mathbf{R}$ knytter tallet

$$P(A) = \int_A p(x)\,dx = \int 1_A(x)p(x)\,dx.$$

Det upræcise ligger naturligvis i, at vi ikke tager stilling til hvad en "pæn" delmængde af $\mathbf{R}$ er for noget. Vi vil nøjes med at anføre, at intervaller og endelige foreningsmængder af intervaller i hvert fald hører til blandt de "pæne". Integranden $1_A(x)p(x)$ bliver i dette tilfælde en stykkevis kontinuert funktion, så integralet er veldefineret.

Selv om vi således ikke gør os helt klart, hvad P's definitionsmængde er, kan vi godt konstatere, at P har mange egenskaber fælles med de mere præcist definerede diskrete sandsynlighedsfordelinger, som blev indført i §1.3. Der gælder således

$$0 \le P(A) \le 1,$$

som man ser ved at integrere uligheden $0 \le 1_A(x)p(x) \le p(x)$. For disjunkte "pæne" mængder A og B gælder

$$P(A \cup B) = P(A) + P(B),$$

hvilket ses ved integration af identiteten $1_{A\cup B}(x)p(x) = 1_A(x)p(x) + 1_B(x)p(x)$. Ligeledes er $P(\mathbf{R}) = 1$, og alle de regneregler for diskrete fordelinger, som er afledt direkte af disse egenskaber, gælder naturligvis også. Ligesom i det diskrete tilfælde skal $P(A)$ fortolkes som sandsynligheden for, at det tilfældige udfald (her den reelle stokastiske variable) havner i mængden A.

Det væsentlige ved ovenstående definition er, at den præciserer tæthedens fortolkning: Hvis en reel stokastisk variabel X har (fordeling med) tæthed p, så er sandsynligheden for, at X havner i et interval A (eller en foreningsmængde A af endeligt mange intervaller) bestemt ved

$$P(X \in A) = \int 1_A(x) p(x)\, dx.$$

Bemærk, at denne fortolkning er i overensstemmelse med den vi hidtil har anlagt, som går ud på, at tætheden er det lokale forhold mellem sandsynlighedsmasse og intervallængde. For et lille interval $[x_0, x_0 + h]$ gælder jo

$$\frac{P(X \in [x_0, x_0 + h])}{h} = \frac{1}{h} \int_{x_0}^{x_0+h} p(x)\, dx \approx p(x_0),$$

forudsat at p er kontinuert i punktet x_0. I de endeligt mange diskontinuitetspunkter, som en "pæn" funktion kan have, har denne approksimation ingen mening; men for kontinuerte fordelinger er det ikke så vigtigt, hvad der sker i endeligt mange punkter. Fordelingen afhænger jo ikke engang af, hvordan p defineres i disse punkter.

Definition. Ved den til p (eller P, eller X) hørende *fordelingsfunktion* forstås funktionen $F\colon \mathbf{R} \to \mathbf{R}$ givet ved

$$F(x) = P(]-\infty, x]) = P(X \leq x) = \int_{-\infty}^{x} p(x')\, dx'.$$

Fordelingsfunktionen er altså tæthedens stamfunktion, eller dens ubestemte integral. Den er voksende (eller rettere: ikke–aftagende) med

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

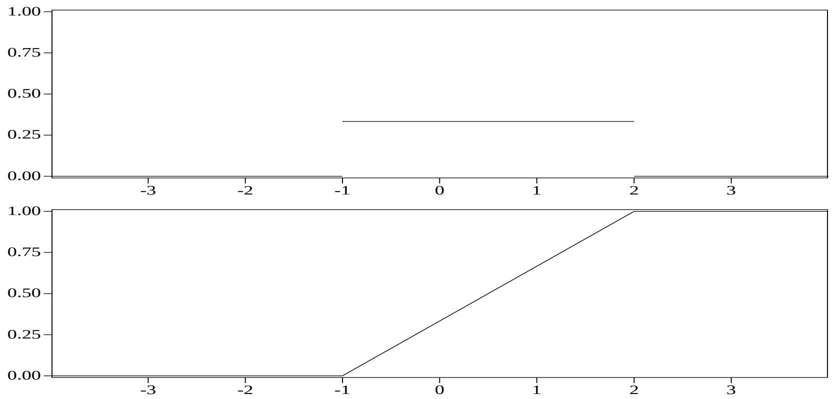
Af sammenhængen mellem differentiation og integration følger, at fordelingsfunktionen i ethvert kontinuitetspunkt for tætheden er differentiabel med

$$\frac{d}{dx} F(x) = F'(x) = p(x).$$

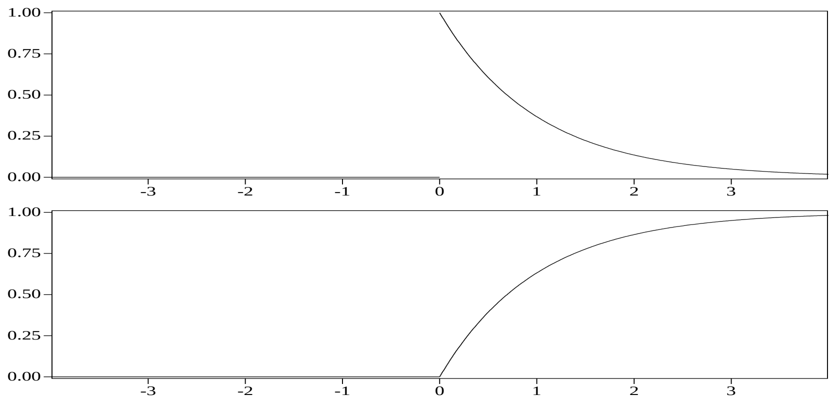
I tæthedens diskontinuitetspunkter vil F normalt ikke være differentiabel (men altid kontinuert).

Eksempel 5.3.1. Ligefordelingen på et interval $[a, b]$ har, opfattet som fordeling på hele aksen, den stykkevis lineære fordelingsfunktion

$$F(x) = \int_{-\infty}^{x} \frac{1}{b-a} 1_{[a,b]}(x')\, dx' = \begin{cases} 0 & \text{for } x < a, \\ \frac{x-a}{b-a} & \text{for } a \leq x \leq b, \\ 1 & \text{for } b < x. \end{cases}$$



Tæthed og fordelingsfunktion for rektangulær fordeling på [-1,2]



Tæthed og fordelingsfunktion for normeret eksponentialfordeling

Eksempel 5.3.2. Den normerede eksponentialfordeling har fordelingsfunktionen

$$F(x) = \int_{-\infty}^{x} 1_{[0,+\infty[}(x')e^{-x'}\,dx' = \begin{cases} 0 & \text{for } x < 0, \\ 1 - e^{-x} & \text{for } 0 \leq x. \end{cases}$$

Af relationen $p(x) = F'(x)$ følger, at sandsynlighedstætheden – og dermed fordelingen – kan rekonstrueres ud fra fordelingsfunktionen. Man kan også mere direkte opskrive fordelingen ved hjælp af fordelingsfunktionen, idet der jo for ethvert interval $[a, b]$ gælder

$$P([a, b]) = F(b) - F(a).$$

På denne måde kan en intervalsandsynlighed udregnes som fordelingsfunktionens tilvækst over intervallet, og additionsreglen gør det herefter muligt at udtrykke sandsynligheder for endelige foreningsmængder af intervaller ved hjælp af fordelingsfunktionen.

