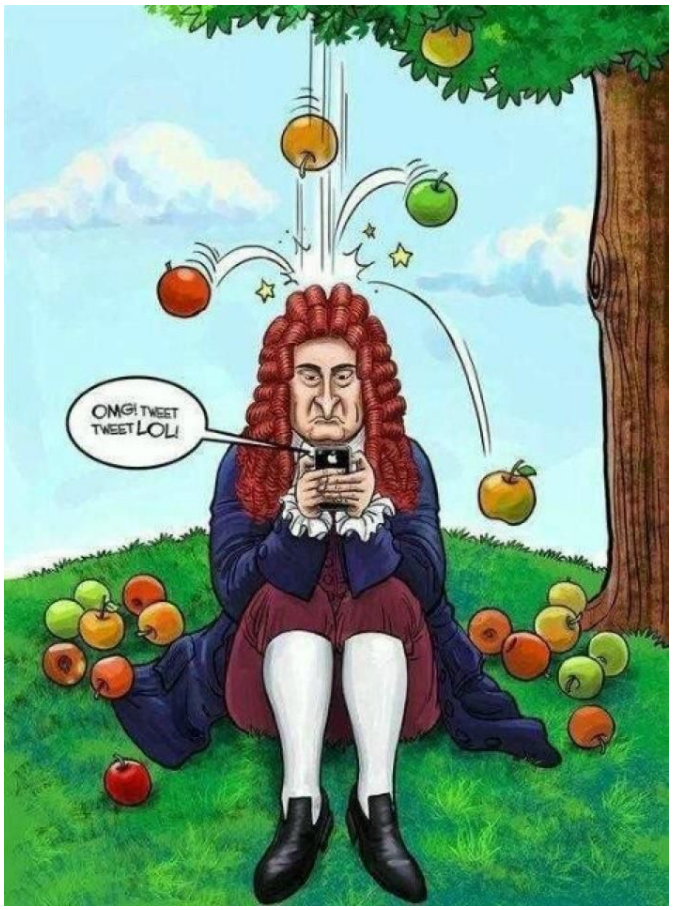
Del 1 - Frekvens, sannsynlighet, deskriptiv statistikk

Læringsmål

- Lære om frekvens, relativ frekvens og sammenheng med sannsynlighet
- Vise gode datapresentasjoner i form av spredningsplott og histogrammer
- Beregne enkle statistiske mål som gjennomsnitt, varians og standardavvik
- Vurdere og kovarians og korrelasjon
- Estimere varians i estimater utledet av flere stokastiske variable
- Utvikle egne Python-koder for å gjøre statistiske estimater

Innledning

Fysikk er en vitenskapelig disiplin der man ut fra observasjoner og eksperimenter etablerer en forståelse av sammenhenger i naturen, oftest ved hjelp av matematiske modeller. Disse modellene kan siden hjelpe oss til å *predikere* (forutsi) hvordan naturen vil oppføre seg når betingelsene endres. Har du noen gang tenkt på hva en 'naturkonstant' slik som G (Newtons gravitasjonskonstant) egentlig representerer? Det riktige ordet på denne 'konstanten' bør være "empirisk fysisk konstant", fordi den er basert på observasjoner, en modell ($F = G\frac{m_1 m_2}{r^2}$) og en tilpasning av modellen slik at G passer til observasjonene. Med andre ord, denne 'naturloven' formulert av Newton er egentlig en matematemisk konstruksjon som viser seg å passe til observasjoner ('loven' er bygget på *empiri*). Når vi fysikere bruker denne modellen i mekanikk glemmer vi nok ofte dette bakteppet og tenker kanskje ikke på at vi gjør en *prediksjon* når vi beregner hvor lang tid eplet bruker på å falle fra grenen ned til bakken. Slik fysikken læres i dag starter man oftest med fysikkmodellene, og ikke med observasjoner av fysiske fenomen. Det antas egentlig at du har en grunnleggende forståelse av for eksempel gravitasjonen som virker mellom eplet og jorda eller hvordan en pendel svinger, og at det er mer effektivt å direkte lære om fysikkens lover som 'naturlover'. Men hva hvis du utvikler en mer komplisert modell som strekker seg utenfor det opprinnelige observasjonsgrunnlaget teorien først ble utviklet for? Vil du akseptere denne modellen slik den er, eller er det behov for nye målinger for å *validere* den (bekrefte gylidgheten)? Og hvis du observerer en forskjell mellom modell og måleresultat, er det da nødvendigvis noe feil med modellen eller målingene?



Vi trenger matematiske analyser og statistikk for å tolke data fra fysikkeksperimenter, men også som hjelp når vi designer eksperimenter. Videre er statistiske konsepter nyttige innen teoretiske fysikkdispliner slik som statistisk mekanikk og kvantemekanikk. Mange av prosessene og målingene vi skal diskutere er såkalte *stokastiske prosesser*, der man ikke kan *eksakt* kan forutsi utfallet. Begrepet stokastisk (=tilfeldig) er kanskje litt annerledes her enn i dagligtalen , siden 'tilfeldig' i denne sammenhengen betyr at man kan si noe om *sannsynligheten* for et gitt utfall. Det å kaste en kron/mynt har et tilfeldig utfall, men sannsynligheten for å få mynt er fortsattt 1/2=0.5 (ett utfall av to mulige). Et annet eksempel er å kaste en terning, som har 6 mulige utfall. Vi kan 'kun' si noe om sannsynligheten for f.eks. å få 3 på terningen (sannynligheten er 1/6). For en radioaktiv atomkjerne kan vi bare si noe om sannsynligheten for at kjernen desintegrer innen et gitt tidsintervall. Vi kan ikke forutsi hastigheten til et enkelt oksygenmolekyl i luft ved romtemperatur, men vi kan si noe om sannsynligheten for at molekylet har en viss hastighet. Når 100 studenter gjør en måling av vekten til et lodd vil man kunne få 100 måleresultater der i prinsippet ingen trenger å være like, og der man for student #101 bare kan anslå sannsynligheten for at han/hun måler en viss verdi. For en gitt måling der måleinstrumentene og prosedyrene er beheftet med usikkerheter må målingen anslås med en total usikkerhet, der usikkerheten reflekter et intervall der det er sannsynlig at den 'sanne' måleverdien befinner seg. Disse eksemplene viser at mange fenomen og målinger har tilfeldige utfall (vi kan ikke forutsi eksakt hva som skal skje), men at vi kan si noe om *sannsynligheten* for utfallet. Noen er iboende tilfeldige (terningen, atomkjernen) mens andre skyldes begrensinger og/eller feil i f.eks. måledata eller i modeller vi bruker til å forklare naturen.

