

Om emnet statistik og stikprøver

Læreplanen for matematik STX A og B

<http://www.uvm.dk/Uddannelse/Gymnasiale%20uddannelser/Love%20og%20regler/Laereplaner/stx.aspx>

– simple statistiske metoder til håndtering af et datamateriale, grafisk præsentation af et statistisk materiale, empiriske statistiske deskriptorer, *stikprøvers repræsentativitet*

Undervisningsvejledningen for matematik STX A og B

<http://www.uvm.dk/Uddannelse/Gymnasiale%20uddannelser/Love%20og%20regler/Vejledninger%20til%20laereplaner/stx.aspx>

2.b Statistik og sandsynlighedsregning

Ifølge læreplanen skal eleverne kunne: ”*anvende simple statistiske eller sandsynlighedsteoretiske modeller til beskrivelse af et givet datamateriale eller fænomener fra andre fagområder, kunne stille spørgsmål ud fra modeller, have blik for hvilke svar, der kan forventes, samt være i stand til at formulere konklusioner i et klart sprog*”.

Statistik er videnskaben om indsamling, håndtering og fortolkning af data fra omverden. Selv på et elementært niveau skal statistik ofte forholde sig til ubearbejdede data, og det ligger i fagområdets natur, at statistiske konklusioner ikke kan opnås og præsenteres med samme grad af sikkerhed, som man ellers er vant til i den øvrige del af matematikundervisningen. Dette skal præge undervisningen, så eleverne får et tydeligt indtryk af statistikkens særlige karakter.

Overalt præsenteres vi for oplysninger og påstande, der baserer sig på forskellige mængder og typer af data. Det kan være formuleringer som: ”Aspirin forebygger hjerteproblemer, viser en ny undersøgelse...”, eller: ”Et rundspørge, som Tv-Avisen har foretaget viser, at 2 ud af 3 danskere mener...”. Eksemplarisk materiale af denne type kan være et godt udgangspunkt for en indledende undervisning, for en diskussion af statistikkens metoder samt af spørgsmål som: Hvor sikre er de konklusioner, vi præsenteres for? Undervisningen skal overordnet set medvirke til, at eleverne bedre bliver i stand til at forholde sig kritisk til en formidling af et givet statistisk materiale, samt at de kan stille spørgsmål til kvaliteten af og håndteringen af statistiske undersøgelser.

Ifølge læreplanen omfatter kernestoffet: ”*simple statistiske metoder til håndtering af et datamateriale, grafisk præsentation af et statistisk materiale, empiriske statistiske deskriptorer, stikprøvers repræsentativitet*”.

Statistik arbejder med metoder til at håndtere usikkerhed. Men de spørgsmål, man søger svar på, må ikke være præget af uklarhed. Tværtimod er det afgørende, at man så præcist som muligt har gjort sig klart, hvad det er, man vil måle, hvad man gerne vil vide og hvilke antagelser, man i øvrigt gør sig, før arbejdet starter.

Det første skridt ind i statistikken vil normalt være en overvejelse om, hvad troværdige data er: Hvorledes vælger man stikprøver af en population (se Eksempel 220) således, at stikprøven kan siges at være repræsentativ.

Hvordan designer man metoder til at skaffe data, således at man med statistiske metoder kan give troværdige svar på givne spørgsmål?

Gennem undervisningen skal eleverne have mødt så mange eksempler på stikprøve-situationer – herunder stikprøver præget af forskellige former for systematiske fejl (bias), stikprøver, hvor der er skjulte variable på spil (konfundering), og stikprøver, hvor forskellige typer blindtest anvendes – at de kan håndtere problemstillinger som:

– I nyhedsudsendelsen på en lokal tv-station fortælles: ”Inden for de sidste timer har en af vo-re journalister gået rundt i Kolding og spurgt 55 tilfældige forbipasserende om holdningen til en aktuel og meget provokerende kunstudstilling. 35 af de adspurgte ønskede udstillingen lukket. Der er således et massivt pres på byrådet om at gribe ind. Nu afventer vi borgmeste-rens reaktion.”

- Hvad er populationen og hvad er stikprøven?
- Kommenter undersøgelsen og tv-kanalens præsentation af denne.

– Et sundhedsmagasin ønsker at undersøge, om store doser vitamintilskud forbedrer sundheds-tilstanden. Bladet anmoder de af læserne, som gennem længere tid har taget store doser vi-tamintilskud, om at skrive ind og fortælle om positive og negative erfaringer med dette. 2754 læsere skriver ind. 93 % fortæller, at de kan spore en vis forbedring af helbredet.

- Hvad er population, og hvad er stikprøve?
- Kommenter undersøgelsens metode, og skriv et lille indlæg herom til en avis.

