6. Kontinuerlige sannsynlighetsfordelinger.

6.1. Definisjoner og begreper.

Vi har tidligere sett på *diskrete stokastiske variabler*. Vi skal nå se på ***kontinuerlige stokastiske variabler***. Dette er stokastiske variabler som kan ha alle reelle tall (eller en delmengde av de reelle tallene) som verdier. Eksempler på slike kontinuerlige stokastiske variabler kan være høyden til en tilfeldig valgt person, tiden en maratonløper bruker på å fullføre distansen, eller liknende.

Vi kan med en gang merke oss en viktig ting: Når X er en kontinuerlig stokastisk variabel, er sannsynligheten for at X skal ha en bestemt verdi lik null!! Hvis du synes at det virker merkelig, kan du prøve å anslå sannsynligheten for at høyden på vår tilfeldig utvalgte person er *nøyaktig* 1.825717… meter, eller at maratonløperen bruker *nøyaktig* 2 timer, 28 minutter og 14.26598… sekunder på tilbakelegge distansen. Du innser sikkert at vi kun kan snakke om sannsynligheten for at en kontinuerlig stokastisk variabel kan ligge i et *intervall*.

Denne erkjennelsen skal vi benytte til å definere en ***sannsynlighetstetthet*** $f(x)$ for en kontinuerlig stokastisk variabel. Du husker sikkert at for en diskret stokastisk variabel X var

$$P(a \le X \le b) = \sum_{x=a}^{b} P(X = x).$$

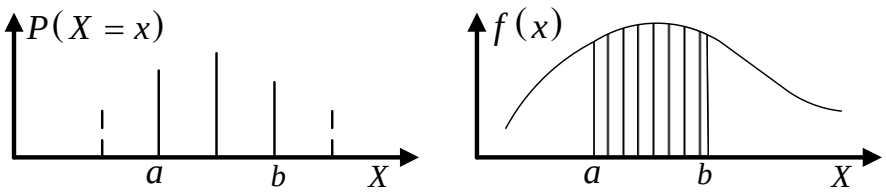
Analogt med dette definerer vi ***sannsynlighetstettheten*** $f(x)$ for en kontinuerlig stokastisk variabel X slik:

Sannsynligheten for at en kontinuerlig stokastisk variabel X skal ligge mellom to verdier a og b er

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

der $f(x)$ er ***sannsynlighetstettheten*** for variabelen X.

Situasjonen er illustrert på figuren nedenfor: For den diskrete fordelingen til venstre er $P(a \le X \le b)$ gitt ved summen av "søylene" fra og med $X = a$ til og med $X = b$. For den kontinuerlige fordelingen til høyre er $P(a \le X \le b)$ gitt ved arealet under grafen til $f(x)$.



Legg merke til at for en kontinuerlig fordeling spiller det ingen rolle om vi skriver $P(a \le X \le b)$ eller bare $P(a < X < b)$.

Vi summerer opp:

La X være en kontinuerlig stokastisk variabel. Til denne variabelen tilordnes en ***sannsynlighetstetthetsfunksjon*** $f(x)$ med disse egenskapene:

1. $f(x) \geq 0$ for alle aktuelle verdier av X.
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.
3. $P(a \leq x \leq b) = \int_a^b f(x)dx$.

De to første egenskapene innebærer at sannsynlighet aldri kan være et negativt tall, og at samlet sannsynlighet for alle mulige verdier av X er lik 1.

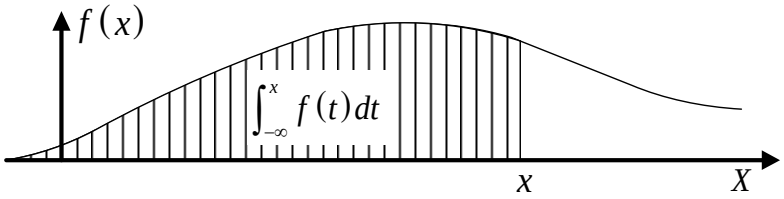
I praksis får vi ofte bruk for ***kumulativ sannsynlighetsfunksjon***. Dette er en funksjon som gir sannsynligheten for at en stokastisk variabel X skal ha en verdi mindre enn eller lik x. For en kontinuerlig stokastisk variabel X har vi:

Den kumulative sannsynlighetsfunksjonen er

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

der $f(x)$ er sannsynlighetstettheten til X.

Grafisk vil $F(x)$ være arealet under grafen til $f(t)$ til venstre for $t = x$.



Noen lærebøker bruker denne sammenhengen til å definere sannsynlighetstettheten $f(x)$. De tar utgangspunkt i den kumulative sannsynlighetsfunksjonen

$$P(X \leq x) = F(x) = \int_{-\infty}^{x} f(t)dt,$$

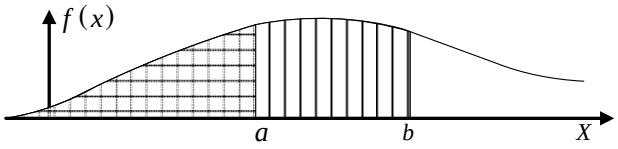
og sier deretter at

$$f(x) = \frac{dF(x)}{dx}.$$

Men den kumulative sannsynligheten er nyttig på andre måter også. Du vil etter hvert oppdage at mange av de nyttigste sannsynlighetstetthetsfunksjonene er vanskelige eller umulige å integrere. Da benytter vi tabeller over de kumulative sannsynlighetsfunksjonene. Når du skal beregne sannsynligheten for at X skal ligge mellom a og b der $b > a$, regner du slik:

$$P(a \leq X \leq b) = \int_a^b f(x)dx = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = F(b) - F(a).$$

Situasjonen er illustrert nedenfor. Det loddrett skraverte arealet gir sannsynligheten for at X skal ha en verdi mindre enn a. Det vannrett skraverte arealet gir sannsynligheten for at X skal ha en verdi mindre enn b. Differansen blir da sannsynligheten for at X ligger mellom a og b.



Hvis du har en tabell som gir *kumulativ* fordeling

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\,dt\,,$$

vil $F(b) - F(a)$ gi arealet mellom linjene $x = b$ og $x = a$, som svarer til sannsynligheten for at X skal ligge mellom a og b.