Eksempel 5.3.3. Lad X være ligefordelt på $[-1,2]$. Så er

$$\begin{aligned} P\left(|X| \geq \frac{2}{3}\right) &= P\left(X \in [-1, -\frac{2}{3}] \cup [\frac{2}{3}, 2]\right) \\ &= \left(F(-\frac{2}{3}) - F(-1)\right) + \left(F(2) - F(\frac{2}{3})\right) \\ &= \left(\frac{1}{9} - 0\right) + \left(1 - \frac{5}{9}\right) = \frac{5}{9}. \end{aligned}$$

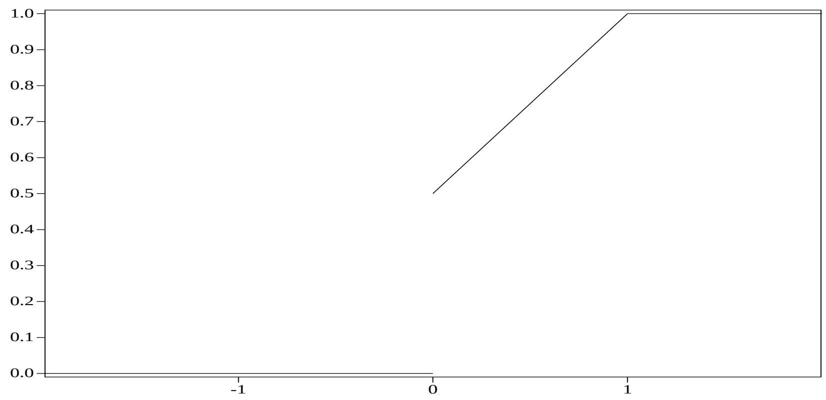
Kontinuerte fordelinger, som vi har defineret dem i denne fremstilling, angives mest naturligt ved deres tætheder, ligesom diskrete fordelinger mest naturligt angives ved deres sandsynlighedsfunktioner. Fordelingsfunktionernes fordel er, at de bringer kontinuerte og diskrete fordelinger på den reelle akse ind i samme begrebsramme. Således har definitionen

$$F(x) = P(X \leq x)$$

jo mening, uanset om X har en diskret eller en kontinuert fordeling. Forskellen består i, at medens fordelingsfunktionen for en kontinuert fordeling er kontinuert overalt (og differentiabel næsten overalt), så er fordelingsfunktionen for en diskret fordeling (f. eks. en binomialfordeling) en stykkevis konstant funktion, som vokser ved spring i de punkter, hvor sandsynlighedsmassen er placeret. Med udgangspunkt i begrebet fordelingsfunktion kan man give en generel definition af sandsynlighedsfordelinger på $\mathbf{R}$, som ud over diskrete og kontinuerte fordelinger omfatter en hel del andre, herunder "hybrider" af diskrete og kontinuerte fordelinger, samt nogle ret sygelige såkaldte "singulære fordelinger" der udmærker sig ved hverken at kunne beskrives ved punktsandsynligheder eller tætheder, og som i øvrigt er uden praktisk interesse.

Eksempel 5.3.4. Lad X være ligefordelt på $[-1,1]$. Fordelingen af $Y = X \vee 0$ er et eksempel på en "hybrid" som omtalt ovenfor. Halvdelen af sandsynlighedsmassen i denne fordeling er koncentreret i punktet 0, og den anden halvdel er jævnt fordelt over enhedsintervallet. Man kan naturligvis hverken opskrive en sandsynlighedsfunktion eller en tæthed i dette tilfælde, men fordelingsfunktionen er veldefineret:

$$F(y) = \begin{cases} 0 & \text{for } y < 0, \\ \frac{1}{2} + \frac{1}{2}y & \text{for } 0 \leq y \leq 1, \\ 1 & \text{for } 1 < y. \end{cases}$$



Fordelingsfunktionen for Y i eksempel 5.3.4

Fordelinger af denne art vil vi ikke beskæftige os med i det følgende. Her nævnes de blot for at pointere, at man ikke får fat i hele sandheden om sandsynlighedsfordelingers “natur” ved at supplere de diskrete fordelinger med de kontinuerte. Angående den udvidelse, man opnår ved at tage udgangspunkt i fordelingsfunktionerne, skal siges, at den heller ikke fører langt. Det er kun på den reelle akse man til nød kan holde ud at arbejde med fordelingsfunktioner, selv om der findes et lignende begreb på $\mathbf{R}^n$. Videregående sandsynlighedsregning tager udgangspunkt i abstrakt mål– og integralteori, som for det meste baseres på additive mængdefunktioner med passende kontinuitetsegenskaber.

5.4. Transformation af kontinuerte fordelinger.

Eksempel 5.3.4 viser, at selv en simpel transformation af en variabel med kontinuert fordeling kan føre os ud af klassen af sådanne fordelinger. Der er således tale om en mere kompliceret situation end den vi oplevede i §1.4 (med de udvidelser, der blev foretaget i §3.5), hvor en vilkårlig transformation $t\colon E \to F$ definerede en afledt stokastisk variabel $Y = t(X)$, hvis fordeling ligeledes blev diskret.

Der er dog to hovedtilfælde, hvor fordelingen af den afledte variabel $Y = t(X)$ umiddelbart kan tillægges en mening:

(1) $t\colon \mathbf{R} \to E$ er stykkevis konstant (E en vilkårlig mængde).

(2) $t\colon \mathbf{R} \to \mathbf{R}$ er monoton og kontinuert differentiabel med $t'(x) \neq 0$.

I begge tilfælde antages her, at X er en stokastisk variabel på $\mathbf{R}$. Men alt hvad der siges i det følgende gælder næsten uden modifikationer, når $\mathbf{R}$ erstattes med et interval på $\mathbf{R}$.

Tilfælde (1). Lad $t\colon \mathbf{R} \to E$ (= en vilkårlig mængde) være stykkevis konstant, f.eks. givet på formen

$$t(x) = y_i \text{ for } x \in]a_i, a_{i+1}]$$

for passende delepunkter $\cdots < a_i < a_{i+1} < \ldots$ og billedpunkter $y_i \in E$. Det er da naturligt at opfatte $Y = t(X)$ som en variabel med diskret fordeling, givet ved punktsandsynlighederne

$$P(Y = y) = P\left(X \in t^{-1}(y)\right) = \int 1_{t^{-1}(y)}(x)p(x)\,dx.$$

Her vil $t^{-1}(y)$ være et af intervallerne $]a_i, a_{i+1}]$ eller en foreningsmængde af nogle af disse, så integralet er veldefineret. Bemærk, at det er ligegyldigt hvorledes t defineres i selve delepunkterne a_i. Vi kunne f.eks. lige så godt have defineret

$$t(x) = y_i \text{ for } x \in [a_i, a_{i+1}[$$

fordi den kontinuert fordelte variabel X antager enhver af værdierne a_i med sandsynlighed 0.

Eksempel 5.4.1. Lad X være ligefordelt på enhedsintervallet, og definér $t\colon [0,1[\to \{1,2,\ldots,N\}$ ved

$$t(x) = [Nx] + 1.$$

Så bliver $Y = t(X)$ ligefordelt på $\{1,2,\ldots,N\}$ (jvf. eksempel 5.1.1).

