Kort om statistikk

Philip Haugen

August 22, 2024

1 Data

Data er en samlebetegnelse for informasjon som observeres i verden. Denne informasjonen kan være **kvantitativ** (numerisk) eller **kvalitativ** (beskrivende og basert på ord). Data kan komme i mange former, som for eksempel:

- Diskret datatype: representert av heltall $\mathbf{Z}$. Eksempler av diskret data er binære data, representert av 0 eller 1, og kategorisk data, representert av flere enn to grupper.
- Kontinuerlig datatype: representert av alle reelle tall $\mathbf{R}$. Brukes for data som kan ta enhver verdi innenfor et intervall. Et eksempel på kontinuerlig data er overlevelsesanalyse.

Det finnes mange forskjellige datatyper, men alle er ofte utsatt for **tilfeldighet**. Tilfeldighet gjør det utfordrende å lage presise prediksjoner basert på data. Spørsmålet blir derfor: Hvordan kan man lære noe nytt fra data til tross for tilfeldighet?

Svaret er statistikk, som betyr å samle og analysere data for å finne tilnærminger. Selv om tilfeldighet kan virke kaotisk og ukontrollert, så går det an å anta at denne tilfeldigheten har en underliggende struktur så man kan lage statistiske tilnærminger.

2 Sannsynlighetsteori

Det sentrale objektet som brukes for å representere data er den tilfeldige (stokastiske) variabelen $\mathbf{X}$. Denne variabelen kan ta forskjellige verdier med forskjellige sannsynligheter. Tilfeldige variabler har spesielle funksjoner som beskriver sannsynligheten tilknyttet hver eneste verdi. Denne sannsynligheten kalles for **tetthetsfunksjon**.

Formen til en tetthetsfunksjon beskriver nettopp strukturen bak tilfeldigheten i dataen. Fordelingen kan ta mange former, men de mest vanlige er:

- Uniform fordeling: alle verdier har samme sannsynlighet. Dette betyr at hver mulig utfall har lik sannsynlighet for å skje, og fordelingen er derfor flat.
- Bernoulli-fordeling: beskriver en fordeling for binære data, og viser sannsynligheten for at 1 (suksess) skjer. Det vil si at den har kun to mulige utfall, 0 (fiasko) og 1 (suksess), med en gitt sannsynlighet for suksess.
- Binomisk fordeling: ligner på Bernoulli, men viser sannsynligheten for flere binære utfall over et fast antall forsøk.
- Poissonfordeling: beskriver antall hendelser som inntreffer innenfor et fast tidsrom eller område, med en kjent gjennomsnitt.
- Normalfordeling: beskriver en kontinuerlig, symmetrisk fordeling som følger gausskurven. Den har sitt toppunkt på gjennomsnittet, og variansen bestemmer spredningen.

Nå er ikke tetthetsfunksjon den eneste måten å beskrive tilfeldigheten i dataen, det finnes også **kumulativ fordelingsfunksjon**. Denne brukes for å finne kvantiler eller prosentiler til en tilfeldig variabel.

Det finnes også andre nyttige verdier for å beskrive tilfeldig data. For eksempel, hva er en typisk verdi for en tilfeldig variabel? Det som er typisk blir funnet i målingen av **sentralmål**, som kan være:

- Gjennomsnitt: den forventende verdien av sentralmålet.

$$\bar{x} = \frac{S_x}{n}$$

- Median: midtpunktet i en kumulativ fordelingsfunksjon.
- Typetall: verdien som finnes på toppunktet av en tetthetsfunksjon.

Det som også er interessant er spredningen i verdier en tilfeldig variabel kan ha. Dette heter **spredningsmål**:

- Varians: gir en antydning på hvor spredningen til verdiene er fra det forventede gjennomsnittet. Det finnes to typer: utvalgsvarians og populasjonsvarians.
- Standardavvik: spredningen til dataen utifra den originale dataen. Standardavviket er alltid kvadratroten av variansen.

Formmål forteller mer spesifikke detaljer om formen til en sannsynlighetsfordeling. Skjevheten viser hvor ubalansert fordelingen mot en side eller annen. Skjevheten antyder at det finnes **ekstremverdier** i en fordeling.

3 Bayes' teorem

Bayes' teorem er et prinsipp i sannsynlighetsteori som gir en beregningsmetode for å oppdatere sannsynligheten for en hendelse utifra ny informasjon. De sentrale begrepene er:

- **Betinget sannsynlighet** $P(A|B)$: Sannsynligheten for hendelse A, gitt hendelse B.
- **Prior sannsynlighet** $P(A)$: Sannsynligheten for at hendelse A skjer før ny informasjon.
- **Posterior sannsynlighet** $P(A|B)$: Sannsynligheten for hendelse A etter hensyn til informasjonen om hendelse B.
- **Likelihood** $P(B|A)$: Sannsynligheten for at hendelse B observeres, gitt A.
- **Marginal sannsynlighet** $P(B)$: Sannsynligheten for hendelse B, uavhengig av A.