Etter disse betraktningene er tiden inne til å definere forventningsverdi (middelverdi) og varians for en kontinuerlig stokastisk variabel:

La X være en kontinuerlig stokastisk variabel med tilhørende sannsynlighetstetthetsfunksjon $f(x)$. Da er

Forventningsverdi $\mathrm{E}(X) = \mu = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$.

Varians $\mathrm{Var}(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\,dx = \int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx - \mu^2$.

Det fins et utall mer og mindre nyttige sannsynlighetstetthetsfunksjoner. Vi skal starte med den enkleste av dem: den ***uniforme*** sannsynlighetstetthetsfunksjonen. Deretter skal vi gå grundig inn på den klart nyttigste: ***normalfordelingen***. I senere notater skal vi skal også se på ***t-fordelingen***, som kan oppfattes som en generalisering av normalfordelingen. Til slutt skal vi se på et par andre funksjoner som brukes en god del i praksis.

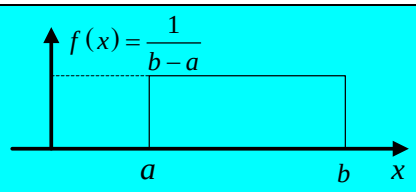
6.2. Uniform sannsynlighetstetthet.

Mange stokastiske variabler har *konstant* sannsynlighetstetthet innenfor et intervall (d.v.s. at det er lik sannsynlighet for å få en verdi innen et område i intervallet uansett hvor dette området plasseres). Et par eksempler:

Du trekker opp ei vanlig klokke og lar den gå inntil den stopper. Da er det rimelig å anta at det er helt tilfeldig hvor langviseren peker når klokka stopper. Med andre ord: Dersom langviserens vinkel med en eller annen start-linje er en kontinuerlig stokastisk variabel X, så er X uniformt fordelt i intervallet $[0, 360^0\rangle$.

Neste eksempel: Du måler temperaturer, og avrunder til nærmeste hele grad. Denne avrundingen gir en unøyaktighet som ligger i området $\left[-0.5^\circ, +0.5^\circ\right\rangle$. Dersom X er unøyaktigheten ved en slik måling, er det rimelig å anta at X er uniformt fordelt i dette intervallet.

Nå er vi klar til en generell definisjon på uniform sannsynlighet:



Den ***uniforme sannsynlighetstettheten*** er gitt ved

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{når } a \le x \le b \\ 0 & \text{ellers} \end{cases}$$

Vi sjekker at kravene for en sannsynlighetstetthet er oppfylt:

1. Forutsatt at $b > a$, er $f(x) \ge 0$ for alle x.
2. $\int_{-\infty}^{\infty} f(x)\,dx = \int_a^b \frac{1}{b-a}\,dx = \frac{1}{b-a}\left[x\right]_a^b = \frac{1}{b-a}(b-a) = \underline{1}$.

For treningens skyld kan vi jo finne forventningsverdien og standardavviket:

$$\mu = \int_{-\infty}^{\infty} x \cdot f(x)\,dx = \int_a^b x \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a}\left[\tfrac{1}{2}x^2\right]_a^b = \frac{1}{b-a} \cdot \tfrac{1}{2}\left(b^2 - a^2\right) = \underline{\underline{\tfrac{1}{2}(b+a)}}.$$

Men dette er jo midtpunktet mellom a og b (og det var vel ikke så uventet?).

Det er litt mer regnearbeid å finne standardavviket. Vi får

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2 = \int_a^b x^2 \cdot \frac{1}{b-a}\,dx - \left(\frac{b+a}{2}\right)^2 = \frac{1}{b-a} \cdot \left[\tfrac{1}{3}x^3\right]_a^b - \left(\frac{b+a}{2}\right)^2$$

$$= \tfrac{1}{3}\frac{b^3 - a^3}{b-a} - \frac{b^2 + 2ab + a^2}{4} = \frac{b^2 + ab + a^2}{3} - \frac{b^2 + 2ab + a^2}{4}$$

$$= \frac{b^2 - 2ab + a^2}{12} = \underline{\frac{(b-a)^2}{12}} \quad \Leftrightarrow \quad \underline{\underline{\sigma = \frac{1}{2\sqrt{3}}(b-a)}}$$

Den uniforme fordelingen er såpass enkel at mange av resultatene virker direkte innlysende, noe eksemplet nedenfor viser.

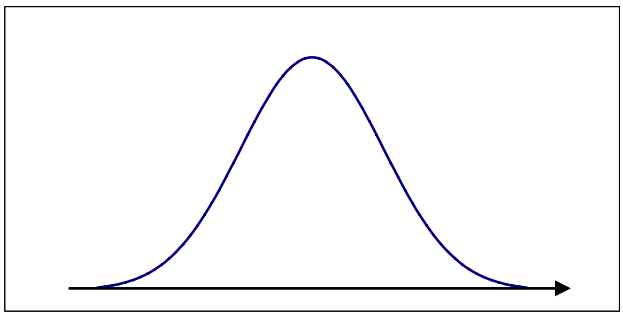
Eksempel 6.1: Anta at X er uniformt fordelt i området $\left[4, 20\right]$. Hva er sannsynligheten for at X skal ligge i området $\left[8, 10\right]$?

Løsning: Vi benytter at

$$P(8 \le x \le 10) = \int_8^{10} f(x)\,dx = \int_8^{10} \frac{1}{20-4}\,dx = \frac{1}{16} \cdot \left[x\right]_8^{10} = \frac{1}{16}(10-8) = \underline{\underline{\frac{1}{8}}}.$$

Og det er jo ventet, siden intervallet $\left[8, 10\right]$ er $\frac{2}{16} = \frac{1}{8}$ av hele intervallet $\left[4, 20\right]$.

6.3. Normalfordelingen.



I praksis har vi ofte bruk for en sannsynlighetstetthetsfunksjon som har sin største verdi nær en middelverdi, og som avtar mot null når vi kommer langt nok vekk fra middelverdien. Grafen har en typisk "klokkeform".

Se et eksempel til venstre.

Nå fins det mange matematiske funksjoner som har en graf som likner grafen ovenfor. Av alle disse er det en som peker seg ut. Ikke bare fordi den gjengir reelle situasjoner på en god måte. Men like mye fordi den har matematiske egenskaper som statistikerne setter stor pris på. Vi skal se på noen av disse egenskapene etter hvert. Denne funksjonen kalles ***normalfordelingen*** og ser slik ut:

Normalfordelingen er gitt ved $$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

der μ og σ er henholdsvis middelverdi og standardavvik for fordelingen.