Eksempel 5.4.2. Lad $X \in [0,+\infty[$ være normeret eksponentialfordelt og definér $t\colon [0,+\infty[\to \mathbf{N}_0$ ved

$$t(x) = \left[\frac{x}{h}\right]$$

for et givet tal $h > 0$. Fordelingen af $Y = t(X)$ er da givet ved

$$\begin{aligned}
P(Y = y) &= P([X/h] = y)\\
&= P(y \le X/h < y+1)\\
&= P(hy \le X < h(y+1))\\
&= \int_{hy}^{h(y+1)} e^{-x}\,dx\\
&= e^{-hy} - e^{-h(y+1)}\\
&= (1 - e^{-h})(e^{-h})^y,
\end{aligned}$$

dvs. Y er geometrisk fordelt med parameter $p = 1 - e^{-h}$ ($\approx h$ for h lille). Bemærk overensstemmelsen mellem dette resultat og den geometriske fordelings rolle som fordeling af en “afrundet ventetid”. I eksempel 5.1.2 bemærkede vi f.eks. at ventetiden til første punktering for vores cyklist C ville være fordelt som $X_0 \times 333.33$ for X_0 normeret eksponentialfordelt. Af overvejelserne i eksempel 3.6.2, i forbindelse med definitionen af den geometriske fordeling (eksempel 3.5.1) følger, at antallet af hele km–strækninger før første punktering er geometrisk fordelt med parameter $p =$ ca. 0.003.

Tilfælde (2). Betragt nu situationen hvor $t\colon \mathbf{R} \to \mathbf{R}$ er en bijektiv, kontinuert afbildning. Antag først at t er strengt voksende og kontinuert differentiabel med differentialkvotient $t'(x) = \frac{d}{dx}t(x) > 0$ for alle x. For et punkt $x \in \mathbf{R}$ er sandsynlighedsmassen i intervallet $[x, x+h]$ approksimativt $p(x)h$. Dette interval føres af t over i intervallet fra $y = t(x)$ til $t(x+h) \approx y + h \times t'(x)$. Der gælder derfor

$$P(Y \in [y, y + h \times t'(x)]) \approx P(X \in [x, x+h]) \approx hp(x),$$

hvilket (ifølge vor oprindelige fortolkning af tætheden som det lokale forhold mellem sandsynlighed og intervallængde) betyder at Y's fordeling har tæthed

$$q(y) = \frac{p(x)}{t'(x)} = \frac{p\left(t^{-1}(y)\right)}{t'\left(t^{-1}(y)\right)}$$

i punktet y. For $t'(x) < 0$ (svarende til at t er strengt aftagende) fås et tilsvarende resultat, idet nævneren i udtrykket for $q(y)$ dog skifter fortegn, fordi t bytter om på nedre og øvre endepunkt for det lille interval. Hermed har vi givet et heuristisk bevis for følgende resultat (som vi, for en ordens skyld, formulerer således at variationsområderne for X og Y kan være vilkårlige intervaller):

SÆTNING 5.4.1. *Lad E og F være intervaller på $\mathbf{R}$, $t\colon E \to F$ en bijektiv, monotont voksende eller aftagende, kontinuert differentiabel afbildning med $t'(x) \neq 0$ for alle x. Lad p være en sandsynlighedstæthed på E, og definér $q\colon F \to \mathbf{R}$ ved*

$$q(y) = \frac{p\left(t^{-1}(y)\right)}{\left|t'\left(t^{-1}(y)\right)\right|}.$$

Da er q en sandsynlighedstæthed på F, og for enhver pæn mængde $B \subseteq F$ (dvs. et interval eller en endelig foreningsmængde af intervaller) gælder

$$\int 1_{t^{-1}(B)}(x)p(x)\,dx = \int 1_B(y)q(y)\,dy.$$

BEVIS. Vi nøjes med at se på tilfældet hvor t er voksende, idet beviset forløber helt analogt for t aftagende. Vi vil endvidere indskrænke os til tilfældet hvor B er et interval. For endelige foreningsmængder af intervaller følger resultatet let ved addition af tilsvarende relationer for disjunkte intervaller.

Sæt $B = [a, y]$ $(y \geq a)$, og betragt (for fast a) de to funktioner

$$f_1(y) = \int_a^y q(y')\,dy',$$

$$f_2(y) = \int_{t^{-1}(a)}^{t^{-1}(y)} p(x)\,dx.$$

Det vi skal vise er præcis at $f_1(y) = f_2(y)$. Da $f_1(a) = f_2(a)$ følger dette, hvis vi kan vise at disse to funktioner er differentiable med samme differentialkvotient. For f_1's vedkommende er

$$\frac{d}{dy} f_1(y) = q(y),$$

og for f_2 fås v.h.a. reglen for differentiation af en sammensat funktion og reglen for differentiation af omvendt funktion

$$\frac{d}{dy} f_2(y) = p\left(t^{-1}(y)\right) \frac{d}{dy} t^{-1}(y) = p\left(t^{-1}(y)\right) \frac{1}{t'\left(t^{-1}(y)\right)} = q(y).$$

Heraf følger sætningens sidste påstand i tilfældet hvor B er et begrænset interval. Man udvider let til tilfældet, hvor B er en endelig foreningsmængde af begrænsede intervaller, og en simpel grænseovergang viser at resultatet også gælder hvis B er et ubegrænset interval. Specielt fås for $B = F$ at $\int_F q(y)\,dy = \int_E p(x)\,dx = 1$, dvs. q er en sandsynlighedstæthed.

Sætningen viser, at den eneste rimelige definition af fordelingen af Y i dette tilfælde går ud på, at Y følger den kontinuerte fordeling givet ved tætheden q. Kun herved får vi jo opfyldt den naturlige relation

$$P(Y \in B) = P\left(X \in t^{-1}(B)\right).$$

Eksempel 5.4.3 (Positions- og skalaparameter). Hvis X har tæthed p vil $Y = a + bX$ åbenbart have tæthed

$$q(y) = \frac{1}{|b|} p\left(\frac{y-a}{b}\right),$$

forudsat at $b \neq 0$. Fordelingen af Y siges her at være fremkommet af X's fordeling ved at denne er forsynet med *positionsparameteren* a og *skalaparameteren* b. Fordelingerne af X og Y siges så at være af samme *type*. Når man taler om positions- og skalaparameter for en fordeling er det oftest med reference til et (underforstået) valg af en *normeret* fordeling af typen. Hvis f.eks. Y er eksponentialfordelt med positionsparameter a og skalaparameter b betyder det, at Y har tæthed

$$q(y) = \frac{1}{|b|} e^{-(y-a)/b} 1_{\{(y-a)/b>0\}},$$

med reference til den *normerede* eksponentialfordeling, der har tæthed e^{-x} på den positive halvakse.

