1. Innledning.

1.1. Hva er statistikk?

Vi kommer daglig i kontakt med statistikk i en eller annen form. Bare tenk på alle de meningsmålingene som vi bombarderes med. Eller ta for deg sportssidene i en avis. De er faktisk fulle av statistikk.

Det er vanlig at man fatter beslutninger på grunnlag av statistiske analyser. Da må man vite hvordan man utfører slike analyser. Men man må også vite hvordan man skal tolke resultatene av slike analyser, og man må innse hvilke usikkerheter og begrensninger slike analyser har.

I dette kurset skal du først og fremst lære deg hvordan du kan utføre slike analyser. Men du må også lære deg være kritisk til påstander basert på statistikk.

Statistiske analyser baserer seg på **sannsynlighetsregning**. Dette er et tema som mange synes er "vanskelig". Men sannsynlighetsregning er også et spennende tema som du forbløffende ofte kommer i kontakt med. Det er derfor bryet verdt å sette seg inn i de grunnleggende prinsippene for sannsynlighetsregningen.

Statistikk og sannsynlighetsregning har en egen terminologi som i begynnelsen kan virke mystisk. Er det nødvendig med alle disse rare begrepene? Men en av hovedgrunnene til at statistikk så ofte feiltolkes eller misbrukes, er at viktige begreper har en upresis betydning i dagligtalen. Det er derfor nødvendig å banke fast en mest mulig klar og entydig terminologi så snart som mulig.

La oss gå rett på sak med to grunnleggende begreper:

- En ***populasjon*** er en samling objekter (personer, biler, målinger, karakterer osv. osv.) som vi ønsker å si noe om. Som regel er det ikke praktisk mulig å undersøke alle objektene i populasjonen.
- Et ***utvalg*** består av *noen* av objektene fra populasjonen. Vanligvis ønsker vi oss et *tilfeldig* utvalg, d.v.s. at vi lar utvalget bestå av *tilfeldige* objekter fra populasjonen.

Og så til poenget:

Svært ofte ønsker vi å skaffe oss informasjon om egenskaper ved *populasjonen* på grunnlag av tilsvarende egenskaper ved *utvalget*.

Eksempel 1.1: Ved en meningsmåling om norsk medlemskap i EU spørres 1000 norske velgere hvordan de ville stemt. Da viser det seg at 35 % av disse er for norsk medlemskap i EU, mens 40 % er mot. Hva er *populasjonen*, og hva er *utvalget* i denne meningsmålingen?

Løsning: Populasjonen er alle norske velgere. De 1000 som blir spurt i målingen, er *utvalget.*

Strengt tatt viser undersøkelsen bare at 35 % av personene i *utvalget* ville stemt for og 40 % mot norsk medlemskap i EU. Men dersom utvalget er "riktig" sammensatt, vil de utgjøre et representativt utsnitt av alle norske velgere. Da kan man med en viss sannsynlighet si at *populasjonen* (alle norske velgere) vil fordele seg på *omtrent* samme måte.

Dersom en journalist prøvekjører en ny bilmodell, vil *utvalget* ofte bestå av en eneste bil. På grunnlag av erfaringene med dette meget begrensede utvalget bedømmer da journalisten *alle* biler av denne modellen. *Populasjonen* består altså av *alle* biler av denne modellen som er produsert (og ofte også biler som *vil bli* produsert en gang i fremtiden).

All sunn fornuft sier oss at resultatene fra et utvalg bare vil gi oss *tilnærmede* verdier for de resultatene vi ville fått dersom vi hadde undersøkt hele populasjonen. Vi sier at vi får *estimater* for de tilsvarende størrelsene i populasjonen. Et fundamentalt spørsmål er da: Hvor god er disse estimatene? Hvor stor er usikkerheten når vi overfører resultatene fra utvalget til hele populasjonen? Her er det at sannsynlighetsregningen kommer inn i bildet.

En vanlig statistisk oppgave er å sammenlikne to populasjoner. I praksis sammenliknes ofte *utvalg* fra de to populasjonene. Hvis vi får forskjeller mellom disse utvalgene, kan vi da være sikre på at det også er forskjeller mellom *populasjonene*, eller er disse forskjellene bare tilfeldige utslag som vi alltid må regne med i slike undersøkelser?

Vi har nå (kort og overfladisk) sett på noen av de mange problemstillingene der statistikk kommer inn i bildet. For å kunne handtere slike problemstillinger, må vi kunne:

1. Samle inn data, bearbeide data, og presentere data.
2. Benytte statistiske metoder til å vurdere resultatene.

Punkt 1 ovenfor utgjør det vi kaller ***deskriptiv (beskrivende) statistikk***.
Punkt 2 består bl.a. av ***estimering*** og ***hypotesetesting***, og krever kjennskap til sannsynlighets-regning.

I dette kurset skal vi derfor starte med deskriptiv statistikk. Deretter ser vi på sannsynlighets-regning før vi avrunder med estimering og hypotesetesting.

1.2. Data.

Vi har brukt uttrykket "data" flere ganger allerede. Det er på tide å definere begrepet:

- **Data** er observasjoner (for eksempel målinger, antall, kjønn, farge, osv. osv.) som er samlet inn.

En samling rådata er lite nyttig for oss. Dataene må bearbeides, slik at vi kan danne oss et inntrykk av viktige egenskaper ved det utvalget som dataene er hentet fra. Vi definerer:

- En **parameter** er en tallstørrelse som beskriver en eller annen egenskap ved en *populasjon*. Eksempel: gjennomsnittsverdi, spredning rundt gjennomsnittet, andel av populasjonen som har en bestemt egenskap, osv.