Vi skal her følge opp en del av tingene du allerede har lært under "statistikk-delen" av STK-FYS1110, men skal bruke en mer empirisk vinkling med utgangspunkt i data og fordelinger. Kapittelet følger i stor grad kapittel 1 og tar også noe fra kapittel 5 i læreboka (Modern Mathematical Statistics with Applications).

Stokastiske variable

Når du gjør en måling av det samme fenomenet gjentatte ganger der man ikke observerer den samme måleverdien hver gang, sier man at måleverdien x_i representerer et mulig utfall av en *stokastisk variabel* X. 'Målingen' kan for eksempel være et terningkast, der variabelen vil være *diskret*. For terningen vil det totale utfallsrommet være $S = (1, 2, 3, 4, 5, 6)$ der gjentatte terningkast for eksempel kan gi observasjonene eller 'målingene' x_1=2, x_2=3, x_3=5 osv. Hvis man har en kontinuerlig positiv variabel kan man for eksempel ha $S = [0, \infty >$. La oss si at du for sistnevnte lager en måleserie med 100 målinger; $\underline{M} = \{x_1, x_2, \ldots, x_{100}\} = \sum_{i=1}^{100} \{x_i\}$ så vil måleserien representere et *utvalg* av utfallsrommet; $\underline{M} \in S$. I det siste tilfellet kan du aldri lage en måleserie som dekker det totale utfallsrommet for den stokastiske kontinuerlige variabelen X, du kan kun måle et (eller flere) utvalg.

O1 Plott ditt første datasett

Last ned datasettet $\underline{M}_{motstand}$, som finnes i filen *motstand.txt* (hint: **numpy.loadtxt**). Dette er en en rekke med 100 tall, som er fra et eksperiment der man gjentatte ganger har målt motstanden i en 10 m lang koppertråd med tverssnitt 1 mm^2 (syntetisk generterte data). Ved 20 °C vil motstanden i tråden være 170 mΩ. Dataene er ordnet kronologisk, slik at x_1 er første måling og x_{100} er siste måling. En fin første måte å få et bedre overblikk av datane på er å plotte disse som et spredningsplott (scatterplot). I dette tilfellet plottes hver måling som et punkt i et kartesisk koordinatsystem, der (i dette tilfellet) abscissen er målenummer og ordinaten er måleverdien. Lag et spredningsplott med de 100 tallene. I dette tilfellet skal du låse y-aksen til en minimums- og maksimumsverdi på hhv 160 og 180 $m\Omega$. Vi skal grave litt dypere i datasettet senere, og da kan vi 'slippe løs' skaleringen av y-aksen. NB; I **Matplotlib** er det standard at punktene er fylte, men det er her hensiktsmessig å plotte hvert punkt som feks en ring for bedre visualisering.

```
#Svar
```

Frekvensfordeling og sannsynlighet

Spredningsplottet ovenfor kan være en fin måte å vise data på, spesielt når man skal se på sammenhenger mellom måleresultat og såkalte forklaringsvariabler (vi skal komme tilbake til dette mange ganger). Det er nå på tide å definere en sentral statistisk størrelse - *frekvens* (hyppighet). Innenfor statistikken er frekvens *antall ganger en dataverdi inntreffer* i en måle-

eller observasjonsserie $\underline{M} = \{x_1, x_2, \ldots, x_N\}$. For diskrete data med en gitt utfallskategori S_j vil frekvensen altså defineres som N_j; hvis du kaster en terning ti ganger og får terningkast 5 to ganger vil altså N_5=2. Mer formelt kan man definere frekvensen av observasjoner innenfor en gitt kategori som:

$$N_j = |\underline{M} \in S_j| \quad , \quad j = 1, 2, \ldots, k \tag{1}$$

der $|\ldots.|$ angir antall tellinger av den gitte kategorien innenfor observasjonsserien ($|\ldots.|$ kalles kardinalitet), og k er antall kategorier (k=6 for terningen).

Fra definisjonen av frekvens ovenfor kan vi videre definere begrepet *relativ frekvens*:

$$f_j = \frac{N_j}{N} \tag{2}$$

Den relative frekvensen angir altså andelen hendelser som inntreffer for en gitt kategori. Mange foretrekker å presentere denne i %, men dette er en smaksak. Videre må vi ha at

$$\sum_{j=1}^{k} f_j = 1 \tag{3}$$

siden $\sum_{j=1}^{k} N_j = N$. Med andre ord, summen av de relative frekvensene må bli 1, hvis frekvensene er beregnet for alle observasjonene. Serien $\overset{k}{_{j=1}}\{f_j\} = \{f_1, f_2, \ldots, f_k\}$ kalles i det videre en *normalisert frekvensfordeling*.

Hvis ikke målebetingelsene endres vil frekvensfordelingen nærme seg en 'sann' fordeling når $N \to \infty$. I dette tilfellet er det slik at den normaliserte frekvensen f_j til en gitt kategori vil approksimere *sannsynligheten* p_j - som noen ganger kalles *empirisk sannsynlighet*. Dette er for øvrig en utgave av *de store talls lov*. La oss gå tilbake til terningen for å eksemplifisere, der sannsynligheten for å få en gitt side er p_j=1/6. Men hva hvis denne sannsynlighetsfordelingen ikke var kjent på forhånd? For terningen vet du at du i gjennomsnitt vil få en sekser i løpet av 6 kast, men noen ganger behøves det mange flere kast (og noen ganger færre). Med andre ord, for å estimere sannsynlighetsfordelingen for en terning eksperimentelt må man kaste terningen kanskje 1000 ganger (vi skal simulere dette senere). For en annen type måling eller observasjon vil tilsvarende gjelde, men vi skal ikke lage noen generell retningslinje for hvor mange målinger som er nødvendige for å estimere sannsynlighetsfordelingen med god nøyaktighet.