– En bestemt sygdom påvirker de røde blodlegemer og forårsager stor smerte. Et medicinsk præparat til behandling af sygdommen er udviklet, og kvaliteten af præparatet ønskes afprø- vet på en population på 300 patienter, der har haft særligt mange smerteanfald.

- Forklar, hvorfor det ville være en dårlig strategi at lade alle 300 få den nye medicin.
- Beskriv et forsøg, der kunne give information om pågældende præparats virkning over for smerteanfald.

På fagets side på emu'en er der placeret en række yderligere opgaver om stikprøver.

Eleverne forventes at kunne anvende simple statistiske deskriptorer og simple grafiske præsentatio- ner i en beskrivelse af et datamateriale. Det drejer sig om middeltal, median og kvartilsæt, om box- plot's og histogrammer, der kan optræde i problemstillinger som:

– For en bestemt gruppe på 15 læger blev det undersøgt, hvor ofte de havde udført et kirurgisk indgreb, der medførte fjernelse af livmoderen. Antallet af operative indgreb var for hver af de 15 læger henholdsvis: 50 33 25 86 25 85 31 37 44 20 36 59 34 28 49

En gruppe på 10 kvindelige læger blev tilsvarende undersøgt, og blandt disse blev der udført indgreb følgende antal gange: 7 14 25 5 33 29 18 31 10 20

- Lav i samme koordinatsystem boxplot af hver af de to datasæt.
- Kommenter undersøgelsen ved hjælp af den grafiske præsentation og størrelserne mid-deltal, median samt kvartilerne for de to datasæt.

- Givet et histogram over matematiklæreres aldersfordeling. Karakteriser dettes form. Ligger medianen til venstre for, til højre for, eller er den lig med middeltallet? Begrund svaret.
- Givet en computerudskrift af et statistisk materiale og beregninger af: maksimum, minimum, øvre og nedre kvartil, median, middelværdi og eventuelle andre karakteristiske størrelser for datasæt for forskellige sammenlignelige produkter, fx madtyper og kalorieindhold. Der ønskes en sammenligning i form af opstilling af boxplots for produkterne i samme koordinatsystem. Der ønskes en kommentar til den grafiske præsentation.

Histogrammer i eksamensopgaver er baseret på lige brede intervaller, mens de til grund liggende intervaller for sumkurver godt kan have forskellig bredde. Opgaven kan i øvrigt dreje sig om at anvende en grafisk repræsentation af et talmateriale til at svare på forskellige spørgsmål, herunder at aflæse og kommentere kvartilsæt. Da kvartilsæt ikke er entydigt defineret, kan forskellige it-redskaber give lidt forskellige svar, hvilket anses for uproblematisk. Sådanne opgavetyper vil ikke indgå i prøven uden hjælpemidler.

De faglige mål vedrørende statistik og sandsynlighedsregning vil hovedsageligt blive udmøntet gennem det supplerende stof. Ifølge læreplanen skal dette omfatte ”*anvendelse af mindst to typer statistiske eller sandsynlighedsteoretiske modeller, indsamling og bearbejdning af data til belysning af en opstillet hypotese*”.

Den sandsynlighedsteoretiske formalisme med udfaldsrum og sandsynlighedsfunktion er ikke en del af det fælles kernestof. Men det enkelte hold kan vælge at gennemføre forløb over elementer af klassisk sandsynlighedsteori og kombinatorik, og så evt. bygge binomialmodeller, urnemodeller, betingede sandsynligheder eller andet ovenpå. Sandsynlighedsbegrebet kan imidlertid også introduceres gennem frekventielle sandsynligheder (i diskrete tilfælde) knyttet til statistiske undersøgelser.

Det formelle begreb ”stokastisk variabel” indgår heller ikke i det fælles kernestof. Men i bestemte forløb kan det være en fordel at introducere stokastisk variabel som et begreb og en notation, der gør det mere enkelt at formulere spørgsmål og opstille formler.

Et datamateriale kan tilvejebringes på mange måder:

- eleverne kan selv via spørgeskemaer, test i idræt eller på anden vis generere det datamateriale, holdet vil underkaste en statistisk analyse
- datamaterialet kan også komme via et samarbejde med andre fag (fx naturvidenskabeligt grundforløb, eksperimentelle fag eller samfunds-fag)
- man kan også vælge at trække på det omfattende materiale af autentiske data, som findes i en række databanker på nettet. Med moderne it-værktøjer kan sådanne data umiddelbart trækkes ind og gøres til genstand for statistisk behandling.

