Statistik og dataanalyse

Peter Heilbo Ratgen - perat17@student.sdu.dk

4. september 2020

Indhold

1 Uge 36

Denne uge er om kapitlet "Introduction to data".

1.1 Introduktion til kurset

Eksamen er en multiple choice. Undervejs er det tællende aktiviteter.

1.2 Statistiske metoder

Hvordan indsamler vi data, i forhold til hvad vi skal vide? Har vores data bias? Vi har alle sammen bias på en eller anden led. Vi skal gerne lande et sted mellem teori, viden og virkelighed.

Generelt for man svar som man spørger. Hvis man putter urelaterede punkter ind og laver regression, får man altid et matematisk svar, men om dette er korrekt er ikke relateret til den matematiske model man anvender. Statistik analyse baserer sig på normalt distribueret data, uafhængighed og stor eller lille stikprøvestørrelse. Ved ikke at følge disse principper kan vi drage forkerte konklusioner fra dårlig data. Den teoretisk model er korrekt nok, men vi kan ikke drage konklusioner fra dårlig data.

En normalfordeling er den fine lille klokkekurve, fx højde, IQ, vægt, mv. Når man har data nok vil det blive normalfordelt. Generelt set, skal man lave være med at arbejde i små populationer. Data må ikke kunne påvirke hinanden, uafhængighed er det vigtigste i statisk. Det er hele præmissen for statetisk.

1.3 Arbejde med data

Vi skal have en stikprøve. Vi starter med en hypotese. Så skal vi finde en model i den statistiske værktøjskasse. Nogengange ligger svære i at finde det rigtige værktøj. Så estimerer vi, hvad vi før ud af den model vi har valgt. Man har selvfølgelig en forventning om hvad der skal komme ud. Derefter evaluerer vi resultatet af modelleringen fx. har jeg fået det ud af det jeg forventede?

1.4 Population til stikprøve

Man skal have en stikprøve fra den samlede population. En stikprøve er korrekt når den ikke er biased eller noget i den retning. Stikprøven skal være repræsentativ for den samlede population. Den skal også være stokastisk, man skal sikre sig at den man vælger, faktisk er tilfældig. Så kan konklusion der drages af stikprøven, anvendes på den større population.

1.5 Data

Vi har forskellige typer af data.

- Kontinuert - numerisk

  en flydende overgang i data, fx hvor gammel nogen er. En person er et vidst antal år, måneder, dage, timer, sekunder, milisekunder, osv.

- Diskret - numerisk

  Enkelte tal, fx en karakterrække

- Ordinær - kategorisk,

  Kategorisk data har en naturlig orden til sig. I bogen er givet eksemplet med forskellige uddannelsesniveauer, her der forskellige kategorier, men disse kategorier har en bestemt orden mellem sig.

- Nominel - kategorisk

  Nominelle variabler er kategoriske, men har ingen særlig orden til sig, modsat ordinære variabler.

Association er ikke det samme som kausalitet. Kausalitet kan kun drages fra randomiserede eksperimenter. Hvis man bare kigger på tal og tænker sig til en sammenhæng, kan man drage forkerte konklusioner. Et eksempel er Minnesota, med bøgerne i trailerparkerne, der var blevet sat ud på grund af at man havde fundet, det gik bedre for børn i hjem med bøger.

1.6 Statistik og programmering

Vi kan bruge mange værktøjer til at lave statetisk. Vi bruger R. Python kan også bruges til den slags. R er lavet specifikt til formålet, det er lavet af statistikere.