Frekvensfordeling - histogram

Når vi har en måle- eller observasjonsserie med gjentatte målinger kan en datavisualisering i form av et *frekvenshistogram* være passende. Frekvenshistogrammet er et plott som viser hvordan dataene *fordeler seg*. I sin enkleste form er det et slags søyeldiagram der høyden av hver søyle bestemmer hvor mange observasjoner man har for en gitt kategori av observasjoner (altså frekvensen). X-aksen beskriver kategoriene. Høyden på hver søyle angir altså

frekvensen til et gitt utfall. Hvis du kaster en terning 102 ganger og skal plotte observasjonsserien i form av et histogram, vil altså abscissen ha kategoriene $S = (1, 2, 3, 4, 5, 6)$ (de mulige utfallende på terningen), mens ordinaten vil vise frekvensen. Hvis du - *tilfeldigvis* - har fått like mange av hvert tall på terningen vil søylen for hver kategori ha høyden 17 (fordi 102/6 gir en frekvens lik 17), og fordelingen av observasjoner fremkommer helt flat. Generelt er altså søylehøyden lik frekvensen N_j. Så y-aksen i et frekvenshistogram vil alltid ha benevning a) "Frekvens" (hvis du rapporterer absolutte tall) eller b) "Relativ frekvens" (hvis du rapporterer andel relativt til totalt antall observasjoner; (2). Summerer du alle søylene (altså tar arealet av histogrammet) får du i tilfelle a) det totale antall observasjoner N eller 1 i tilfelle b).

Frekvenstabell / frekvensmatrise

Et alternativ til å plotte frekvenshisotgram er å vise data i en tabell som kalles *frekvenstabell*, som eksplisitt viser frekvensene til hver kategori. Hvis du har *k* kategorier (k=6 for terningen) vil frekvenstabellen kunne lagres som en $k \times 2$ *frekvensmatrise* (rader, kolonner) der første kolonne er kategori S_j og andre kolonne frekvens N_j . Du skal nedenfor generere data og lagre disse i form av en frekvensmatrise før du plotter frekvenshistogrammer.

Tilfeldige tall i simuleringer

På en datamaskin kan vilkårlige tall genereres ved bruk av såkalte "pseudorandom number generators" (PRNGs). Disse algoritmene genererer tallsekvenser som er deterministiske og ikke virkelig tilfeldige, men som er tilstrekkelig komplekse og uforutsigbare til å bli brukt i de fleste tilfeller der man trenger tilfeldige tall. En vanlig metode for å generere tilfeldige tall på en datamaskin er å bruke en seed-verdi (startverdi) som input til en PRNG-algoritme. Seed-verdien kan for eksempel være tiden på datamaskinen, eller en brukerdefinert verdi som blir generert på annen måte. Når PRNG-algoritmen får seed-verdien som input, vil den generere en sekvens med tall basert på en matematisk formel. Hvis man bruker samme seed-verdi flere ganger, vil man få den samme tallsekvensen som resultat hver gang.

Tilfeldige tall brukes i fysikksimuleringer for å modellere tilfeldige hendelser som kan påvirke systemet vi simulerer. For eksempel kan tilfeldige tall brukes til å modellere partikkelspredning eller termisk bevegelse i gasser. Tilfeldige tall kan også brukes til å introdusere usikkerhet i simuleringen, noe som kan være nyttig når vi vil studere hvordan en modell oppfører seg i ulike situasjoner. NumPy inneholder funksjoner for å generere tilfeldige tall via *random*-modulen , som inneholder en rekke funksjoner som kan brukes til å generere tilfeldige tall av forskjellige typer. Modulen bruker PC-klokka til å finne en seed-verdi. Seed-verdien kan settes av brukeren med *random.seed()* - dette gjør at samme tilfeldige tallsekvens generes i hver simulering. Men det anbefales at du normalt *ikke* bruker denne, og lar NumPy automatisk sette seed-verdien. For eksempel kan vi bruke funksjonen *random.rand()* til å generere tilfeldige tall uniformt mellom 0 og 1. Hvis vi vil generere tilfeldige heltall, kan vi bruke funksjonen *random.randint()*. Du kan 'sample' fra normalfordelingen ved å bruke *random.normal()*.

O2 Simuler terningkast og lag frekvenshistogram

Simuler at du kaster terning, feks 102 ganger. Eventuelt kan du ta en virkelig terning, kaste denne et visst antall ganger og lagre måleserien for innlesning i Python. Men det kan være kjekt å sette seg litt inn i simuleringer allerede her. Hint: bruk *numpy.random.choice*. Her må du definere utfallsrommet eller kategoriene ('a') og sannsynligheten ('p') for hvert utfall. Deretter lagrer du hvert terningkast i en liste/vektor. Lag videre en kode der du simulerer en 'jukseterning' som havner på tallet 3 eller 4 to ganger hyppigere en resten av utfallene. Ut fra simuleringene skal du nå telle opp hvor mange treff du fikk for hver side av terningen (beregn frekvens). Lagre data i form av 6×2 frekvensmatriser. Skriv ut matrisene. Sjekk at summen av alle frekvensene blir lik det totale antall terningkast (du kan feks bruke *numpy.sum*). Plott deretter begge frekvensmatrisene i form av histogrammer. ***NB! I denne oppgaven er det ikke lov å bruke innebygde histogramfunksjoner i Python/Numpy/Matplotlib!*** Du kan f.eks. bruke *matplotlib.pyplot.bar* til å plotte histogrammet. Hvordan er fordelingen av terningkastene - er antall enere osv slik som forventet? Hva skjer når du kjører koden på nytt - blir histogrammene like?

```
#Svar
```

O3 Frekvens $\rightarrow$ empirisk sannsynlighet

Histogrammet du genererer ovenfor er opplagt ikke basert på tilstrekkelig med data for at det gjenspeiler den teoretiske frekvensfordelingen. Lag en kode nedenfor der du simulerer hvor mange terningkast som behøves for at den simulerte *relative* frekvensfordelingen skal bli approksimativt lik den teoretiske sannsynlighetsfordelingen. Du skal selv definere et kriterium for hva som defineres som 'approksimativt lik'. Plott både den simulerte og teoretiske fordelingen i samme graf.

```
#Svar
```

Frekvensfordeling, frekvensstetthet og sannsynlighetstetthet for kontinuerlige data