Bare slapp av. Du skal slippe å huske denne formelen. Du får neppe bruk for den heller.

Ta for deg en variabel *X* som er normalfordelt. Sannsynligheten for at *X* skal ha en verdi som ligger mellom *a* og *b* er gitt ved

$$P(a\le x\le b)=\int_a^b f(x)\,dx=\int_a^b \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx\,.$$

Her kommer en stor nedtur: *Dette integralet lar seg ikke løse eksakt!!* Hva gjør du da?

Du kan naturligvis finne en tilnærmet verdi ved å løse integralet numerisk. Men det gir tidkrevende og tungvinte regninger. I dag kan vi riktignok la dataverktøy foreta utregningene. Men det er ekkelt nok å taste inn funksjonsuttrykket. Og jeg har jo lovd at du skal slippe å huske det stygge uttrykket.

Vi følger heller en framgangsmåte som jeg har antydet tidligere, og som ble utviklet i tiden før man fikk dataverktøy: *Vi bruker standardiserte tabeller over den tilhørende kumulative sannsynlighetsfunksjonen*. Før vi gyver løs på detaljene, vil jeg nok en gang presisere at:

Sannsynligheten for at *X* skal ligge mellom *a* og *b* er

$$P(a\le x\le b)=\int_a^b f(x)\,dx=P(x\le b)-P(x\le a)=F(b)-F(a)$$

der $F(x)$ er den kumulative sannsynlighetsfordelingen.

Når vi nå har banket fast dette grunnlaget, skal vi se hvordan vi jobber i praksis. Vi skal starte med en normalfordeling som har forventningsverdi $\mu = 0$ og standardavvik $\sigma = 1$. Denne spesielle normalfordelingen kalles forresten ***standard normalfordeling***. Når vi bruker standard normalfordeling, er det vanlig å bruke symbolet *Z* (ikke *X*) for den stokastiske variabelen (jeg skal snart fortelle hvorfor). Tabell over kumulativ standard normalfordeling finnes i alle lærebøker i statistikk. I vår lærebok finner du den både inne i permen og i det vedlagte formelarket. Tabellen for negative verdier av *z* er gjengitt nedenfor:

NEGATIVE *z* Scores

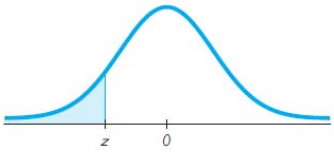


TABLE A-2 Standard Normal (*z*) Distribution: Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.50 and lower	.0001									
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	* .0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	* .0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

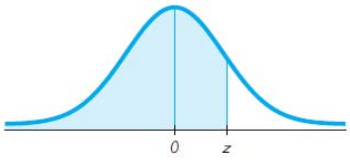
NOTE: For values of *z* below −3.49, use 0.0001 for the area.
*Use these common values that result from interpolation:

z score	Area
−1.645	0.0500
−2.575	0.0050

Av denne tabellen finner vi bl.a.:

- Sannsynligheten for å få en *z*-verdi som er mindre enn eller lik -1.00 er $P(z \le -1.00) = \underline{0.1587}$.
- Sannsynligheten for å få en *z*-verdi som er mindre enn eller lik -0.55 er $P(z \le -0.55) = \underline{0.2912}$.

Tabellen for positive verdier av *z* er gjengitt nedenfor:



POSITIVE *z* Scores

TABLE A-2 (*continued*) Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495 *	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	* .9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.50 and up	.9999									

NOTE: For values of *z* above 3.49, use 0.9999 for the area.

*Use these common values that result from interpolation:

z score	Area
1.645	0.9500
2.575	0.9950

Common Critical Values

Confidence Level	Critical Value
0.90	1.645
0.95	1.96
0.99	2.575

Av denne tabellen ser vi bl.a. at sannsynligheten for å få en *z*-verdi som er mindre enn eller lik 1.52 er

$$P(z \le 1.52) = \underline{0.9357}\,.$$

Dersom du først skjønner dette grunnprinsippet, går det greit å finne for eksempel:

- Sannsynligheten for å få en *z*-verdi som er *større* enn eller lik -1.00 er
  $P(z \ge -1.00) = 1 - P(z \le -1.00) = 1 - 0.1587 = \underline{0.8413}\,.$
- Sannsynligheten for å få en *z*-verdi som mellom -0.55 og 1.52 er
  $P(-0.55 \le z \le 1.52) = P(z \le 1.52) - P(z \le -0.55) = 0.9357 - 0.2912 = \underline{0.6445}\,.$

Disse sannsynlighetene kan du også finne med dataverktøy. I Excel har du funksjonen
=NORMSFORDELING(z) (merk *S*-en)
der du taster inn den *z*-verdien du ønsker å bruke.

Noen ganger ønsker vi å gå motsatt vei: Vi kjenner sannsynligheten, og ønsker å finne *z*-verdien. Vanligvis må vi da interpolere i tabellen. Slik interpolering tas gjerne "på øyemål", slik eksemplet nedenfor viser.

Eksempel 6.2: La *z* være standard normalfordelt. Bestem *a* slik at $P(z \le a) = 0.88$.

Løsning: Av tabellen ser vi at
$P(z \le 1.17) = 0.8790$,
$P(z \le 1.18) = 0.8810$.
Da sier vi at
$P(z \le \underline{1.175}) = 0.8800$.

Vi kan også bruke Excel. Da har vi funksjonen
=NORMSINV(p)
der *p* er den aktuelle sannsynligheten.

Så langt har vi kun sett på *standard normalfordeling*, d.v.s. en normalfordeling med middelverdi $\mu = 0$ og standardavvik $\sigma = 1$. Hva gjør vi dersom vi har en generell normalfordeling? Vi kan jo ikke lage egne tabeller for alle kombinasjoner av μ og σ. Løsningen kommer med denne setningen:

Dersom *X* er normalfordelt med middelverdi μ og standardavvik σ, er
$$Z = \frac{X - \mu}{\sigma}$$
normalfordelt med middelverdi 0 og standardavvik 1.