Eksempel 5.4.4. Lad X være normeret eksponentialfordelt, og betragt den afledte stokastiske variabel $Y = X^2$. Da afbildningen $t(x) = x^2$ har differentialkvotient $t'(x) = 2x$, får Y ifølge sætningen tætheden

$$q(y) = \frac{p\left(t^{-1}(y)\right)}{\left|t'\left(t^{-1}(y)\right)\right|} = \frac{e^{-\sqrt{y}}}{2\sqrt{y}}$$

på intervallet $[0, +\infty[$. Bemærk at sætningens antagelser ikke helt er opfyldt, fordi $t'(0) = 0$. Det fører til at tætheden q for Y bliver udefineret eller $+\infty$ i punktet 0. Men det er klart, at q er en tæthed under de udvidelser af integralbegrebet vi har givet i §5.2. Overvejelser af denne art har vi valgt at gå let hen over (også i selve beviset for sætning 5.4.1, som den kritiske læser måske har bemærket).

Eksempel 5.4.5. Antag $p(x) > 0$ for alle $x \in E$ (= et interval på $\mathbf{R}$), og betragt transformationen $t\colon E \to [0, 1]$ givet ved

$$t(x) = F(x) = \int_{-\infty}^{x} p(x)\, dx$$

(hvor den nedre grænse $-\infty$ for integralet strengt taget skal erstattes med E's nedre endepunkt, hvis E er nedad begrænset). Så er $t'(x) = p(x)$, hvoraf følger at tætheden q for $Y = t(X)$ bliver konstant lig med 1. Hermed har vi indset, at hvis X er en (kontinuert fordelt) reel stokastisk variabel med fordelingsfunktion F, så vil $F(X)$ være ligefordelt på enhedsintervallet. Det kan man i øvrigt også let indse direkte ud fra fordelingsfunktionens definition. Omvendt gælder det, at en stokastisk variabel med fordelingsfunktion F kan konstrueres ud fra en ligefordelt variabel Y på enhedsintervallet ved

$$X = F^{-1}(Y)$$

(bemærk: F^{-1} er veldefineret i hvert fald på $]0, 1[$, da F er kontinuert og strengt voksende). Dette udnyttes ofte ved simulation, hvor de algoritmer man har til generering af (pseudo–) ligefordelte variable hermed kan benyttes til generering af variable med en given kontinuert fordeling, forudsat at man har en procedure til beregning af funktionen F^{-1}. Normerede eksponentialfordelte variable $X_1, X_2, \ldots$ kan for eksempel konstrueres ud fra ligefordelte variable $R_1, R_2, \ldots$ efter opskriften

$$X_i = -\log\left(R_i\right),$$

jvf. opgave 5.1.1.

Sætning 5.4.1 udtaler sig kun om tilfældet, hvor transformationen t er bijektiv. I virkeligheden gælder et tilsvarende resultat for tilfældet, hvor variationsområdet for X kan deles op i intervaller, på hvilke t er monoton og i øvrigt opfylder sætningens betingelser. I praksis betyder dette, at fordelingen af en afledt variabel $Y = t(X)$ altid kan tillægges en kontinuert fordeling, med mindre t netop er konstant på et interval (hvilket, som vi har set, fører til koncentration af positiv sandsynlighedsmasse i et enkelt billedpunkt). Tætheden for Y bliver sum af bidrag af samme form som i sætning 5.4.1, idet man blot skal summere over de x der afbildes over i det pågældende y:

$$q(y) = \sum_{x \in t^{-1}(y)} \frac{p(x)}{|t'(x)|}.$$

Eksempel 5.4.6 (Arcus–sinus fordelingen). Lad X være ligefordelt på $[0, 2\pi[$, og sæt $Y = \sin(X) \in [-1, 1]$. Her svarer hvert y til to mulige værdier af x (bortset fra $y = \pm 1$). Differentialkvotienten $t'(x)$ har samme absolutte værdi i de to punkter svarende til et givet y, idet der jo for $y = \sin(x)$ gælder

$$|t'(x)| = |\frac{d}{dx} \sin(x)| = |\cos(x)| = \sqrt{1 - \sin(x)^2} = \sqrt{1 - y^2}.$$

Tætheden for Y's fordeling bliver derfor

$$q(y) = \sum_{x : \sin(x) = y} \frac{p(x)}{|t'(x)|} = 2 \frac{1/2\pi}{\sqrt{1 - y^2}} = \frac{1}{\pi \sqrt{1 - y^2}}.$$

Fordelingen kaldes *arcus–sinus fordelingen* efter sin fordelingsfunktion, som ses at være $F(y) = \frac{1}{2} + \frac{1}{\pi} \sin^{-1}(y)$ (hvor $\sin^{-1}$ eller arcsin betegner den inverse funktion til $\sin \colon [-\frac{\pi}{2}, \frac{\pi}{2}] \to [-1, 1]$).

Opgave 5.4.1. Lad X være en reel stokastisk variabel med tæthed p. Opskriv tætheden for fordelingen af

(a) $\sqrt{X}$ (forudsat $P(X \geq 0) = 1$)

(b) $1/X$

(c) X^2

(d) $\exp(X)$

Opgave 5.4.2*. Lad X være en stokastisk variabel med tæthed p, og lad A være en pæn mængde med $P(X \in A) > 0$ (i det følgende kan man tænke på et interval, hvis det gør tingene lettere). Den *betingede fordeling* af X, givet $X \in A$, kan defineres som fordelingen med tæthed

$$p(x|A) = \begin{cases} \frac{p(x)}{P(X \in A)} & \text{for } x \in A, \\ 0 & \text{ellers.} \end{cases}$$

Vis at $p(\cdot|A)$ faktisk er en sandsynlighedstæthed, og gør rede for, at den tilsvarende fordeling $P(\cdot|A)$ er givet ved

$$P(B|A) = \frac{P(X \in A \cap B)}{P(X \in A)}.$$

Hvad bliver fordelingsfunktionen for den betingede fordeling, udtrykt ved fordelingsfunktionen for den oprindelige fordeling? (her bliver resultatet kun til at holde ud at se på i tilfældet hvor A er et interval).

OPGAVE 5.4.3. I opgave 3.6.1 (A venter på en taxa ...) vil ventetiden X til første taxaankomst naturligvis være eksponentialfordelt med skalaparameter 30 min. Gør rede for følgende nedslående udsagn: Den betingede fordeling (jvf. opgave 5.4.2) af restventetiden $X - x$, givet at A uden held har ventet i x minutter, er præcis den samme som den oprindelige (ubetingede) fordeling af X.

OPGAVE 5.4.4*. (Cauchy fordelingen). Lad X være ligefordelt på intervallet $]-\frac{\pi}{2}, \frac{\pi}{2}[$, og sæt $Y = \tan(X)$. Vis at fordelingen af Y er givet ved tætheden

$$q(y) = \frac{1}{\pi(1+y^2)}.$$

Denne fordeling kaldes en (normeret) *Cauchy fordeling.*

(a) Hvad er fordelingsfunktionen for denne fordeling?

(b) Hvis X er normeret Cauchy–fordelt, hvad er fordelingen af $|X|$? og af X^2? og $1/X$?

OPGAVE 5.4.5*. (fraktiler). Lad X følge en kontinuert fordeling (på $\mathbf{R}$ eller et interval $E \subseteq \mathbf{R}$) med strengt positiv tæthed p. For $a \in]0,1[$ defineres fordelingens *a–fraktil* som den værdi af x for hvilken der gælder $P(X \leq x) = a$. Her angives a normalt i procent, således at f.eks. 50%–fraktilen (som også kaldes *medianen*) er det tal x, der giver $P(X \leq x) = \frac{1}{2}$.