For kontinuerlige data blir en kategorisering slik som for f.eks. sider på terningen ikke like rett fram. Her må du definere en fornuftig måte å kategorisere (*'binne'*) dataene på, og dette via *verdiintervaller*. Normalt velges en fast *søylebredde* Δx slik at *høyden på hver søyle angir antall observasjoner innenfor et gitt intervall*. Men merk at du kan variere Δx slik at bredden på hver søyle varierer, men at dette er ikke så vanlig. Hvis du skal lage et standard frekvenshistogram er det kanskje enklest å starte med å definere et antall søyler k. En optimal k for best visualisering av fordelingen avhenger blant annet av hvor mange observasjoner du har og

hvordan fordelingen av data ser ut. Men du kan f.eks. starte med å prøve $k = \sqrt{N}$. Når *k* er bestemt kan du dermed finne søylebredden ved

$$\Delta x = \frac{(x_{max} - x_{min})}{k} \tag{4}$$

der x_{min} og x_{max} er hhv minimum og maksimum i observasjonsrekken. Kategorien S_j kan nå defineres som et intervall:

$$S_j = [x_{min} + (j-1)\Delta x, x_{min} + j\Delta x > \quad , \quad j = 1, 2, \ldots, k \tag{5}$$

Vi kan definere midtpunktet til hver kategori som

$$x_j = x_{min} + (j - 1/2)\Delta x \tag{6}$$

Over *k* intervaller skal du nå telle *antall observasjoner* fra obervasjonsserien $\underline{M}$ som faller innenfor hvert intervall. Definisjonen blir dermed den samme som for terningen vist ovenfor:

$$N_j = |\underline{M} \in S_j| \quad , \quad j = 1, 2, \ldots, k \tag{7}$$

N_j angir igjen frekvensen til kategori S_j, som er definert av intervallet i (5). Dermed har vi nok en gang at den relative frekvensen blir:

$$f_j^* = \frac{N_j}{N} \tag{8}$$

Vi setter $*$ for å markere at dette er for en kontinuerlig variabel. Alternativt til flyten ovenfor kan du starte med å definere en søylebredde Δx, og fra dette bestemme antall søyler. Du kan også 'klippe' bredden på histogrammet ved å sette x_{max} til en verdi x''_{max} som er mindre enn den globale maksverdien (tilsvarende for x_{min}). Dette kan være hensiktsmessig hvis man feks har 'uteliggere' i datasettet som ikke passer inn sammen med de andre observasjonene. Hvis du klipper histogrammet må imidlertid dette spesifiseres siden du dermed utelater data . En annen mulighet er å legge alle observasjoner > x''_{max} til den svarende søylen osv.

Vi kan nå definere *frekvenstettheten* for en kontinuerlig variabel:

$$f_j = \frac{f_j^*}{\Delta x} \tag{9}$$

Tenk deg at vi nå lar Δx gå mot en infitesimal grense dx. Da blir hver kategori et infitesimalt intervall $[x, x + dx]$ og vi får:

$$f(x)dx = f^*(x, x + dx) \tag{10}$$

$f(x)dx$ angir den inkrementale andelen av obervasjoner i det gitte intervallet. Ved tilstrekkelig med observasjoner ($\lim N \to \infty$) kan vi dermed definere den empiriske sannsynlighetstettheten ved $p(x) = f(x)$. Dermed, en lang dataserie (mange målinger) og små intervaller vil kunne gi deg en god approksimasjon på $p(x)$. I praksis er det et spørsmål om å finne en Δx som dekker tilstrekkelig med observasjoner for å gi et godt estimat på $f(x)$,

men samtidig må ikke Δx være for stor slik at den sanne fordelingen smøres ut. For en kontinuerlig variabel med frekvenstetthet $f(x)$ må vi ha at

$$\sum_{j=1}^{k} f_j \Delta x = 1 \underset{\Delta x \to 0}{=} \int_{x_{min}}^{x_{max}} f(x)dx \tag{11}$$

der x kan anta alle mulige verdier i intervallet $[x_{min}, x_{max}]$ (som ofte, men ikke alltid, settes til $[-\infty, \infty]$). Tilsvarende vil gjelde for $p(x)$. Merk at en variabel bredde Δx_j kan innføres og (11) er fortsatt gyldig.

O4 Lag frekvenshistogram for kontinuerlige data

For måleserien $M_{motstand}$ er det i utgangspunktet ikke noen klar sammenheng mellom måleverdi og målenummer, og vi antar i første omgang at hver måling er uavhenging av de andre (vi skal avsløre en sammenheng mellom måleverdi og målenummer senere). Lag et frekvensmatrise (evt kun en liste med frekvenser) og plott frekvenshistogram for denne måleserien. Du må selv eksperimentere med få finne et antall søyler som viser datafordelingen på en hensiktsmessig måte. Du kan bruke **numpy.amax** e.l.for å finne maksimum og tilsvarende for minimum. Du kan videre benytte elementer fra koden under **O3** for å generere frekvensmatrisen osv. Men hvordan skal verdiene på abscissen se ut - hva slags tall vil du bruke til å representere hver søyles posisjon? Kategorinummer, slik som for terningen, vil ikke være optimalt her. Du får selv tenke ut hvordan dette skal implementeres i koden. Se på denne **linken**. Hva slags sannsynlighetsfordeling tror du måleserien approksimativt følger?

```
#Svar
```

Nå som du har laget din egen histogramkode og du forstår hva et histogram er kan du begynne å bruke innebygde funksjoner i Python/Numpy. En mulighet er **numpy.histogram**. Bruk denne til å generere samme histogram som i koden du laget ovenfor. Får du samme resultat? Hva med verdiene på x-aksen?

```
#Svar
```

Størrelser som beskriver datasett og fordelinger

For en gitt måleserie gir det ikke alltid mening å presentere *alle* dataene. Ofte dropper man også frekvenstabeller/histogrammer. Hva slags tall kan vi bruke til å *representere* måleserien, og som vi rapporterer videre? En viktig størrelse du alltid bør rapportere er *antall observasjoner*. Videre benytter man ofte størrelsene *gjennomsnitt* og *standardavvik*. I det videre skal vi også definere *median*, *persentil* og *variasjonsbredde*.

Gjennomsnitt og forventningsverdi

For en tallserie $\{x_1, x_2, \ldots, x_N\}$ definerer vi gjennomsnittet som:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{12}$$

Dette *aritmetiske* gjennomsnittet definert ovenfor, som er den vanligste formen for gjennomsnitt, kan representere datasettet godt hvis dataene er noenlunde jevnt fordelt rundt $\bar{x}$. Hvis vi går tilbake til frekvensfordelingen ser vi at man kan skrive gjennomsnittet som

$$\bar{x} = \sum_{j=1}^{k} x_j f_j \qquad , \qquad \bar{x} = \int_{x_{min}}^{x_{max}} x f(x) dx \tag{13}$$

for henholdsvis en diskret og kontinuerlig fordeling.