Dette innbærer at dersom *X* er normalfordelt med kjent middelverdi og standardavvik, regner vi ut *z*-scoren til *X* og bruker tabell for standard normalfordeling som før. Du husker sikkert at *z*-scoren angir hvor mange standardavvik *X* er fra μ (d.v.s. fra midtpunktet i fordelingen).

Du husker sikkert også at jeg tidligere (i punkt 3.6) har påstått at ca. to tredeler av alle dataverdier ligger mindre enn ett standardavvik fra middelverdien, mens ca. 95 % av alle dataverdier ligger mindre enn to standardavvik fra middelverdien dersom dataene er "normalt" fordelt. Disse påstandene kan vi nå begrunne. Av tabellen ser vi at dersom dataeneverdiene er normalfordelt, vil 15.87 % av alle dataene ha *z*-verdi mindre enn -1.00, og 84.13 % av alle dataene har *z*-verdi større enn +1. Da må $84.13\% - 15.87\% = 68.26\%$ av dataene ha *z*-verdier mellom disse to verdiene. Og det er omtrent to tredeler. På samme måte kan du vise at ca 95 % av dataene har *z*-verdi mellom -2 og +2.

Eksempel 6.3: I en populasjon av voksne menn er kroppshøyden normalfordelt med middelverdi $\mu = 175.3$ cm og standardavvik $\sigma = 7.1$ cm. Bestem sannsynligheten for at en tilfeldig uttrukket mann skal ha:

a) Kroppshøyde under 169.0 cm
b) Kroppshøyde over 182.3 cm
c) Kroppshøyde mellom 171.0 cm og 184.5 cm.

Løsning: Vi går i gang og beregner *z*-scorer, og bruker tabellen. Vi får:

a) $$z = \frac{169.0 - 175.3}{7.1} = \underline{-0.887}\,.$$

Vi går inn i tabellen på $z = -0.89$ og får en sannsynlighet på 0.1867. I praksis runder vi av til 0.187. Hvis du vil være svært nøye, kan du prøve å interpolere i tabellen for å få med ett siffer til, eller du kan bruke Excel og får 0.18754.

b) $$z = \frac{182.3 - 175.3}{7.1} = \underline{0.9859}\,.$$

Vi går inn i tabellen på $z = 0.99$ og får en sannsynlighet på 0.8389. Verdien $z = 0.99$ er litt for stor, så vi runder nedover til en sannsynlighet på 0.838. Med Excel får vi 0.8379. Men dette er sannsynligheten for å få en høyde *under* 182.3. Sannsynligheten for å få en høyde *over* 182.3 er da $1 - 0.838 = \underline{0.162}$.

c) Her trenger vi to *z*-verdier:

$$z_L = \frac{171.0 - 175.3}{7.1} = \underline{-0.6056}$$

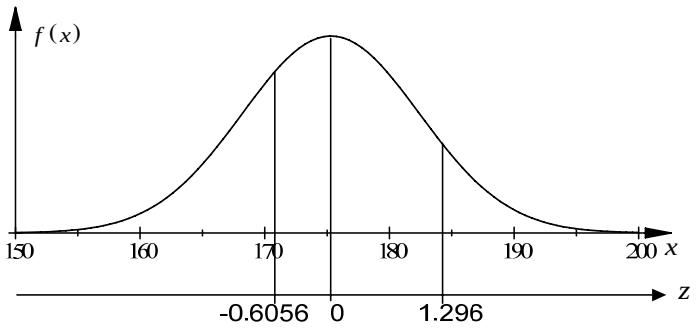
og

$$z_H = \frac{184.5 - 175.3}{7.1} = \underline{1.296}\,.$$

Sannsynligheten for en høyde mellom 171.0 cm og 184.5 cm blir da

$$P(z \le 1.296) - P(z \le -0.6056) = 0.9025 - 0.2724 = \underline{0.6301}\,.$$

Når vi jobber med slike oppgaver, synes jeg at det er greit å lage en grovskisse av en normalfordeling og tegne inn både *x*-verdier og samhørende *z*-verdier. Da er det lettere å holde oversikten. En slik skisse for deloppgave c) er vist nedenfor:



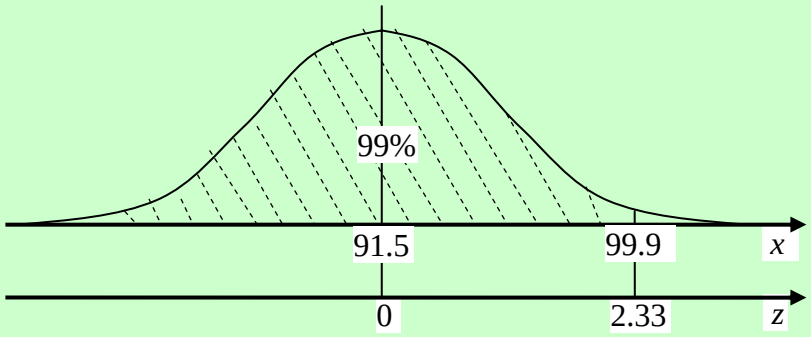
La meg avslutte med et eksempel der vi tar utgangspunkt i sannsynlighet og skal beregne *X*.

Eksempel 6.4: Jeg har stadig erfart at personbiler har for liten takhøyde. Vi kan ta utgangspunkt i at voksne menn har en sittehøyde som er normalfordelt med $\mu = 91.5\,\text{cm}$ og standardavvik $\sigma = 3.6\,\text{cm}$. Hvor stor må da høyden fra setet til biltaket være for at 99 % av alle voksne menn skal kunne sitte uten å stange hodet i biltaket?

Vi finner fram en *z*-verdi-tabell, og ser at det 99.01 % sannsynlighet for at en tilfeldig *z*-verdi er lavere $z = 2.33$. Vi kan da sette opp

$$z = \frac{x-\mu}{\sigma} \quad \Leftrightarrow \quad x = \mu + \sigma z = 91.5 + 3.6 \cdot 2.33 = \underline{99.9}\,.$$

Avstanden fra setet til biltaket må altså være minst 99.9 cm for at 99 % av alle voksne menn skal kunne sitte uten å stange hodet i taket. Når vi illustrerer med figur, får vi:



I resten av dette kurset vil du stadig komme i kontakt med problemer som går ut på å bruke normalfordelings-tabell slik det er gjort i eksemplene ovenfor. Du *MÅ* kunne denne teknikken! Jeg skal derfor repetere hovedtrekkene:

Når *X* er normalfordelt, finner du *z*-score med

$$Z = \frac{X-\mu}{\sigma}$$

der μ og σ er henholdsvis middelverdi og standardavvik for *X*. Denne *z*-scoren er standard normalfordelt, slik at du kan bruke tabell som gir sannsynlighet for at *z* er *mindre* enn eller lik den beregnede verdien.