(a) Angiv 5%–fraktil, median og 95%–fraktil i den normerede eksponentialfordeling.

(b) Angiv 5%–fraktil, median og 95%–fraktil i den normerede Cauchy fordeling (opgave 5.4.4).

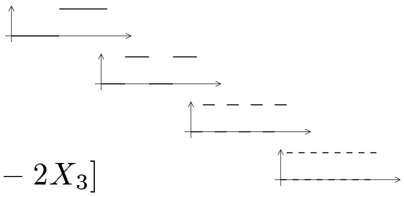
(c) Hvilke regler gælder for fraktiler i en transformeret fordeling? Specielt, hvordan afhænger fraktilerne i en fordeling med positions– og skalaparameter af disse parametre?

OPGAVE 5.4.6. Lad X følge en fordeling med tæthed p på $\mathbf{R}$. Hvad er tætheden for fordelingen af $|X|$?

OPGAVE 5.4.7. C stiller sin cykel fra sig efter en længere tur. Beskriv fordelingen af baghjulsventilens højde over jorden.

OPGAVE 5.4.8. (Simulation af et vilkårligt antal møntkast ud fra én rektangulært fordelt variabel). Med $[x]$ betegnes som sædvanligt *heldelen* af det reelle tal x, dvs. det største hele tal som er $\leq x$. Lad X være rektangulært fordelt på enhedsintervallet, og definer diskrete stokastiske variable $X_1, X_2, \ldots, X_n$ med værdier i $\{0,1\}$ ved

$$\begin{aligned} X_1 &= [2X] \\ X_2 &= [4X - 2X_1] \\ X_3 &= [8X - 4X_1 - 2X_2] \\ X_4 &= [16X - 8X_1 - 4X_2 - 2X_3] \\ &\vdots \end{aligned}$$



Vis at $X_1, X_2, \ldots, X_n$ er uafhængige, ligefordelte på $\{0,1\}$ (Vink: Bemærk at der i totalssystemet gælder $X = 0.X_1X_2X_3\ldots X_n$, på nær afrunding nedad til n betydende cifre).

5.5. Middelværdi og varians.

DEFINITION. Lad $X \in \mathbf{R}$ følge en kontinuert fordeling med tæthed p. Ved *middelværdien* af X forstås da størrelsen

$$\mathrm{E}X = \int p(x)x\,dx.$$

Definitionen forudsætter naturligvis at integralet er defineret, dvs. (jvf. §5.2)

$$\int p(x)|x|\,dx < +\infty.$$

Hvis denne betingelse ikke er opfyldt tillægges X ikke nogen middelværdi. Dog vil man ofte, i tilfælde af en fordeling på den positive halvakse, skrive $\mathrm{E}X = +\infty$ hvis $\int p(x)|x|\,dx = \int p(x)x\,dx = +\infty$.

EKSEMPEL 5.5.1. Lad X være ligefordelt på $[a,b]$. Da er (ikke forbavsende)

$$\mathrm{E}X = \frac{1}{b-a}\int_a^b x\,dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}.$$

EKSEMPEL 5.5.2. Middelværdien af en normeret eksponentialfordelt variabel X kan udregnes ved delvis integration:

$$\mathrm{E}X = \int_0^{+\infty} e^{-x}x\,dx = \left[\left(-e^{-x}\right)x\right]_0^{+\infty} - \int_0^{+\infty}\left(-e^{-x}\right)\,dx = 1.$$

Eksempel 5.5.3. Middelværdien i Cauchy fordelingen (opgave 5.4.4) er ikke defineret, idet der gælder

$$\int \frac{1}{\pi(1+x^2)}|x|\,dx = +\infty.$$

Dette følger f.eks. af at

$$\int_{-n}^{n} \frac{1}{\pi(1+x^2)}|x|\,dx \geq \frac{1}{\pi}\int_{1}^{n} \frac{|x|}{1+x^2}\,dx$$

$$\geq \frac{1}{\pi}\int_{1}^{n} \frac{|x|}{2x^2}\,dx = \frac{1}{2\pi}\int_{1}^{n} \frac{1}{|x|}\,dx = \frac{1}{2\pi}\log n.$$

De i §4.2 udledte regneregler for middelværdier (sætning 4.2.2 (a), (b) og (c), samt fortolkning af sandsynligheden for en hændelse som middelværdien af dens indikator) gælder uændret i det kontinuerte tilfælde. Men nogle af dem har ingen interesse i øjeblikket, fordi konstruktionen af flere kontinuerte variable, afledt af en fælles baggrundsvariabel, normalt forudsætter en flerdimensional baggrundsvariabel med en kontinuert fordeling (som vil blive indført i kapitel 6). Vi nøjes med at bemærke, at regnereglen

$$\mathrm{E}(a+bX) = a + b\mathrm{E}X$$

her kan opfattes som en formel til udregning af middelværdien i en fordeling med positions– og skalaparameter. Den kan bevises ved direkte udregning, baseret på tætheden for fordelingen af $a+bX$ (se eksempel 5.4.3), men den er også en konsekvens af det generelle princip, at middelværdien af en afledt stokastisk variabel kan beregnes enten i sin egen fordeling eller i fordelingen af den variabel, som den er funktion af:

Sætning 5.5.1. *Lad X have tætheden p, og lad $Y = t(X)$ være en afledt stokastisk variabel. Da er $\mathrm{E}|Y| < +\infty$ hvis og kun hvis*

$$\int p(x)|t(x)|\,dx < +\infty,$$

og i så fald er

$$\mathrm{E}Y = \int p(x)t(x)\,dx.$$

Her forudsættes naturligvis, at vi er i en situation hvor det har mening at tale om Y som en stokastisk variabel, svarende til et af de to hovedtilfælde som er behandlet i §5.4. Vi vil dog nøjes med at bevise sætningen i tilfælde (2), hvor t antages monoton og kontinuert differentiabel og Y igen får en kontinuert fordeling. Beviset i tilfælde (1), hvor t er stykkevis konstant og fordelingen af Y bliver diskret, overlades til læseren.

I tilfælde (2) vil vi endda nøjes med at skitsere et bevis: Rent formelt er sætningen en konsekvens af reglen for integration ved substitution. Hvis t antages voksende er

$$\begin{aligned}\int q(y)y\,dy &= \int \frac{p\left(t^{-1}(y)\right)}{t'\left(t^{-1}(y)\right)} y\,dy \\ &= \int \frac{p(x)}{t'(x)} t(x)\,dt(x) \\ &= \int \frac{p(x)}{t'(x)} t(x)t'(x)\,dx \\ &= \int p(x)t(x)\,dx.\end{aligned}$$

Denne omskrivning er i første omgang kun gyldig når p er kontinuert på et begrænset interval $[a, b]$, og t er voksende med $t'(x) > 0$. Men det er let at udvide formlen til de situationer, som vi berørte i §5.2. Her må man blot sikre sig eksistens af disse "udvidede" integraler ved at gennemføre samme udregning med $|y|$ i stedet for y. Og mere vil vi ikke gøre ud af det.