Summenotasjonen over frekvensene f_j er en metode du kan bruke hvis du kun har tilgang på frekvensdata (og ikke rådata), og er også ment som hjelp til å forstå integraldefinisjonen av gjennomsnitt.

Hvis man nå kjenner den sanne sannsynlighetsfordelingen $p(x)$for en gitt variabel x kan *forventningsverdien* defineres som:

$$E[x] = \mu = \int_{-\infty}^{\infty} x p(x) dx \tag{14}$$

Forventningsverdien (kjært barn har mange navn; vi vil bruke både $E[x]$ og μ) er altså det forventede gjennomsnittet av observasjonene, og verdien er altså avhengig av at sannsynlighetstettheten $p(x)$ er kjent.

For en diskret variabel vil gjennomsnitt og forventningsverdi enkelt bli:

$$\bar{x} = \sum_{j=1}^{k} x_j f_j \quad , \quad \mu = \sum_{j=1}^{k} x_j p_j \tag{15}$$

Vi bør ha et teoretisk skille mellom μ og $\bar{x}$, der førstnevnte er den sanne verdien mens sistnevnte representerer et estimat ut fra et gitt forsøk ('sample' på engelsk). Hvis man gjentar forsøket mange nok ganger kan man til slutt få:

$$\mu \underset{N \to \infty}{=} \bar{x} \tag{16}$$

Dette er et eksempel på det som kalles *Store talls lov*. I oppgave **04** ovenfor gjorde vi en variant av dette, der vi brukte et høyt antall 'målinger' (terningkast) for å estimere sannsynlighetsfordelingen. Innenfor eksperimentalfysikken er det er viktig å presisere at (15) *ikke nødvendigvis betyr at du får et korrekt estimat hvis du gjør mange målinger*. Dette kan for eksempel være tilfellet hvis målebetingelsene er feil eller det er systematiske feilkilder i måleoppsettet. Vi skal diskutere dette i mer detalj siden.

Varians og standardavvik

Gjennomsnittet er kanskje den mest rapporterte størrelsen i forbindelse med målinger og simuleringer. Men man ønsker også å vite *spredningen* i datagrunnlaget. Hvis spredningen er veldig stor er det ikke sikkert at gjennomsnittet egentlig er et representativt tall på forventningsverdien. For en enkeltobservasjon x_i kan vi definere *avviket* eller *residualet* i målingen ved $\epsilon_i = x_i - \mu$. Ut fra dette kan vi definere *variansen* som

$$Var[x] = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\epsilon_i^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{17}$$

Variansen er altså det gjennomsnittlig kvadrerte avviket mellom enkeltobservasjoner og gjennomsnittverdi. Ved å bruke tilsvarende argumentasjon som for (13) ovenfor ser vi at variansen kan skrives som:

$$Var[x] = \int_{-\infty}^{\infty}(x-\mu)^2 p(x)dx \qquad , \qquad Var[x] = \sum_{j=1}^{k} p_j(x_j - \mu)^2 \tag{18}$$

for henholdsvis kontinuerlige og diskrete variable. Ved å ekspandere $(x-\mu)^2$, ta integralet ledd for ledd og bruke definisjonen av forventningsverdi i (13) ser vi at variansen blir:

$$Var[x] = E[x^2] - E[x]^2 = E[x^2] - \mu^2 \tag{19}$$

Variansen er et kvadrert utrykk og vil dermed ikke ha samme enhet som målstørrelsen X. Vi definerer *standardavviket* som

$$\sigma = \sqrt{Var[x]} \tag{20}$$

Standardavviket er et mål på spredning og vil alltid være positiv. For en gitt måleserie (som er et subsett av et hele utfallsrommet) vil du kunne estimere varians og standardavvik ved:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{21}$$

Hvorfor $N-1$? Når s estimeres i (20) benyttes $\bar{x}$, som allerede er beregnet ved hjelp av den samme måleserien $\{x_i\}$. Residualet for observasjon k er $\epsilon_k = (x_k - \bar{x})$, og i summen i (20) vil bare $N-1$ av disse residualene være uavhengige størrelser. Hvis du feks har måleserien {1,3,8}, så er $\bar{x}$ =4 og residualene bli {-3, -1, 4}. Med andre ord, det siste residualet er bestemt av de to første. Vi sier derfor at vi har $N-1$ *frihetsgrader* for estimatet av standardavviket, og at s slik formulert i (20) er en *forventningsrett* ('unbiased') estimator. I praksis, ved relativt store N, vil imidlertid ikke denne korreksjonen for antall frihetsgrader bety så mye.

Spesialtilfelle: uniform fordeling

Vi snakket ovenfor om at terningkastet følger en (diskret) uniform fordeling, der $p_i = 1/6$ for alle kategorier i. Men hva med en kontinuerlig variabel X som kan variere i intervallet [a,b] - hvordan blir formen på sannsynlighetsfunksjonen for en uniform fordeling $p(x)$? Hvis alle utfall skal være like sannsynlig (hvilket er definisjonen på *uniform*) for alle verdier x så må $p(x) = konstant = C$. Fra definisjonen har vi:

$$\int_a^b p(x)dx = \int_a^b Cdx = (b-a)C \overset{!}{=} 1 \qquad \rightarrow C = \frac{1}{b-a} \tag{22}$$

Dermed, sannsynlighetstettheten for en uniform fordeling er $p(x) = 1/(b-a)$ innenfor intervallet $[a, b]$. Utenfor intervallet er $p(x)$=0. Forventningsverdien blir:

$$\mu = \int_{-\infty}^{\infty} xp(x)dx = \frac{1}{b-a}\int_a^b xdx = \frac{a+b}{2} \tag{23}$$

Er svaret som du ville forvente? Sjekk at du også får dette svaret når du regner ut integralet!

Videre vil vi ha at variansen er gitt ved:

$$\sigma^2 = E[x^2] - E[x]^2 = \int_a^b x^2 p(x)dx - \mu^2 = \frac{(b-a)^2}{12} \tag{24}$$

Verifiser at utrykket for variansen er korrekt.

Sympy / WolframAlpha

Generelt, når vi skal regne på fordelingsfunksjoner og andre problemer analytisk slik som ovenfor, kan det være kjekt å bruke **Sympy**. Dette er et program for symbolsk algebra, og kan gi deg støtte hvis du feks sitter fast med et integral. Et veldig bra alternativ utenfor Python er **WolframAlpha**.