6.4. Hvordan fordeles utvalgs-parametre? Sentralgrensesetningen.

Jeg skal ta utgangspunkt i eksemplet med at en andel $p = 40\% = 0.40$ av en *populasjon* som består av alle norske velgere er mot norsk medlemskap i EU. Vi trekker et *tilfeldig utvalg* fra denne populasjonen. Vi må da være temmelige heldige dersom den andelen *i utvalget* som er mot norsk medlemskap i EU også er nøyaktig lik $p = 0.40$. Men vi vil nok finne en andel som er *omtrent* lik *p*. Vi sier at vi finner *estimat* av *p*. Jo større utvalget er, jo bedre blir vårt estimat av *p*.

La oss nå anta at vi trekker mange (like store) utvalg, og beregner estimat av *p* for hvert utvalg. Da vil vi se at de estimatene vi finner, klumper seg rundt rett verdi av *p*. Mer presist: Dersom vi beregner riktig mange slike estimater av *p*, vil disse estimatene være *normalfordelt*

med middelverdi lik *p*. Vi sier gjerne at *forventningsverdien* for våre estimater er lik *p*, eller at estimatene våre av *p* er *forventningsrett*.

Tilsvarende forhold kan vi finne for *middelverdi*. Anta at i en *populasjon* er dataene normalfordelt med middelverdi μ og standardavvik σ. Du trekker et tilfeldig utvalg fra denne populasjonen, og beregner middelverdien $\bar{x}$ fra dette utvalget. Da vil du finne at $\bar{x}$ ligger i nærheten av μ. Dersom du trekker mange slike utvalg, og beregner $\bar{x}$ for hvert av dem, vil du se at de beregnede verdiene av $\bar{x}$ vil fordele seg rundt μ på en slik måte at middelverdien av $\bar{x}$ er lik μ. Med andre ord: $\bar{x}$ er et forventningsrett estimat av μ.

Før vi går videre, vil jeg presisere hovedpoenget: Anta at dataene i en populasjon har en eller annen egenskap (for eksempel en middelverdi) som kan gis en numerisk verdi. Vi trekker *tilfeldige utvalg* fra denne populasjonen, og *estimerer* verdien for denne egenskapen for hvert utvalg. Da vil disse estimatene fordele seg rundt (og forhåpentlig nær) verdien i populasjonen. Dersom middelverdien av våre estimater er lik verdien i populasjonen, sier vi at estimatene er *forventningsrett*.

En liten presisering: Når vi beregner middelverdien av våre estimater, bør vi ha "uendelig mange" slike estimater til disposisjon. I praksis er ikke dette mulig. Men skarpe teoretikere kan resonnere seg fram til hvordan estimatene vil fordele seg. På grunnlag av disse teoretiske analysene kan vi avgjøre om våre estimat er forventningsrette eller ikke.

Vi skal i første omgang se på sammenhengen mellom middelverdien μ i en populasjon og beregninger av middelverdi $\bar{x}$ i et utvalg fra populasjonen. Da kan vi vise:

Anta at en stokastisk variabel *X* er normalfordelt med forventningsverdi μ og standardavvik σ. Vi trekker tilfeldige utvalg på *n* verdier fra populasjonen, og beregner $\bar{x}$ for disse utvalgene. Da vil $\bar{x}$ for utvalgene være normalfordelt med:

Forventningsverdi $\mu_{\bar{x}} = \mu$.

Standardavvik $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$.

Les nøye hva som står i denne setningen. For det første forutsettes det at *X* er *normalfordelt* i populasjonen. Deretter sies det at når du beregner $\bar{x}$ for tilfeldige utvalg fra populasjonen, vil $\bar{x}$ være et forventningsrett estimat for μ. Videre sies det at dersom $\bar{x}$ beregnes for mange tilfeldige utvalg, vil *spredningen* i de beregnede verdien av $\bar{x}$ være slik at standardavviket for $\bar{x}$ er lik standardavviket i populasjonen delt på kvadratrota av antall elementer i utvalget.

Dette må illustreres med et eksempel:

Eksempel 6.5: Kroppshøyden for voksne menn er normalfordelt med middelverdi $\mu = 175.3$ cm og standardavvik $\sigma = 7.1$ cm. Vi trekker ut et tilfeldig utvalg på $n = 25$ slike kroppshøyder, og beregner gjennomsnittet $\bar{x}$ av dem. Finn forventningsverdi og standardavvik for $\bar{x}$.

Løsning: Dersom vi gjentar dette veldig mange ganger, vil våre verdier av $\bar{x}$ være normalfordelt slik at:

Forventningsverdien for $\bar{x}$ blir $\mu_{\bar{x}} = \mu = \underline{\underline{175.3\text{cm}}}$.

Standardavviket for $\bar{x}$ blir $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{7.1\text{cm}}{\sqrt{25}} = \underline{\underline{1.42\ \text{cm}}}$.

Hittil har vi forutsatt at populasjonen var normalfordelt. Hva skal vi gjøre dersom dette kravet ikke er oppfylt? Her kommer ***sentralgrense-setningen*** inn i bildet. Den fins i flere varianter. Men vi skal klare oss med denne:

Sentralgrense-setningen:
Anta at en stokastisk variabel X har middelverdi μ og standardavvik σ.
Vi trekker tilfeldige utvalg på n verdier fra populasjonen, og beregner $\bar{x}$ for disse utvalgene.
Da vil fordelingen av $\bar{x}$ for utvalgene gå mot en normalfordeling med forventningsverdi $\mu_{\bar{x}} = \mu$ og standardavvik $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$ når $n \to \infty$.