Eksempel 5.5.4. En normeret eksponentialfordelt variabel X kan tænkes fremkommet af en ligefordelt variabel R på $[0, 1]$ ved

$$X = -\log(R).$$

Det er derfor også muligt at udregne middelværdien i eksponentialfordelingen som

$$\mathrm{E}X = \int_0^1 -\log(r)\,dr \qquad (= 1).$$

Varians og standardafvigelse kan defineres præcis som i §4.3. Kovarians og korrelation må vente til kapitel 6, fordi disse begreber igen forudsætter at vi kan tale om flere kontinuerte variable, afledt af samme baggrundsvariabel. Alt hvad der er sagt om varians og standardafvigelse kan (for så vidt det ikke vedrører flere variable) overføres uændret. For eksempel er relationen

$$\mathrm{var}(a + bX) = b^2\mathrm{var}(X)$$

gyldig, og vi kan her tænke på den som en formel til beregning af variansen i en fordeling med positions- og skalaparameter.

Eksempel 5.5.5. For en normeret eksponentialfordelt variabel X er

$$\begin{aligned}\mathrm{E}(X^2) &= \int_0^{+\infty} e^{-x}x^2\,dx \\ &= \left[(-e^{-x})x^2\right]_0^{+\infty} - \int_0^{+\infty} (-e^{-x})2x\,dx = 0 + 2\mathrm{E}X = 2,\end{aligned}$$

hvoraf følger at

$$\mathrm{var}(X) = \mathrm{E}(X^2) - (\mathrm{E}X)^2 = 2 - 1 = 1.$$

Variansen i en eksponentialfordeling med skalaparameter b er således b^2.

Eksempel 5.5.6. Variansen i den "normerede rektangulære fordeling", dvs. ligefordelingen på enhedsintervallet, er

$$\int_0^1 \left(x - \frac{1}{2}\right)^2 dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} x^2\, dx = 2\int_0^{\frac{1}{2}} x^2\, dx = 2\frac{(1/2)^3}{3} = \frac{1}{12}.$$

Standardafvigelsen for en ligefordeling på intervallet $[a, b]$ er således $(b - a)/\sqrt{12}$ (idet denne fordeling kan opfattes som en rektangulær fordeling med skalaparameter $b - a$ og positionsparameter a, når den normerede rektangulære fordeling defineres som ovenfor).

Opgave 5.5.1. Beregn middelværdi og varians i arcus–sinus fordelingen (se eksempel 5.4.6). Tegn fordelingens tæthed. Sammenlign variansen med variansen for en rektangulær fordeling på $[-1, 1]$.

Opgave 5.5.2. Den *tosidede eksponentialfordeling* har tætheden

$$p(x) = \frac{1}{2}e^{-|x|}.$$

Tegn tæthed og fordelingsfunktion, og beregn fordelingens middelværdi og varians.

Opgave 5.5.3. Lad X være normeret exponentialfordelt. Hvad er

(a) $\mathrm{E}X^3$?

(b) $\mathrm{E}\exp(-aX)$?

Opgave 5.5.4. (Γ–fordelingen med heltallig formparameter, jvf. kapitel 7). Sæt

$$I_n = \int_0^{+\infty} x^n e^{-x}\, dx.$$

(a) Vis at $I_n = n!$ (Vink: Vis at $I_n = nI_{n-1}$).

(b) Funktionen $p_n\colon [0, +\infty[\to \mathbf{R}$ givet ved $p_n(x) = \frac{1}{n!}x^n e^{-x}$ er således en sandsynlighedstæthed. Udregn middelværdi, varians og standardafvigelse i den tilsvarende fordeling.

5.6. Den normale fordeling.

Betragt funktionen

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

Det er klart, at denne funktion har endeligt integral, da $e^{-x^2/2} \to 0$ for $x \to \pm\infty$ hurtigere end f.eks. $e^{-|x|}$. Knapt så indlysende er det, at integralet netop er 1, altså at

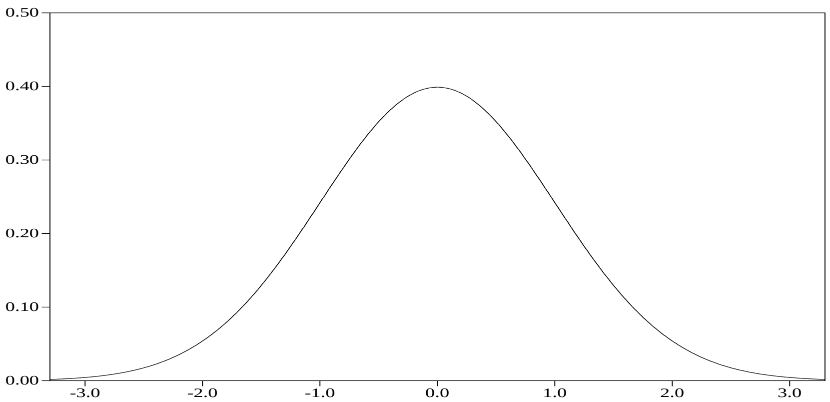
$$\int_{-\infty}^{+\infty} e^{-x^2/2}\, dx = \sqrt{2\pi}.$$

Det vil blive bevist i kapitel 6. Fordelingen med tæthed φ kaldes den (normerede) *normale fordeling* eller *Gauss fordelingen*. Den tilsvarende fordelingsfunktion betegnes

$$\Phi(x) = \int_{-\infty}^{x} \varphi(x')\, dx'.$$

Den normale fordeling med positionsparameter μ og skalaparameter $\sigma > 0$ har åbenbart tætheden

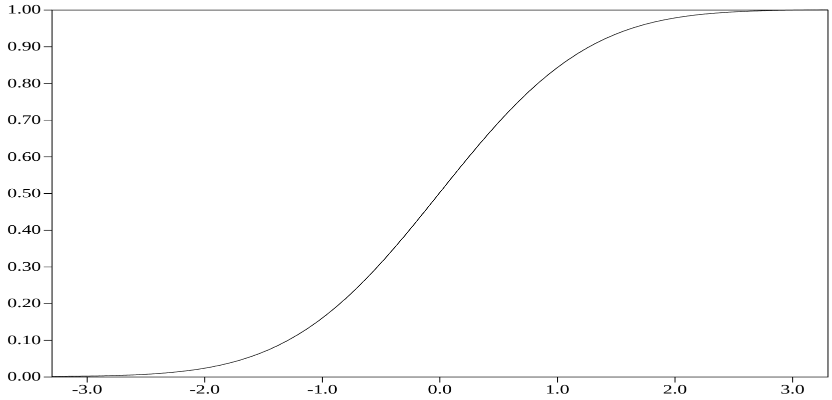
$$\frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$



Den normale fordelings tæthed

Middelværdi og varians i normalfordelingen ses let at være veldefinerede. Faktisk gælder der

$$\frac{1}{\sqrt{2\pi}} \int |x|^n e^{-x^2/2}\, dx < +\infty$$



Den normale fordelings fordelingsfunktion

for ethvert $n \in \mathbf{N}$. For en normeret normalt fordelt stokastisk variabel X er

$$\mathrm{E}X = \frac{1}{\sqrt{2\pi}} \int x e^{-x^2/2}\, dx = 0$$

(fordi integranden er en ulige funktion). Variansen kan udregnes på følgende måde. Differentiation af tætheden φ giver

$$\varphi'(x) = \frac{d}{dx}\left(\frac{1}{\sqrt{2\pi}} e^{-x^2/2}\right) = \frac{1}{\sqrt{2\pi}}(-x)e^{-x^2/2},$$

og differentiation endnu en gang giver

$$\varphi''(x) = \frac{d}{dx}\left(\frac{1}{\sqrt{2\pi}}(-x) e^{-x^2/2}\right) = \frac{1}{\sqrt{2\pi}}(x^2-1)e^{-x^2/2}.$$