O5 Standardavvik sammen med histogram

Last inn datasettet $\underline{M}_{motstand}$. Du kan bruke NumPy til å beregne gjennomsnitt, varians og standardavvik (husk $N-1$!). Gjenta histogram-plottet i **O5**, og indiker gjennomsnittet og et intervall $\bar{x} \pm s$ i figuren. Hvor stor andel av dataene er innenfor dette intervallet? Hvor stor andel er innenfor intervallet $\bar{x} \pm 2s$? Lag et plott der du viser andelen av data innenfor $\bar{x} \pm Cs$ der du lar C variere mellom feks 0 og 4. Indiker frekvensene ved $C = \{1, 2\}$. Vis Hvor stor må C være for at du får med alle observasjonene innenfor intervallet?

```
#Svar
```

Din første trendanalyse: Glidende gjennomsnitt og standardavvik

Tenk at du har en måleserie $\underline{M}(x) = \{x_{t_1}, x_{t_2}, \ldots, x_{t_N}\}$ som er en tatt opp over et visst tidsrom $[t_1, t_N]$. Vi vet at gjennomsnittet for hele dataserien kan estimeres ved (11). Men hva hvis dataserien viser en tidsutvikling? Et glidende gjennomsnitt ('moving average') tar gjennomsnittet over et 'vindu' (eller maskevidde, her ΔN) av observasjoner, og flytter vinduet 'glidende' med økende observasjonsindeks (her: tid t_i for indeks i). Vi definerer det glidende gjennomsnittet som:

$$\bar{x}_{g,i} = \frac{1}{\Delta N} \sum_{j=i-\Delta N/2}^{j=i+\Delta N/2} x_j \tag{25}$$

Tiilsvarende kan du definere et standardavvik over samme maske. Dette er en superenkelt eksempel på en *konvolusjon*.

O6 Kode for glidende gjennomsnitt

Lag en algoritme der du leser inn serien $\underline{M}_{motstand}$ og lager et glidende gjennomsnitt over ett visst antall målinger. Finn et passende maskevidde for å ta gjennomsnittet over. Maskevidden bør ikke være for smal (da får du liten glatteeffekt) eller for bred (da vil 'trenden' bli en horisontal kurve svarende til globalt gjennomsnitt). Samtidig skal du estimere standardavviket over samme maskevidde. Plott $\bar{x}_g$ og $\bar{x}_g \pm s_g$ som funksjon av indeks i måleserien.

```
#Svar
```

Kumulativ fordeling

Vi har tidligere definert frekvensfordelingen $f(x)$, der $f(x)dx$ angir andelen av obervasjoner i intervallet $[x, x+dx]$ og alternativt $p(x)dx$ sannsynligheten for et utfall i det samme intervallet. Noen ganger kan det være hensiktsmessig å representere data eller sannsynlighetsfordeling i form av en *kumulativ fordeling*. En slik fordeling viser andelen av data fra en minimumsverdi opp til en gitt verdi x:

$$F(x) = F[X \leq x] = \int_{-\infty}^{x} f(x')dx' \quad , \quad P(x) = P[X \leq x] = \int_{-\infty}^{x} p(x')dx' \tag{26}$$

$F(x)$ *er dermed antall datapunkter i utvalget* X *med verdi mindre enn eller lik* x. $P(x)$ kalles den kumulative sannsynlighetsfordelingen, og viser samlet sannsynlighet for å få et utfall i intevallet $[-\infty, x]$. For fordelingen av kategoriske utfall slik som for terningen blir formuleringen av kumulativ fordeling:

$$P_j = \sum_{i=1}^{j} p_i \tag{27}$$

For terningen husker du at sannsynlighetstettheten var flat; $p_i = 1/6$ for alle i. Den svarende kumulative fordelingen blir nå $P = \{1/6, 2/6, 3/6, 4/6, 5/6, 6/6\}$ $\approx \{0.167, 0.333, 0.500, 0.667, 0.833, 1.000\}$. Med andre ord, for eksempel P_3 forteller om den totale sannsynligheten for å få 1, 2, eller 3 på terningen. Her ser vi også at $F(\infty) = 1.0$ for kontinuerlige variable og $F_{j_{max}} = 1.0$ for kategoriske data der j_{max} er siste (evt. største) kategori. Husk at dette betinger at den underliggende frekvensfordelingen er normalisert.

For å lage et kumulativt histogram for kontinuerlige data kan du for eksempel starte med frekvenshistogrammet, og summere de relative frekvensene for hver kategori. Vi kan kalle dette **metode 1**. Men det fine med fremstilling av en kumulativ fordeling er at du egentlig ikke er avhengig av noen maskevidde Δx slik som for frekvenshistogrammet for kontinuerlige måledata. I praksis kan du sortere alle dataverdier fra lavest til høyest i en liste $\underline{X}_{sort}$ og så registrere hvert datapunkt kumulativt. Hvis du har en måleserie med N målinger kan du dermed lage en kumulativ funksjon $F(x)$ tilsvarende som i (27), der hvert steg (f_i) i den normaliserte funksjonen tilsvarer $1/N$. Dette bestemmer altså F-verdiene til det kumulative histogrammet, uavhengig av hvordan de underliggende dataene fordeler seg. Men formen på det kumulative histogrammet vil fremkomme når du plotter F som funksjon av den sorterte dataserien $\underline{X}_{sort}$. Dette kaller vi **metode 2**.Hvis du for eksempel har måleserien $X = \{4.7, 7.1, 1.3, 11.2, 9.8\}$ så sorterer du denne slik at $\underline{X}_{sort} = \{1.3, 4.7, 7.1, 9.8, 11.1\}$. Ditt svarende kumulative histogram blir dermed $F(x) = \{0.2, 0.4, 0.6, 0.8, 1.0\}$ (fem målepunkter gir 0.2 relativ frekvens i"hopp" for hvert punkt). For eksempel vil andelen av datapunkter med verdier opp til og med $x = 9.8$ dermed bli $F(9.8) = 0.8$ (med andre ord, 80% av dataene har verdier lavere enn eller lik 9.8). Dette eksempelet er basert på veldig få datapunkter, men er altså ment å illustrere hvordan du kan lage det kumulative histogrammet ved hjelp av metode 2. Hvis du sammenlikner de to metodene vil du se at metode 1 gir deg ekvidistante steg i x (definert av Δx), mens metode 2 gir ekvidistante steg i F (der stegene blir $1/N$).