Merk deg forskjellen mellom denne setningen og den forrige. I sentralgrensesetningen er kravet om at X skal være normalfordelt droppet. Til gjengjeld må vi finne oss i at fordelingen for $\bar{x}$ *går mot* en normalfordeling når n blir stor istedenfor å *være* normalfordelt.

Første gang du leser dette, vil du antakelig trekke på skuldrene og si "Ja vel, og hva så? Er nå dette noe å mase om?" Men dette er faktisk svært viktig. Denne setningen gjør at du med god samvittighet kan bruke alt du har lært (og skal lære) om normalfordelingen *også når du behandler data som ikke er normalfordelt*, forutsatt at du holder deg til middelverdier av tilstrekkelig store utvalg. Og la oss bare innse det: De fleste dataene du behandler i praksis, er ikke normalfordelte.

Hva mener vi med "tilstrekkelig store utvalg?" Det avhenger helt av hvor store krav du stiller til nøyaktighet, og ikke minst av hvordan dataene i populasjonen er fordelt. Dersom populasjonsfordelingen ikke er alt for ekstrem, er det vanlig å si at fordelingen av $\bar{x}$ er normalfordelt dersom $n \geq 30$. Er populasjonsfordelingen nær normalfordelt, kan dette kravet slakkes.

Vi avrunder med en oppsummering, som også angir den praktiske arbeidsmåten:

Anta at X har middelverdi μ og standardavvik σ. Dersom betingelsene i sentralgrenseteoremet er oppfylt, har vi at:

$\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ er normalfordelt med $\mu_{\bar{x}} = \mu$ og $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$.

Da er

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$ standard normalfordelt.

Eksempel 6.6: Anta at *X* er *uniformt* fordelt i området $[0,10]$. Hva er sannsynligheten for at gjennomsnittet av 25 tilfeldig uttrukne tall fra denne populasjonen skal være større enn 6.00?

Løsning: Vi har tidligere vist at når *X* er uniformt fordelt i området $[a,b]$, så er

$$\mu = \frac{a+b}{2} = \frac{0+10}{2} = \underline{5}$$

og

$$\sigma = \frac{1}{2\sqrt{3}}(b-a) = \frac{10-0}{2\sqrt{3}} \approx \underline{2.89}\,.$$

For gjennomsnittet $\bar{x}$ av $n = 25$ tilfeldig uttrukne tall i området $[0,10]$ er

$$\mu_{\bar{x}} = \mu = \underline{5.00}$$

og

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.89}{\sqrt{25}} = \underline{0.578}\,.$$

Videre er *n* såpass stor at vi tar sjansen på at $\bar{x}$ er normalfordelt. Vi finner *z*-scoren til $\bar{x} = 6.00$ slik:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{6.00 - 5.00}{0.578} = \underline{1.73}\,.$$

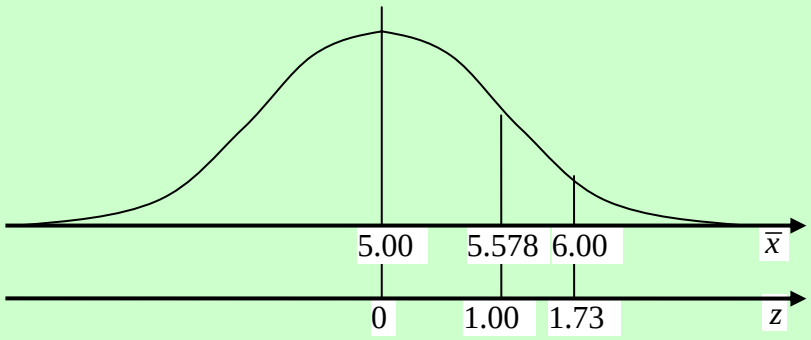
Av tabell for kumulert standard normalfordeling finner vi at

$$P(\bar{x} \le 6.00) = P(z \le 1.73) = \underline{0.9582}$$

slik at

$$P(\bar{x} > 6.00) = 1 - 0.9582 = \underline{\underline{0.0418}}\,.$$

Situasjonen er altså slik:



Sentralgrensesetningen er faktisk en av grunnene til at normalfordelingen har fått en så sentral plass i statistikken. Det er også derfor setningen ovenfor har fått navnet "The Central Limit Theorem". På norsk burde setningen hett noe likt som "den sentrale grensesetningen". Navnet "sentralgrense-setningen" er en språklig lapsus.

6.5. Normalfordeling som tilnærming til binomisk fordeling.

Det er ikke bare sentralgrensesetningen som gjør normalfordelingen så anvendelig, tross sitt stygge funksjonsuttrykk. Vi kan nemlig vise at den binomiske fordelingen nærmer seg normalfordelingen når *n* blir stor slik setningen nedenfor angir:

Når $n \cdot p$ vokser og $n(1-p)$ vokser, så vil den binomiske fordelingen nærme seg en normalfordeling med $\mu = n \cdot p$ og $\sigma = \sqrt{n \cdot p(1-p)}$.

Jo større $n \cdot p$ og $n(1-p)$ er, jo bedre er tilnærmingen. Vår lærebok sier at tilnærmingen er tilstrekkelig bra når $n \cdot p \geq 5$ og $n(1-p) \geq 5$. Andre lærebøker kan stille litt andre krav, for eksempel at $n \cdot p \cdot (1-p) \geq 5$. Videre kan vi si at tilnærmingen er best når $p \approx 0.5$, slik at den binomiske fordelingen i utgangspunktet er relativt symmetrisk.

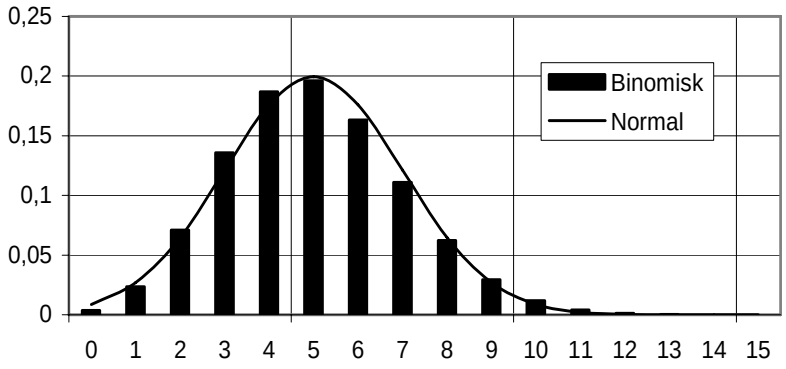
Som eksempel skal vi se på en binomisk fordeling med $n = 25$ og $p = 0.2$, slik at vi befinner oss på grensen for hva som er "brukbar" tilnærming. Figuren nedenfor viser den binomiske fordelingen sammen med en normalfordeling med

$$\mu = n \cdot p = 25 \cdot 0.2 = 5.0$$

og

$$\sigma = \sqrt{n \cdot p(1-p)} = \sqrt{25 \cdot 0.2 \cdot 0.8} = 2.00.$$

Du ser at normalfordelingen er en rimelig bra tilnærming, men langt fra perfekt. Hovedgrunnen til forskjellene er at mens normalfordelingen er symmetrisk om $\mu = 5$, så er ikke binomialfordelingen symmetrisk.