Bayes' teorem er som følger:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Dette teoremet gir en måte å oppdatere **priorer**, altså tidligere antagelser, basert på ny data. Hvis ny informasjon oppstår som endrer oppfatningen av sannsynligheten, kan Bayes' teorem brukes til å beregne den **posterior sannsynligheten**, som er en oppdatert sannsynlighet etter å ha observert dataen.

En av de viktigste bruksområdene for dette er i **Bayesiansk inferens**. I motsetning til frekventistisk statistikk, hvor det brukes p-verdier til å avgjøre om observasjonen er statistisk signifikant i forhold til en hypotese om en parameterverdi, bruker Bayesiansk statistikk priorinformasjon kombinert med data for å beregne en posteriorfordeling som gir en sannsynlighetsfordeling over parameteren. Det vil si at er både parametere og data blir sett på som tilfeldige variabler, og inferens gjøres gjennom å beregne en **posterior fordeling** over parametrene basert på tidligere informasjon og ny data. Dette blir sett mer på i detalj i de neste seksjonene.

4 Frekventistisk inferens og hypotesetesting

I statistikk begynnes det med å definere en populasjon, som er en delmengde av en større superpopulasjon som undersøkes. Siden populasjonen ofte er vanskelig å observere direkte, samles data fra et utvalg. Deretter antas det at dataene

følger en sannsynlighetsfordeling eller en matematisk formel. Målet er å "oversette" et aspekt av populasjonen til en parameter innenfor denne modellen. Siden populasjonen ikke kan observeres direkte, er også denne parameteren ukjent. Prosessen med å estimere denne parameteren kalles inferens, da det forsøkes å trekke konklusjoner om den ukjente populasjonen basert på utvalgsdata.

4.1 Estimater

En **estimator** er en funksjon av utvalgsdataen som gir et anslag for en ukjent parameter. En god estimator er:

- Utvalgsgjennomsnitt: Denne har mange gode kvaliteter som estimator. Ifølge **store talls lov** vil utvalgsgjennomsnittet nærme seg gjennomsnittet for populasjonen når det samles store nok mengder data, noe som betegnes som **consistency**.

Ettersom utvalg er tilfeldige, er også utvalgsgjennomsnittet tilfeldig, og dermed er estimatorer i tillegg også tilfeldige. For å forstå egenskapene til en estimator, er det viktig å forstå dens fordeling, kjent som **utvalgsfordeling**. Denne fordelingen lar sannsynligheten bli vurdert for at en gitt estimator oppnår en bestemt verdi.

4.2 Asymptotiske teoremer

To viktige teoremer i statistikk er **sentralgrenseteoremet** og **store talls lov**. Sentralgrenseteoremet sier at en funksjon av utvalgsgjennomsnittet er normalfordelt når utvalget er stort nok, og derfor er den asymptotisk.

4.3 Hypotesetesting

I statistikk så oversettes antagelser om en populasjon til en påstand om verdien til en parameter. Utvalgsfordeling er viktig for å forstå hypotesetester. For å teste disse påstandene, brukes det hypotesetester:

- **Nullhypotese (H_0)**: Antydning om at det ikke er noen forskjell.
- **Alternativ hypotese (H_1 eller H_a)**: Antydning om at det finnes en forskjell.

Eksempel: I en studie som sammenligner en gruppe med pasienter som får placebo med en gruppe som får medisiner, vil nullhypotesen være at det ikke er noen forskjell mellom gruppene.

$$H_0 : \theta = 0$$

Når mer data blir samlet inn, må det avgjøres om nullhypotesen skal forkastes eller ei. Dette blir gjort ved å beregne en p-verdi, som er sannsynligheten for å

observere utvalgsgjennomsnittet eller mer ekstreme verdier gitt at nullhypotesen er sann. Hvis p-verdien er lav nok, for eksempel under et signifikansnivå på 0.05, antyder det at nullhypotesen sannsynligvis er feil. Her kan man også bruke konfidensintervaller, som er et "spekter" av parameterverdier som realistisk kan ha fått utvalgsgjennomsnittet. Hvis dette intervallet ikke inkluderer verdien foreslått av nullhypotesen, kan nullhypotesen forkastes.

Rent konkret for dette eksempelet, anta at det er 100 pasienter, hvor 50 får placebo og 50 får medisiner. Hvis 5 av de 50 som fikk medisinen, har en positiv respons, mens ingen av de 50 som fikk placebo gjør det, vil en statistisk test som for eksempel kjikvadrattest gi en p-verdi på omtrent 0.02.

Siden p-verdien er lavere enn signifikansnivået på 0.05, kan nullhypotesen forkastes. Konfidensintervallet for forskjellen i responsen mellom gruppene kan være (0.02, 0.10), som ikke inkluderer 0. Dette viser også at nullhypotesen kan forkastes.

