# Statistisk sandsynlighed

Sandsynlighed kan opdeles i to dele statistisk sandsynlighed og kombinatorisk sandsynlighed.

- Statistisk sandsynlighed er når man kun kender et udsnit/stikprøve af en ukendt mængde og ved at se på udsnittet/stikprøven kan man konkludere noget om den ukendte mængde.
  F.eks. hvis man har en skæv terning/figur med en ukendt sandsynlighedsfordeling kan man blive nød til at foretage en undersøgelse for at finde den statistiske sandsynlighed. Ved at udføre en stor mængde eksperimenter kan man ud fra observationerne forudsige en sandsynlighed. Når man har fundet denne kan man begynde at regne med den.
  Eksempler: Kastegris, Knappenåle, Tændstikæsker osv.

- Kombinatorisk sandsynlighed er når man kender antallet af mulige udfald og derved kan beregne sandsynligheden.
  F.eks. hvis man har en 6-sidet terning, kan man beregne den kombinatoriske sandsynlighed, hvis man som udgangspunkt antager at der er 1/6 sandsynlighed for at lande på hver af terningens sider.
  Eksempler: Mønter, Terninger osv.

I denne tekst præsenteres en fremgangsmåde ved statistisk sandsynlighed.

**1) Undersøgelse og opstilling af hypoteser**

For det første skal man gøre sig klart hvad det er man ønsker at undersøge. Ved indsamling af data til sin undersøgelse, skal man også vide hvilken skalatype man vil indsamle data efter. Herunder er opstillet fire skalaer, hvor den skalatype med flest detaljer står øverst og den med mindst detaljer nederst:

- Ratio-interval skala: På denne skala kan værdierne forholde sig til hinandens størrelser, da afstanden mellem naboværdier er ens gennem hele skalaen. F.eks. længde (mm, cm, m), vægt (mg, g, kg) eller, rumfang (ml, dl, l).

- Interval skala: Ikke alene rangordnen kan måles, med også afstanden imellem dem. F.eks. karakterer eller VAS smerte (Visuel Analog Skala).

- Ordinal skala: En ragnordning i en entydig rækkefølge. Her kan man ikke sige noget størrelses forholdet imellem dem. F.eks. ”ikke tilfreds”, ”tilfreds” og ”meget tilfreds”.

- Nominal skala: Er klassifikationer uden indbyrdes aritmetisk sammenhæng. F.eks. ”Mand” og ”Kvinde” eller en rangorden som ”Officer”, ”Befalingsmand” eller ”Menig”, men også ord som ”Ja”, ”Nej” eller ”Ved ikke”.

Ud over skalatyperne kan man opstille retningsbestemte og ikke-retningsbestemte hypoteser. Ved retningsbestemte hypoteser har man en forhåndsindstilling til hvordan det må forholde sig. F.eks. der er sammenhæng mellem antal minutters daglig fysisk træning og kropsvægten. Desto flere minutter desto lavere kropsvægt.

**2) Indsamling af data**

Ved indsamling af data til en undersøgelse, er det sjældent muligt at spørge alle i forbindelse med en undersøgelse. Derfor foretager man undersøgelser på baggrund af en stikprøve, altså et udsnit af virkeligheden. Det er meget vigtigt at udvælge denne stikprøve repræsentativt, da undersøgelsen ellers er ubrugelig. I den forbindelse er der fordele og ulemper ved forskellige udvælgelsesformer. Herunder er der opstillet fire udvælgelsesformer, hvor den udvælgelsesform der er bedst står øverst og den dårligste nederst:

- Simpel tilfældig udvælgelse: Alle elementer har samme sandsynlighed for udvælgelse.

Kort før præsidentvalget i USA, 1936, foretog man en meningsmåling, hvor stikprøven blev udtaget blandt bilejere og husstande med telefon. Derefter konkluderede man, at den republikanske guvernør Landon ville vinde i langt de fleste stater. Men demokraten Franklin D. Roosevelt blev genvalgt med den hidtil største valgsejr i USA’s historie.