Jeg skal bruke dette eksemplet til å vise hvordan vi bruker normalfordelingen til å finne tilnærmede sannsynligheter for en binomisk fordeling.

Eksempel 6.7: Vi har en binomisk fordeling med $n = 25$ og $p = 0.2$. Bestem:

a) $P(x > 6)$

b) $P(3 \leq x \leq 8)$

med binomisk fordeling, og deretter tilnærmede sannsynligheter med normalfordeling.

Løsning: Vi vet allerede at $\mu = n \cdot p = 5.0$ og at $\sigma = \sqrt{n \cdot p(1-p)} = 2.00$.

Når vi jobber med *diskrete* fordelinger, må vi være nøye med om det står bare "større enn" eller "større enn eller lik" (og tilsvarende for "mindre enn"). Når det i oppgave a) står "større enn 6", må det for binomisk fordeling oppfattes som 7, 8, 9, …, 25. Istedenfor å regne ut alle disse sannsynlighetene, bruker vi kumulativ binomisk fordeling og beregner

$$1 - P(x \le 6) = 1 - 0.7800 = \underline{\underline{0.2200}},$$

der beregningen er foretatt med Excel.

Når vi bruker normalfordelingen, må vi være enda mer på vakt. Vi må da anta at "6" i binomisk fordeling tilsvarer intervallet $[5.5, 6.5\rangle$ i normalfordelingen. "Større enn 6" må da tolkes som "fra 6.5 og oppover" når vi går fra binomisk til normalfordeling. Vi må altså finne $P(x \ge 6.5)$ med normalfordeling. Vi går (som vanlig) veien om *z*-score:

$$z = \frac{x - \mu}{\sigma} = \frac{6.5 - 5.0}{2.0} = \underline{0.75}.$$

Av *z*-tabellen finner vi at

$$P(x \ge 6.5) = P(z \ge 0.75) = 1 - P(z \le 0.75) = 1 - 0.7734 = \underline{\underline{0.2266}}.$$

Dette er en rimelig bra tilnærming til den eksakte verdien som vi fant ovenfor.

I oppgave b) må vi være enda mer på vakt. $P(3 \le x \le 8)$ må oppfattes som $P(3) + P(4) + \cdots + P(8)$. Jeg foretrekker å bruke kumulativ binomisk fordeling, og får

$$P(3 \le x \le 8) = P(x \le 8) - P(x \le 2) = 0.9532 - 0.0982 = \underline{\underline{0.8550}}.$$

Når jeg skal finne tilnærmede verdier med normalfordeling, må $P(3 \le x \le 8)$ betraktes som

$$P(2.5 \le x < 8.5) = P(x < 8.5) - P(x < 2.5).$$

Vi beregner *z*-scorer som før:

$$z_H = \frac{8.5 - 5.0}{2.0} = 1.75,$$

og

$$z_L = \frac{2.5 - 5.0}{2.0} = -1.25.$$

Da får vi at

$$P(2.5 \le x < 8.5) = P(z < 1.75) - P(z < -1.25) = 0.9599 - 0.1056 = \underline{\underline{0.8543}},$$

som også må sies å være en brukbar tilnærmelse.

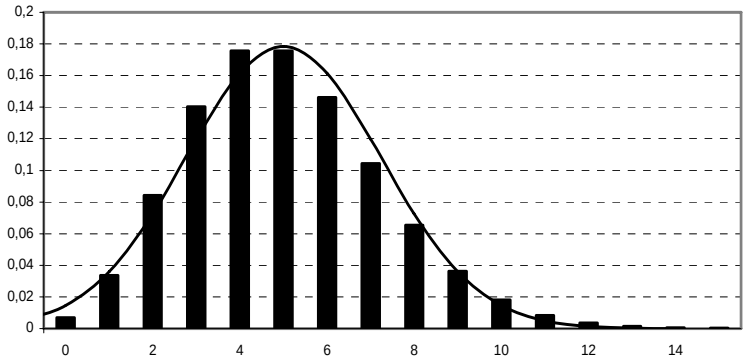
Den justeringen jeg gjør ved at jeg legger til eller trekker fra 0.5, kalles *kontinuites-korreksjon* (eller heltalls-korreksjon). Vi må alltid foreta en slik korreksjon når vi går over fra en diskret fordeling til en kontinuerlig fordeling. Men dersom *n* er stor får korreksjonen liten betydning.

6.6. Normalfordeling som tilnærming til Poisson-fordelingen.

Vi har tidligere sett at Poisson-fordelingen kan være en brukbar tilnærmelse til binomisk fordeling. Da er det rimelig at normalfordelingen kan være brukbar tilnærmelse til Poisson-fordelingen når visse krav er oppfylt. Vi kan vise at

Når μ øker, vil Poisson-fordelingen nærmer seg en normalfordeling med samme middelverdi μ, og med standardavvik $\sigma = \sqrt{\mu}$.

Det er vanlig å sette at normalfordelingen er en brukbar tilnærmelse dersom $\mu \geq 5$. Jeg hadde helst sett et litt strengere krav. Figuren nedenfor viser en Poisson-fordeling med $\mu = 5$ sammen med den tilhørende normalfordelingen.



Når du bruker normalfordelingen som tilnærming for Poisson-fordelingen, må du bruke samme kontinuitets-korreksjon som vi brukte for binomisk fordeling.