Nu er

$$\int_{-\infty}^{+\infty} \varphi'(x)\, dx = \lim_{n\to\infty} \int_{-n}^{+n} \varphi'(x)\, dx = \lim_{n\to\infty} \left(\varphi(n) - \varphi(-n)\right) = 0.$$

Tilsvarende fås, da $\varphi'(n) \to 0$ for $n \to \pm\infty$,

$$\int_{-\infty}^{+\infty} \varphi''(x)\, dx = \lim_{n\to\infty} \int_{-n}^{+n} \varphi''(x)\, dx = \lim_{n\to\infty} \left(\varphi'(n) - \varphi'(-n)\right) = 0.$$

Men det betyder jo at

$$\frac{1}{\sqrt{2\pi}} \int (1-x^2)e^{-x^2/2}\, dx = 0,$$

hvilket er det samme som $1 - \mathrm{E}X^2 = 0$, altså $\mathrm{var}(X) = \mathrm{E}X^2 = 1$. Den normale fordeling med positionsparameter μ og skalaparameter σ har således middelværdi μ og varians σ^2.

Den normale fordeling er langt den vigtigste af de kontinuerte fordelinger. Traditionelt opfattes den som den naturlige (eller "normale") fordeling af målefejl o.l., hvilket man begrunder med *den centrale grænseværdisætning* (som følger nedenfor). Det er også den centrale grænseværdisætning som er skyld i, at normalfordelingen hyppigt optræder som approksimativ fordeling af estimatorer for parametre i statistiske modeller. Dertil kommer, at visse statistiske modeller, baseret på normalfordelingen, har en særligt simpel struktur, som har relation til n-dimensional Euklidisk geometri og lineær algebra. Disse *lineære normalfordelingsmodeller* er de vigtigste og mest anvendte statistiske modeller overhovedet, især hvis man medregner de utallige modifikationer og generaliseringer af dem.

Den centrale grænseværdisætning går groft sagt ud på, at hvis man danner summen af et stort antal uafhængige stokastiske variable, som hver for sig er små i forhold til summen, så vil summens fordeling være approksimativt normal. Som vi så i §4.5 vil gennemsnittet af mange (identisk fordelte) uafhængige variable følge en fordeling, der koncentrerer sig mere og mere snævert omkring middelværdien. Den centrale grænseværdisætning handler altså om, hvad der mere detaljeret går for sig i nærheden af dette "centrum", når man forstørrer billedet så det er til at se hvad der sker. Det kan vel ikke overraske, at den "forstørrelse" der skal til, går ud på at normere fordelingen af summen (eller gennemsnittet) så den får middelværdi 0 og varians 1. Den simpleste version af den centrale grænseværdisætning ser sådan ud:

SÆTNING 5.6.1. *Lad $X_1, X_2, \ldots$ være uafhængige, identisk fordelte stokastiske variable med middelværdi μ og varians σ^2. Da vil fordelingen af*

$$U_n = \frac{\sqrt{n}}{\sigma}\left(\frac{X_1 + \cdots + X_n}{n} - \mu\right)$$

konvergere mod en normeret normalfordeling, i den forstand at

$$P(U_n \leq u) \to \Phi(u).$$

Sætningen gælder uanset om den fælles fordeling af X'erne er diskret eller kontinuert. I det kontinuerte tilfælde har vi endnu ikke præciseret, hvad vi mener med "uafhængige variable" (det kommer i kapitel 6); men det gør ikke så meget, da vi ikke vil forsøge at bevise sætningen, mindst af alt i det kontinuerte tilfælde. Beviset for den centrale grænseværdisætning kræver et matematisk–analystisk apparatur, som vi slet ikke har til rådighed her. Selv den simpleste, klassiske version, *de Moivre's sætning*, som siger at binomialfordelingen konvergerer i form mod normalfordelingen for $n \to \infty$ (p fast), er svær at bevise. Vi vil indskrænke os til at give et (meget) heuristisk bevis for følgende version af denne sætning:

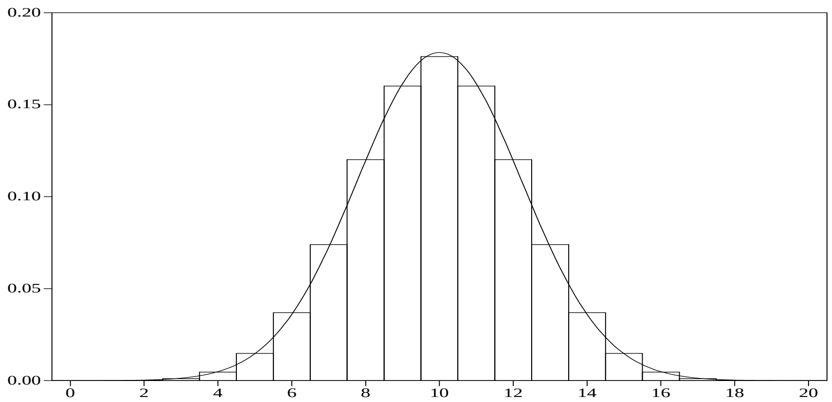
SÆTNING 5.6.2. *Lad (x_n) være en følge af hele tal, $0 \le x_n \le n$, således at (for $n \to \infty$)*

$$\frac{x_n - np}{\sqrt{npq}} \to u$$

hvor $u \in \mathbf{R}$, $p \in]0,1[$ og $q = 1 - p$ er givne tal. Da vil

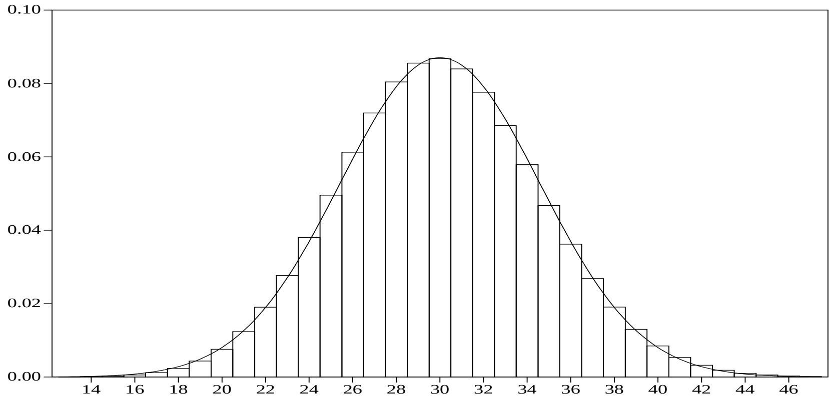
$$\sqrt{npq}\binom{n}{x_n} p^{x_n} q^{n-x_n} \to \varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}.$$

Bemærkninger. Hvis X_n er binomialfordelt (n,p) vil $\frac{X_n - np}{\sqrt{npq}}$ have middelværdi 0 og varians 1. Venstre side i sætningens sidste linie kan derfor, bortset fra normeringsfaktoren $\sqrt{npq}$, fortolkes som en punktsandsynlighed i binomialfordelingen i et punkt x_n der (mere og mere præcist for $n \to \infty$) er placeret u standardafvigelser fra middelværdien np. Vi kunne have valgt $x_n = [u\sqrt{npq} + np]$. Normeringsfaktoren $\sqrt{npq}$ skyldes omskaleringen. En punktsandsynlighed kan jo ikke direkte sammenlignes med værdien af en tæthed, fordi tæthed angiver sandsynlighed pr. intervallængde, jvf. "approksimationen" af en rektangulær fordeling med diskrete fordelinger i eksempel 5.1.1, hvor punktsandsynligheder $1/N$ i punkter med indbyrdes afstand $1/N$ svarede til tætheden 1. På samme måde skal punktsandsynligheden her divideres med afstanden $\frac{1}{\sqrt{npq}}$ mellem punkterne i den "standardiserede binomialfordeling" for at svare til en tæthed.