En statistisk undersøgelse af et materiale har normalt flere trin, hvoraf første trin er af mere deskriptiv karakter. Denne fase giver samtidig bedre muligheder for at kunne stille præcise spørgsmål til det givne materiale. I statistiske undersøgelser formuleres sådanne spørgsmål ofte som hypoteser. Belysning af en opstillet hypotese kan gennemføres på mange måder. Man kan vælge en eksperimentel tilgang med anvendelse af statistiske it-værktøjer. Men et hold kan naturligvis også vælge at fordybe sig i elementer af klassisk hypotesetest. Mange af de forløb, man kunne vælge at gennemføre, har et teoretisk fundament, der går betydeligt ud over det gymnasiale niveau. Det er ikke tanken, at man i sådanne forløb skal søge at nå til bunds i en forståelse af den fordeling, man arbejder med, eller eksempelvis af det formelle grundlag for hypotesetest. I de fleste tilfælde vil det være

mere hensigtsmæssig at inddrage eksperimentelle metoder og præsentere det matematiske begrebsapparat, så det spiller sammen med intuitionen.

Statistik og sandsynlighedsregning har så mange berøringsflader med omverdenen og med andre fag, at der er et stort og varieret antal emner inden for dette område, som kan være genstand for et samarbejde med andre fag, eller som kan dyrkes på rent matematikfagligt grundlag:

- Et forløb om opinionsmålinger (se Eksempel 220) kan både tilrettelægges i grundforløbet og i en studieretning i et samarbejde med samfundsfag. Datamaterialet kan indsamles i spørgeskemaer eller hentes fra databaser eller opinionsmålingsinstitutter. Gennem forskellige former for simuleringer kan der arbejdes med test af hypoteser og med graden af sikkerhed, hvormed resultater præsenteres. Både randomfunktioner i lommeregner og regneark og mere avancerede it-værktøjer kan anvendes.
- Datamateriale fra større spørgeskemaundersøgelser (se Eksempel 221) kan give grundlag for, at eleverne selv opstiller hypoteser om sammenhænge, og om hvorvidt forskellige data-serier udtrykker reelle forskelle mellem køn, aldersgrupper, meningsgrupper osv. eller er udtryk for statistiske tilfældigheder. Et sådant forløb indebærer anvendelse af it-værktøjer.
- Måleserier genereret fra forsøg i eksperimentelle fag, eller målinger foretaget på eleverne i idræt eller biologi – (fx om følsomhed på ryg og hænder, om evnen til at smage forskel, om reaktionshastighed, træfsikkerhed osv.) – kan give anledning til en statistisk sammenligning af måleserier (se Eksempel 280) på grundlag af hypoteser, som eleverne formulerer. Sammenligninger kan både ske ved eksperimenterende arbejdsformer på it-værktøjer eller i nogle andre sammenhænge ved et binomialtest, eller ved at introducere χ^2 -test.
- Et forløb om ventetider (fx i fysik om radioaktivt henfald) kan føre til en introduktion og anvendelse af Poisson-fordelingen og evt. et matematisk studium af sider ved denne.
- Et forløb med random-walk modeller kan give en eksperimenterende tilgang til at undersøge normalfordelinger. Man kan også vælge at anvende statistisk it-værktøj og her lade eleverne eksperimenter med en række forskellige datamaterialer på en sådan måde, at der via et stort antal simuleringer og beregninger af tilhørende middelværdier tilvejebringes et troværdigt indtryk af, hvorledes normalfordelingen optræder som grænsetilfælde.
- Et forløb om arvelighed (se Eksempel 210) kan tilrettelægges, så fokus er på simple betingede sandsynligheder og Hardy-Weinbergs lov. Men man kan også gennemføre et lidt længere forløb om retsgenetik og dna-profiler (se Eksempel 211), hvor der teoretisk bygges på Bayes' sætning, men samtidig udnyttes it-værktøjer til analyse af datamaterialet.
- Et forløb om Simpsons paradoks kan gennemføres alene i matematik eller sammen med andre fag, der også ønsker at sætte fokus på denne side af præsentationen af simple statistiske oplysninger. Forløbet kan allerede tilrettelægges i 1. g, og bygger hovedsageligt på vejet gennemsnit.
- Men man kan også vælge på et hold at gennemføre forløb med en sandsynlighedsteoretisk undersøgelse af forskellige former for spil og lotto. Både binomialmodeller og urnemodeller kan komme i spil.
- Risikovurdering (se Eksempel 273) indgår på mange steder i et moderne samfund og kan både give materiale til forløb om forsikringsmatematik og forløb om vurdering af sikkerheden på avancerede virksomheder.

På fagets side på emu'en er der placeret en række yderligere forløb og materialer til inspiration:

<http://www.emu.dk/gym/fag/ma/undervisningsforloeb/paradigmatiske/indeks.html>