6.7. Noen generaliseringer.

En del av de reglene som vi har vært gjennom, er egentlig spesialtilfeller av mer generelle setninger. Her kommer en slik mer generell setning:

Anta at X_1, X_2, …, X_n er *n* uavhengige, normalfordelte variabler med middelverdier henholdsvis μ_1, μ_2, …, μ_n og varianser henholdsvis σ_1^2, σ_2^2, …, σ_n^2.
Definer

$$Y = a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n .$$

Da er *Y* normalfordelt, og har:

Forventning $\mathrm{E}(Y) = \mu_Y = a_0 + a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$.

Varians $\mathrm{Var}(Y) = \sigma_Y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2$.

La oss se et par eksempler på bruken av denne setningen:

Eksempel 6.8: Middagen din består av 4 kjøttkaker og 3 poteter som du plukker tilfeldig ut av gryta. Etter inngående studer har du funnet ut at vekten av kjøttkakene er normalfordelt med middelverdi $\mu_k = 50$ gram og standardavvik $\sigma_k = 5$ gram, mens potetenes vekt er normal-fordelt med middelverdi $\mu_p = 75$ gram og standardavvik $\sigma_p = 10$ gram.
Hva er sannsynligheten for at middagen din skal veie mer enn 400 gram?

Løsning: Kall vekten av en tilfeldig kjøttkake for *K*, mens vekten av en tilfeldig potet kalles *P*. Kall vekten av hele middagen for *M*. Da kan vi sette opp:

$$M = K_1 + K_2 + K_3 + K_4 + P_1 + P_2 + P_3 .$$

Merk at dette ikke kan skrives $M = 4K + 3P$. Hvis du skriver $M = 4K + 3P$, betyr det at du tar *en* kjøttkake og ganger vekten av den med 4, og *en* potet og ganger vekten av den med 3. Men det er jo ikke det du gjør. Du tar 4 kjøttkaker *uavhengig av hverandre*, og tilsvarende for potetene. Derfor må du bruke den litt tungvinte skrivemåten. Da får du:

$$\mu_M = \mathrm{E}(M) = 50 + 50 + 50 + 50 + 75 + 75 + 75 = \underline{425} .$$

$$\sigma_M{}^2 = \mathrm{Var}(M) = 5^2 + 5^2 + 5^2 + 5^2 + 10^2 + 10^2 + 10^2 = \underline{400} \quad \Leftrightarrow \quad \sigma_M = \sqrt{400} = \underline{20} .$$

Du har altså funnet ut at i gjennomsnitt vil middagen din veie 425 gram, med et standardavvik på 20 gram. Videre vet du at vekten av hele middagen, *M*, er normalfordelt. Nå kan du finne sannsynligheten for at en tilfeldig middag skal veie mindre enn 400 gram:

$$z = \frac{400 - 425}{20} = -1.25 ,$$

som gir at sannsynligheten for at middagen skal veie under 400 gram er 0.1056. Da er sannsynligheten for at middagen skal veie minst 400 gram $1 - 0.1056 = \underline{0.8944}$.

Eksempel 6.9: Vi plukker tilfeldig ut en mann fra en populasjon der kroppshøyden er normalfordelt med middelverdi $\mu_M = 175.3\,\mathrm{cm}$ og standardavvik $\sigma_M = 7.1\,\mathrm{cm}$. Så plukker vi tilfeldig ut en kvinne fra en populasjon der kroppshøyden er normalfordelt med middelverdi $\mu_K = 161.6\,\mathrm{cm}$ og standardavvik $\sigma_K = 6.4\,\mathrm{cm}$. Hva er sannsynligheten for at kvinnen er høyere enn mannen?

Løsning: Kall mannens høyde *M* og kvinnens høyde *K*. Høydeforskjellen blir da

$$D = M - K .$$

Vi skal altså finne sannsynligheten for at $M < K \quad \Leftrightarrow \quad D < 0$. Vi beregner derfor middelverdi og standardavvik for høydeforskjellen *D*:

$$\mu_D = \mu_M - \mu_K = 175.3 - 161.6 = \underline{13.7} .$$

$$\sigma_D{}^2 = \sigma_M{}^2 + (-1)^2 \sigma_K{}^2 = 7.1^2 + 6.4^2 = \underline{91.37} \quad \Leftrightarrow \quad \sigma_D = \sqrt{91.37} = \underline{9.56} .$$

Høydeforskjellen *D* er altså i gjennomsnitt 13.7 cm med standardavvik 9.56 cm. Dessuten vet vi at *D* er normalfordelt. Da finner vi *z*-score og bruker *z*-tabellen:

$$z = \frac{0 - 13.7}{9.56} = \underline{-1.433} ,$$

som gir at sannsynligheten for at kvinnen skal være høyere enn mannen er 0.0759.

La meg avslutte med litt teori. Du husker sikkert at middelverdien av *n* objekter er

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}x_1 + \frac{1}{n}x_2 + \cdots \frac{1}{n}x_n .$$

Dersom alle disse objektene er plukket tilfeldig fra samme *normalfordelte* populasjon med middelverdi μ og standardavvik σ, har vi at:

$$\mathrm{E}(\overline{x}) = \mu_{\overline{x}} = \frac{1}{n}\mu + \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu = \frac{1}{n}\mu(1+1+\cdots+1) = \frac{1}{n}\mu \cdot n = \underline{\mu}.$$

Videre er:

$$\mathrm{Var}(\overline{x}) = \sigma_{\overline{x}}{}^2 = \left(\frac{1}{n}\right)^2 \sigma^2 + \left(\frac{1}{n}\right)^2 \sigma^2 + \cdots + \left(\frac{1}{n}\right)^2 \sigma^2$$

$$= \left(\frac{1}{n}\right)^2 \sigma^2 (1+1+\cdots+1) = \left(\frac{1}{n}\right)^2 \sigma^2 \cdot n = \frac{\sigma^2}{n} \quad \Leftrightarrow \quad \sigma_{\overline{x}} = \sqrt{\frac{\sigma^2}{n}} = \underline{\frac{\sigma}{\sqrt{n}}}$$

Dessuten er $\overline{x}$ også normalfordelt, men det skal vi ikke vise nå.

Vi har altså vist at den setningen som vi innledet kap. 6.4 med, følger direkte av den mer generelle setningen som vi startet dette underkapitlet med.