Binomialfordeling for n=20, p=0.5, samt approksimerende normal tæthed

Et egentligt bevis for de Moivres sætning vil vi som sagt ikke give. Men vi vil give et heuristisk argument, der gør sætningen til en nogenlunde troværdig påstand, og som forklarer, hvorfor det netop er funktionen $\exp(-x^2/2)$, der beskriver binomialfordelingens asymptotiske udseende. Som det ses af ovenstående tegning og tegningen på næste side er denne approksimation faktisk næsten så god som det overhovedet er muligt, når man skal approksimere en diskret fordeling med en kontinuert kurve.



Binomialfordelingen for n=100, p=0.3, samt approksimerende normal tæthed

Betragt de succesive forhold mellem binomialfordelingens punktsandsynligheder

$$\frac{\binom{n}{x+1}p^{x+1}q^{n-x-1}}{\binom{n}{x}p^x q^{n-x}} = \frac{n-x}{x+1} \times \frac{p}{q}.$$

For x nær ved np (vi kan f.eks. tænke på $|x-np| \leq 10\sqrt{npq}$, hvilket ifølge Chebychev's ulighed (sætning 4.5.1) gælder punkterne i mindst 99% af binomialfordelingens sandsynlighedsmasse) er logaritmen til dette forhold approksimativt

$$\begin{aligned}
\log\left(\frac{n-x}{x+1} \times \frac{p}{q}\right) &= \log\left(\frac{n-x}{nq}\right) - \log\left(\frac{x+1}{np}\right) \\
&= \log\left(\frac{nq-(x-np)}{nq}\right) - \log\left(\frac{np+(x+1-np)}{np}\right) \\
&= \log\left(1 - \frac{x-np}{nq}\right) - \log\left(1 + \frac{x+1-np}{np}\right) \\
&\approx -\frac{x-np}{nq} - \frac{x+1-np}{np} \\
&\approx -\left(\frac{1}{nq} + \frac{1}{np}\right)(x-np) \\
&= -\frac{x-np}{npq} = -\frac{u}{\sqrt{npq}},
\end{aligned}$$

hvor $u = \frac{x-np}{\sqrt{npq}}$ er den til x svarende "standardiserede variabel".

Lad nu φ_n betegne en funktion som (i det mindste approksimativt) følger binomialfordelingens sandsynlighedsfunktion. Vi opfatter φ_n som funktion af den standardiserede variabel u, så kravet til φ_n er at der (approksimativt) gælder

$$\varphi_n\left(\frac{x-np}{\sqrt{npq}}\right) = \binom{n}{x}p^x q^{n-x}$$

for $x = 0, 1, \ldots, n$. Det forudsættes i det følgende, at disse funktioner φ_n kan vælges tilpas glatte, uden at vi helt kan præcisere denne forudsætning. Den approksimation vi udledte af logaritmen til forholdet mellem succesive punktsandsynligheder i binomialfordelingen kan nu skrives

$$\log\left(\frac{\varphi_n\left(\frac{x+1-np}{\sqrt{npq}}\right)}{\varphi_n\left(\frac{x-np}{\sqrt{npq}}\right)}\right) \approx -\frac{u}{\sqrt{npq}}$$

eller

$$\frac{\log\varphi_n\left(u+\frac{1}{\sqrt{npq}}\right)-\log\varphi_n(u)}{1/\sqrt{npq}} \approx -u$$

hvilket er omtrent det samme som

$$\frac{d}{du}\log\varphi_n(u) \approx -u.$$

Heraf ses, at hvis følgen (φ_n) af funktioner konvergerer (på nær en multiplikativ konstant, som er uden betydning for $\frac{d}{du}\log\varphi_n(u)$) mod en grænsefunktion φ, så må denne grænsefunktion (igen under passende betingelser som tillader ombytning af differentiation og grænseovergang) opfylde betingelsen

$$\frac{d}{du}\log\varphi(u) = -u,$$

hvilket som bekendt er ækvivalent med at $\log\varphi(u) = \text{const } - u^2/2$, eller

$$\varphi(u) = \text{const} \times e^{-\frac{u^2}{2}}.$$

Samtidig har vi sandsynliggjort, at en sådan konvergens finder sted. Vi har jo på det nærmeste bevist, at $\frac{d}{du}\log\varphi_n(u)$ konvergerer mod $-u$, og heraf følger – igen via passende ombytninger af grænseovergange – at $c_n\varphi_n(u) \to \varphi(u)$ for passende normeringskonstanter c_n. At man netop skal vælge $c_n = \sqrt{npq}$ følger ikke af dette argument, men det har vi tidligere argumenteret for.

Opgave 5.6.1. Tabellér og tegn sandsynlighedsfunktionen for binomialfordelingen med parametre (10,0.5) sammen med den approksimerende normalfordelings tæthed for $x = 0, 1, \ldots, 10$.

Opgave 5.6.2. Tegn sandsynlighedsfunktionen for summen Y ved et slag med tre terninger (jvf. opgave 1.4.4) sammen med den approksimerende normalfordelingskurve $\frac{1}{\sqrt{8.75}}\varphi\left(\frac{x-10.5}{\sqrt{8.75}}\right)$ (og begrund valget af konstanter i dette udtryk).

OPGAVE 5.6.3. En mønt kastes 1000 gange. Hvad er (sådan cirka) sandsynligheden for at få mindst 550 "plat"?

OPGAVE 5.6.4*. (den logaritmiske normalfordeling). Lad X være normalfordelt med parametre (μ, σ^2). Fordelingen af $Y = \exp(X)$ kaldes den *logaritmiske normalfordeling* med parametre (μ, σ^2).

(a) Opskriv tætheden for fordelingen af Y.

(b) Gør rede for, at klassen af logaritmiske normalfordelinger er skalainvariant, i den forstand at hvis Y er logaritmisk normalfordelt så er også βY logaritmisk normalfordelt for $\beta > 0$. Opskriv den formel, som viser hvorledes den logaritmiske normalfordelings parametre ændrer sig ved en sådan skalatransformation.

(c) Vis, at middelværdien i den normerede logaritmiske normalfordeling $(\mu = 0, \sigma^2 = 1)$ er $\sqrt{e} = 1.6487$.