Merk at begrepet *parameter* egentlig beskriver en egenskap ved *populasjonen*. Ofte har vi bruk for å angi egenskaper ved et *utvalg*. På engelsk brukes da begrepet "**statistic**". Jeg kjenner ikke noe tilsvarende godt norsk uttrykk. Noen ganger brukes "observator". Vi kan også bruke uttrykk som "utvalgs-parameter" og populasjons-parameter".

Eksempel 1.2: Dersom 40 % av de som spørres i en spørreundersøkelse sier NEI til norsk medlemskap i EU, er dette tallet da en parameter?

Løsning: Nei, dette tallet er en *observator* siden det er basert på et *utvalg*. Men dersom 40 % av alle velgerne svarer NEI i en folkeavstemning, så må 40 % (eller 0.40) oppfattes som en *parameter*. (Dette forutsetter egentlig at alle velgerne avgir stemme.)

La oss se nærmere på hva slags data vi kommer i kontakt med. Først kan vi skille mellom *kvantitative* og *kvalitative* data:

- ***Kvantitative*** data kan angis ved en eller annen *tallverdi* (målestørrelse, antall).
- ***Kvalitative*** data kan angis ved en ikke-numerisk verdi (kjønn, farge, nasjonalitet …).

I dette kurset skal vi (nesten) bare befatte oss med kvantitative data. Slike data kan videre deles inn i to hovedgrupper:

- ***Kontinuerlige*** data kan anta en hvilken som helst verdi (i praksis innenfor et visst område). Eksempler kan være høyde, vekt, temperatur, fart, strekning osv. osv.
- ***Diskrete*** data kan kun anta visse bestemte verdier, vanligvis heltall. Når vi teller opp antall objekter som har en bestemt egenskap, får vi diskrete data.

For eksempel vil antall personer som står inne i en heis være *diskrete* data. Samlet vekt til disse personene er *kontinuerlige* data.

Rent matematisk må kontinuerlige og diskrete data handteres på ulike vis. Det er derfor viktig å skille mellom disse typene.

1.3. Vær kritisk!

Det sies at det fins tre typer løgn: uskyldig løgn, grov løgn og statistikk. Faktisk er det neppe noen vitenskap som er så lett å misbruke som nettopp statistikk. Gjennom hele kurset skal vi legge stor vekt på å unngå feil bruk av statistikk. Men noen vanlige feil er så elementære at vi

kommer langt med bare sunn fornuft og kritisk sans. La oss se på noen vanlige feil som alt for ofte forekommer i praksis.

Den vanligste feilen er nok ***bruk av uegnede utvalg***. Noe av det verste er utvalg der respondentene (de som svarer) selv avgjør om de vil være med i utvalget eller ikke (eksempel: TV-program der seerne stemmer ved å ringe eller sende SMS). Resultatene fra slike "utvalg" er sjelden representative for en større populasjon. Slike avstemninger kan nok brukes til å lage useriøse oppslag i avisene. Men det er også omtrent det eneste de kan brukes til.

Det er faktisk temmelig vanskelig å skaffe gode utvalg. Skal vi gå ut på gata og kapre folk? Da vil befolkningsgrupper som sjelden går rundt på gata bli underrepresentert. Skal vi plukke ut folk tilfeldig fra telefonkatalogen og ringe dem opp? Da vil grupper som ikke står i telefonkatalogen og grupper som er vanskelig å treffe på telefon bli underrepresentert. Hvis du skal undersøke helsetilstanden i et fiskeoppdrett og skaffer deg et utvalg ved å håve opp noen tilfeldige fisk, kan du risikere at friske fisk lettere spretter ut av håven din enn syke fisk. Da blir det en større andel syke fisk i utvalget ditt enn i populasjonen (som består av alle fiskene i anlegget). Det er utviklet mange forskjellige teknikker til å sette sammen gode utvalg, og til å korrigere for at utvalg ikke er "perfekte". Vi skal ikke komme inn på slike teknikker her.

En annen vanlig feil er å bruke ***for små utvalg***. Jo større utvalget er, jo mer sikker kan vi være på at resultatene fra utvalget er representative for populasjonen. Vi skal etter hvert se på hvor store slike utvalg bør være.

Man skal være svært forsiktig med å påstå at man har påvist en *årsakssammenheng* i det øyeblikk man har funnet en *statistisk sammenheng* mellom to størrelser. Slike statistiske sammenhenger kan skyldes tilfeldigheter. Eller det kan være helt andre sammenhenger enn den som framgår direkte av den statistiske undersøkelsen. Et par eksempler:

- I mange europeiske land er det en klar *statistisk* sammenheng mellom reduksjon i storkebestanden og reduksjon i fødselstall. Har man da påvist at antall storker har betydning for fødselstallet?
- Flere ulike undersøkelser tyder på at barn av foreldre med høy inntekt gjør det bedre på skolen enn barn av foreldre med lav inntekt. Har man da påvist at foreldrenes *inntekt* er viktig for om barna lykkes på skolen? Eller kan det være bakenforliggende årsaker som påvirker både foreldrenes inntekt og barnas suksess på skolen?

En annen vanlig feilslutning går ut på at dersom man kan påvise en statistisk sammenheng mellom to egenskaper i et utvalg, så gjelder denne sammenhengen for *alle objektene* i den populasjonen som utvalget er hentet fra. Hvis man for eksempel påviser en statistisk sammenheng mellom foreldres inntekt og barns suksess på skolen, så betyr ikke det at *alle* foreldre med høy inntekt har barn som gjør det godt på skolen, og at *alle* foreldre med lav inntekt har barn som blir skoletapere.

Statistikk er et svært nyttig redskap. Men da må både de som foretar undersøkelsen og de som vurderer resultatene ha visse grunnleggende kunnskaper. Formålet med dette kurset er å gi deg slike kunnskaper.