Etter at det har blitt tatt en beslutning om nullhypotesen, kan det skje to typer feil:

- **Type I feil**: Nullhypotesen er korrekt, men den blir forkastet (falsk positiv).
- **Type II feil**: Det omvendte; nullhypotesen er feil, men blir ikke forkastet (falsk negativ).

Paradokset her er at når sannsynligheten av en av disse feilene minimeres, så øker den andre. Løsningen på dette er å definere en så lav sannsynlighet (0.05 eller 0.01) som kan tolereres. Som sagt er denne sannsynligheten signifikansnivået for p-verdien.

4.4 Statistiske tester

Valget av statistisk test avhenger av hva dataen inneholder:

- **One-sample test**: Brukes for å karakterisere en enkelt populasjon.
- **Two-sample test**: Brukes for å sammenligne to populasjoner.

Ifølge sentralgrenseteoremet vil utvalgsfordelingene for slike tester være normalfordelte, og derfor kan brukes z-skår. Z-skår blir brukt for å betegne en standard normalvariabel med null i gjennomsnitt og enhetsvarians. Z-skår antyder at populasjonsvariansen er kjent, men dette er urealistisk i praksis. Hvis denne variansen må estimeres så konverteres Z-skåren til Students t-test som kommer fra Students t-fordeling. Hvis det sammenlignes tre grupper så bruker man variansanalyse, også kjent som ANOVA, for å sjekke om alle har samme gjennomsnitt. Hypotesetestene Z, t, og ANOVA vil at man bruker kontinuerlig data eller store nok utvalgsstørrelser.

Når det gjelder binær data så kan man bruke en krysstabell og utføre en kjikvadrattest. De tre nevnte hypotesetestene er univariate analyser siden de fokuserer på single, tilfeldige variabler. For å undersøke forhold mellom variabler brukes regresjonsanalyser. Lineær regresjon viser hvordan en variabel påvirker en annen, mens en generalisert lineær modell kan brukes hvis utdataen er binær. Estimering av parametere i regresjonsmodeller gjøres ofte med **sannsynlighetsmaksimering**.

5 Bayesiansk inferens og hypotesetesting

Denne tilnærmingen til hypotesetesting handler om bruken av **a priori**, og gir på mange måter en bedre metode for å håndtere tilfeldighet på.

5.1 Estimater

For å anslå den ukjente parameteren θ, brukes det en modell som beskriver observasjonene. Punktestimatet er det beste anslaget som kan gis for verdien av θ, representert som en verdi.

De tre viktigste punktestimatene er:

- **MAP (Maksimum A Posteriori) punktestimat**: punktestimatet for θ som maksimerer sannsynlighetsfordelingen.
- **Medianen**: $\tilde{\theta}$ minimerer $l(t)$ fordi den gir det minste gjennomsnittlige avviket i absolutt verdi når den sammenlignes med den tilfeldige variabelen θ.
- **Forventningsverdi**: den forventede verdien av θ gitt dataene, som er:

$$\mu_\theta = E[\theta]$$

Intervallestimatet representerer intervallet hvor parameteren θ har en viss sannsynlighet gitt dataene. Hvis $P(a < \theta < b \mid \mathbf{X}) = 0.95$, så er $[a, b]$ et 95% kredibilitetsintervall for θ. Både punktestimatene og intervallestimatet brukes for å finne estimater til forskjellige sannsynlighetsfordelinger.

5.2 Hypotesetesting

Her brukes en nyttefunksjon, $u_\theta(x)$, for å finne verdien av ulike utfall gitt en parameter θ og en observasjon x. Når man ser både på sannsynlighet og nytte, så finner man hvor gode de forskjellige alternativene er (forventet nytte).

5.2.1 Sammenligning mot en fast verdi

Anta at målet er å teste hypotesen om at en stokastisk variabel θ er større eller mindre enn en gitt referanseverdi θ_0. Her defineres to stegvise nyttefunksjoner, $u_0(x)$ og $u_1(x)$, som representerer nytten under hypotesene $H_0 : \theta \leq \theta_0$ og $H_1 : \theta > \theta_0$, henholdsvis. Nyttefunksjonene kan være stegvise slik at de tar ulike verdier avhengig av om θ faller under eller over θ_0.

5.2.2 Hypotesetest basert på nyttefunksjoner

For å bestemme hvilken hypotese som skal forkastes, vurderes forskjellen mellom de to nyttefunksjonene, $u_1(x) - u_0(x)$. Denne differansen er en 1-sidig hypotesetest, hvor målet er å finne ut om θ er større eller mindre enn θ_0. Når differansen er positiv, gir det større nytte å forkaste hypotesen H_0, mens en negativ differanse sier at H_1 bør forkastes.