Median og persentiler

Det fine med en kumulativ fordeling er at man også kan få størrelser slik som *median* og *persentiler* direkte. Vi definerer medianen m som verdien som deler et utvalg i to like store deler. Fra definisjonen av den kumulative fordelingen fremkommer medianen som:

$$\int_{-\infty}^{m} f(x)dx = 0.5 \tag{28}$$

Alternativt kan man starte integralet på m og integrere opp til ∞ . I praksis, for et gitt utvalg, kan du sortere utvalget og plukke verdien i midten. Hvis antall verdier i utvalget er et partall kan man summere de to midterste verdiene og dele på 2.

Medianen kalles noen ganger *50-persentilen*. En persentil er verdien en gitt prosentandel av et utvalg er mindre enn eller lik. Med andre ord, persentilen x_p er gitt ved:

$$\int_{-\infty}^{x_p} f(x)dx = P \tag{29}$$

der P kan være mellom 0 og 100 (evt 0 og 1.0 hvis man jobber relativt og ikke i %). Generelt vil vi ha at $x_{p0} = x_{min}$ og $x_{p100} = x_{max}$, det vil si at 100-persentilen defineres av maksimumsverdien i datatsettet. Dette gir mening, fordi 100 % av dataene er inneholdt fra x_{min} til x_{max}. Mer generelt, hvis du har en gitt observasjonsserie kan du ut fra **metode 2** over lage et kumulativt histogram og finne x-verdien der $F(x) = P$, der P er den spesifiserte kumulative prosenten. Ofte vil du ikke treffe nøyaktig på en gitt persentil i utvalget ditt, men da kan du feks bruke *interpolasjon* (tema som kommer senere). I sammenheng med persentiler kan man introdusere første, annen, tredje og fjerde *kvartil* Q_1, Q_2, Q_3, Q_4 som verdiene som definerer 25%, 50%, 75% og 100% av alle dataene. Annen kvartil er lik medianen, og fjerde kvartil er maksimumsverdien. Men første og tredje kvartil er også interessante størrelser, og interkvartilavstanden *IQR* (interquartile range) er definert som

$$IQR = Q_3 - Q_1 \tag{30}$$

Denne kan brukes som et alternativt mål på spredning; jfr. standardavviket ovenfor.

O7 Kode for kumulativ fordeling

Last inn datasettet $\underline{M}_{motstand}$. Lag din egen kode for å genere et kumulativt normalisert histogram. Du skal først ta utgangspunkt i frekvenshistogramkoden du lagde i **O5**, og bruke **metode 1** til å lage det kumulative histogrammet. Deretter skal du bruke **metode 2**, og her kan bruke for eksempel **numpy. sort** for å sortere datasettet. Husk at y-verdiene fremkommer enkelt ved steg $1/N$, mens x-verdiene er din sorterte liste $\underline{X}_{sortert}$. Presenter de kumulative histogrammene i et plott. Kommenter forskjellene, og hvilken metode du foretrekker. Ut fra ditt kumulative histogam beregnet med metode 2 skal du estimere median og første og tredje kvartil (**ikke** bruk ferdigfunksjoner i Numpy!). Spørsmålet er her hvordan du finner x-verdiene som svarer til de gitte prosentene (25, 50, og 75%)? En mulighet er å bruke 'smart' indeksering via **numpy.argwhere**, siden det kumulative histogrammet F og $X_{sortert}$ har korresponderende indekser. Med andre ord, finn feks indeksen ind_{25} der F er lik 0.25 og plukk ut svarende x-verdi i $\underline{X}_{sort}$. Dette vil gi deg $Q1$ osv. Når du har laget dette fra 'scratch' kan du deretter sammenlikne estimatene dine med feks **numpy.percentile**. Plott deretter m, $Q1$ og $Q3$ inn i samme figur. Estimer IQR. Hvor stor andel av data finnes innenfor IQR sammenliknet med intervallet $\bar{x} \pm s$?

```
# Svar
```

Kovarians og korrelasjon

La oss si vi har to måleserier $\underline{X}$ og $\underline{Y}$ med forventningsverdi μ_X og μ_Y tatt opp i forbindelse med et eksperiment. Vi kan kvantifsere graden av samvariasjon, eller kovariansen, med:

$$Cov[X, Y] = E[(X - \mu_x)(Y - \mu_y)] \tag{31}$$

Det går an å vise (se f.eks. likn. (18)) at:

$$Cov[X,Y] = E[XY] - \mu_x \mu_y \tag{32}$$

Videre har vi:

$$E[XY] = \frac{1}{N^2} \sum_{j=1}^{N} \sum_{i=1}^{N} x_i y_j \tag{33}$$

Hvis $\underline{X}$ og $\underline{Y}$ er uavhengige størrelser ser vi at $E[XY] = \mu_x \mu_y$, og dermed $Cov[X,Y] = 0$.

Kovariansen er imidlertid proporsjonal med skalaene til $\underline{X}$ og $\underline{Y}$ slik at resultatet ikke er så lett å tolke. *Korrelasjonen* defineres ved:

$$\rho_{X,Y} = \frac{Cov[X,Y]}{\sigma_x \sigma_y} \tag{34}$$

Korrelasjonen vil dermed være et tall som er uavhengig skalaene til variablene som inngår. Hvis vi har $\underline{X} = \underline{Y}$ ser vi at vi kan ta (25) via (18) og vise at $\rho_{X,Y} = 1$, mens $\underline{X} = -\underline{Y}$ gir $\rho_{X,Y} = -1$. Dette markerer hemholdsvis maksimum og minimum korrelasjon. Korrelasjonen kan være en veldig fin metode å evaulere sammenheng mellom variable på før man bruker mer avanserte analyseverktøy.

O8 Kode for *autokorrelasjon*

Autokorrelasjon er et begrep som estimerer til sammenhengen mellom verdier av en variabel over tid. Autokorrelasjon oppstår når verdier av en variabel i en serie er avhengig av tidligere verdier i samme serie. Med andre ord, en variabel er autokorrelert når den viser en tendens til å korrelere med seg selv på forskjellige tidspunkter. For eksempel, hvis du måler temperaturen i et bestemt område hver dag, kan du se at temperaturen i dag har en tendens til å korrelere med temperaturen i går og dagen før det. For en tidsserie $\underline{X}_t = \{X_{t1}, X_{t1}, \ldots, X_{tn}\}$ vil autokorrelasjonen defineres som:

$$\rho_{X_t, X_{t+\Delta t}} = \frac{Cov[X_t, X_{t+\Delta t}]}{\sigma^2_{x_t}} \tag{35}$$

der Δt er en forflytning i tid langs tidserien. Lag kode som beregner autokorrelasjonen for en gitt tidsserie, der Δt varieres. Lag en sinus-tidsserie som du tester koden på. Test også koden på $\underline{M}_{motstand}$. I begge tilfeller skal $\rho_{X_t, X_{t+\Delta t}}$ plottes som funksjon av Δt. *NB!* Du får ikke lov til å bruke ferdigfunksjoner i Python som beregner kovarians, korrelasjon eller autokorrelasjon.

```
#Svar
```

Sammensatt varians i funksjon av stokastiske variable

Tenk at vi måler to serier $\underline{X}$ og $\underline{Y}$ og estimerer en størrelse F ut fra disse målingene; $F(x, y)$. Hvordan kan vi nå estimere variansen i F? Vi kan starte med å approksimere F ved Taylorutvikling rundt gjennomsnittet $\bar{F}$:

$$F(x, y) \approx \bar{F} + \frac{\partial F}{\partial x}(\bar{x} - x) + \frac{\partial F}{\partial y}(\bar{y} - y) \tag{36}$$

Hvis vi nå ser på variansen i F:

$$s_f^2 = \frac{1}{N}\sum (F_i - \bar{F})^2 \tag{37}$$

og kombinerer dette med (36) får vi:

$$s_f^2 \approx \left(\frac{1}{N}\sum \frac{\partial F}{\partial x}(\bar{x} - x_i) + \frac{1}{N}\sum \frac{\partial F}{\partial y}(\bar{y} - y_j) \right)^2 \tag{38}$$

og videre manipulering vil gi den 'sammensatte' variansen:

$$s_f^2 \approx \left(\frac{\partial F}{\partial x} s_x \right)^2 + \left(\frac{\partial F}{\partial y} s_y \right)^2 + 2 \frac{\partial F}{\partial x} \frac{\partial F}{\partial y} s_{xy} \tag{39}$$

der s_{xy} er estimert kovarians mellom $\underline{X}$ og $\underline{Y}$.

Likning (39) er et viktig resultat, fordi den sier at man kan summere variansene av $\underline{X}$ og $\underline{Y}$ vektet med endringen av den resulterende funksjonen mhp hver variabel + en kovarians-term. Hvis man antar at de stokastiske variablene er uavhengige kan man videre generalisere til en funksjon med m variable:

$$s_f^2 = \sum_{k=1}^{m} \left(\frac{\partial F}{\partial x_k} s_{x_k} \right)^2 \tag{40}$$

Denne skal vi bruke i såkalt feilpropagering fremover.

Eksempler

Tenk at vi har funksjonen $f = ax + b$. Dette kan for eksempel være en relasjon mellom luft-temperatur og høyde på kvikksølvet i et termometer. Variansen i f er dermed $s_f^2 = (as_x)^2$, der s_x^2 er variansen i kvikksølvhøyden. I dette tilfellet er ikke resultatet en approksimasjon (kan sjekkes ved å bruke definisjonen av varians). En annen funksjon kan være $f = ax + by + c$, og variansen blir her $s_f^2 = (as_x)^2 + (bs_y)^2$. Dette er også eksakt så lenge x og y er uavhengige. Dermed, hvis du for eksempel summerer massen av to lodd du har målt vekten på (og estimert

tilhørende standardavvik), der m_X=1.00 kg, s_X=0.02 kg og m_Y=2.00 kg , s_Y=0.03 kg, så vil F =3 .00 kg (trivielt) mens $s_f = \sqrt{(s_x^2 + s_y^2)}$=0.04 kg. Vi skal komme tilbake til slike *usikkerhetsberegninger* senere.

Varians i gjennomsnitt

Standardavviket gir et estimat på *presisjonen* - variasjonen i enkeltmålinger- men sier egentlig ikke noe om hvor sikre vi kan være på estimatet av gjennomsnittet (se neste kapitell for mer om presisjon). Husk at gjennomsnittet er $\bar{x} = \sum x_i/N = (x_1 + x_2 + \ldots + x_N)/N$. Relasjon (40) kan dermed generaliseres til en sum over mange målinger der hver måling x_i kommer fra samme fordeling med varians s_x^2. Dermed, hvis vi nå ønsker å estimere variansen i gjennomsnittet:

$$s_m^2 = \sum_{i=1}^{N} \left(\frac{s_x}{N}\right)^2 = \frac{s_x^2}{N} \tag{41}$$

der indeks m indikerer 'mean'. Dette betyr at presisjonen i gjennomsnittsestimatet forbedres med kvadratroten av antall målinger, $s_m \sim 1/\sqrt{N}$, som er et viktig resultat. Det er videre viktig å presisere at for en serie med enkeltmålinger gir standardavviket variasjonen over enkeltmålingene, mens standardavviket i gjennomsnittet gir forventet variasjon i gjennomsnittet hvis målingene gjentas.

O9 Glidende gjennomsnitt med standardavvik i gjennomsnitt

Her skal du fortsette på koden du lagde i **O6**. Plott estimer standardavviket i middelverdien $s_{m,g}$ over hver Δx og plott $\bar{x}_g \pm s_{m,g}$ som funksjon av indeks i måleserien sammen med det du plottet i **O6**.

```
#Svar
```

Tilleggsoppgaver

T1 Lag en kode der du simulerer en femkantet terning. Generer frekvensfordelingen.

T2 Lag en kode slik som i O4, men der søylebredden Δx øker med avstanden fra gjennomsnittsverdien.

T3 Last inn dataserien *outlier.txt*. Beregn gjennomsnitt, standardavvik og for eksempel 1- og 99-persentil. Gjør en vurdering om serien inkluderer uteliggere; diskuter. Plott gjerne data og forklar.

T4 Bruk definisjonene av medianen m slik som i (28), og vis at m blir lik forventningsverdien μ for en uniform fordeling.

T5 Du har en måleserie på 10 målinger som gir deg en presisjon i gjennomsnittet på 5%. Hvor mange målinger må du gjøre for å få presisjonen i gjennomsnittet ned til 0.5%?

T6 Hvor stor andel av et datasett befinner seg innenfor IQR?

T7 Hvorfor er ikke gjennomsnittlig feil, altså $\overline{\epsilon - \mu}$, et godt mål på spredning?

T8 Du har målingene 0.655 ± 0.024, 0.590 ± 0.080 og 0.789 ± 0.071 (standardavvik for hver størrelse angitt). Estimer summen av målingene og tilhørende standardavvik.