- Systematisk udvælgelse: Samme sandsynlighed for udvælgelses, men kun første element udtages tilfældigt.

- Stratificeret udvælgelse: Hvert stratum har sin egen udvælgelses sandsynlighed. Stratificeret udvælgelse sikrer, at stikprøven ligner befolkningen med hensyn til andelen af enheder i mindre grupper. Målet er at have en stikprøve, der på nogle kendte egenskaber ligner befolkningen. Stratificerede stikprøver laves for at sikre, at analyser af mindre grupper inde i stikprøven kan lade sig gøre. F.eks. Køn, Alderstrin mv.

- Klyngeudvælgelse: Med klyngeudvælgelsen kan omkostninger og besvær reduceres, men det øger til gengæld usikkerheden ved stikprøvens generalisering. F.eks. geografiske grupperinger.

Der kan naturligvis også benyttes en kombination af flere af ovenstående.

**Kort om spørgeskemaer**
Har man f.eks. lavet et spørgeskema, skal man være opmærksom på hvor man spørger, hvilke tidspunkt på dagen og hvem man henvender sig til. Husk også at overveje hvor lang tid folk skal bruge på et spørgeskema, opfør dig venlig og vær forberedt på et afslag.

**3) Deskriptiv statistik - Beskrivelse af data**
Indledende behandling af data, det kan være beregning af få relevante tal som f.eks. gennemsnit, variationsbredde eller median. Det kan også være fremstilling af relevante diagrammer f.eks. søjlediagram, cirkeldiagram eller kurvediagram. Dette afhænger af undersøgelsens data og opdeles i enkeltobservationer og grupperede observationer.

**3.1) Deskriptiv statistik – Enkeltobservationer**
1. Tallene som der skal føres statistik over kaldes som helhed et observationssæt og de enkelte tal kaldes observationer.
2. Opstil en tabel over observationerne indeholdende:
    a. Hyppighed h(x) viser antal gange den enkelte observation forekommer.
    b. Summeret hyppighed H(x) viser hvor mange observationer der er mindre end eller lig med den givne observation.
    c. Frekvens f(x) viser hyppigheden i procent.
    d. Summeret frekvens F(x) viser hvor mange procent af det samlede observationssæt, der ligger på den givne observation eller derunder.

**Scan QR-koden for at se en videovejledning til enkeltobservationernes diagrammer i regneark.**

| | Hyppighed | Summeret hyppighed | Frekvens | Summeret frekvens |
|---|---|---|---|---|
| 1999 | 5 | 5 | 23 % | 23 % |
| 2000 | 4 | 9 | 18 % | 41 % |
| 2001 | 5 | 14 | 23 % | 64 % |
| 2002 | 3 | 17 | 13,5 % | 77,5 % |
| 2003 | 3 | 20 | 13,5 % | 91 % |
| 2004 | 2 | 22 | 9 % | 100 % |

3. Størsteværdien er den observation, som har den største værdi.
4. Mindsteværdien er den observation, som har den mindste værdi.
5. Variationsbredden er forskellen mellem den største og mindste observation.
6. Medianen finder man ved at opstille alle observationerne i stigende rækkefølge og derefter finde den midterste.
7. Typetallet er det tal som forekommer hyppigst – altså flest gange.
8. Middeltallet/gennemsnit findes ved at lægge alle observationerne sammen og dividere med antallet af observationer.
9. Der findes mange forskellige diagrammer. Herunder er der angivet hvad man kan vise med hvilke diagrammer. (Scan QR-koden)
    a. Pindediagram kan vise tallene for hyppighed.
    b. Trappediagram kan vise summeret hyppighed eller summeret frekvens.
    c. Blokdiagram eller cirkeldiagram kan vise frekvens.

10. Kvartiler findes ud fra den summerede frekvens F(x). 1. Kvartil er ved 25 %, 2. Kvartil er ved 50 % (dvs. denne kaldes også medianen) og 3. Kvartil er ved 75 %.

**3.2) Deskriptiv statistik – Grupperede observationer**

1. Forskellen er her at observationssættet er grupperet i intervaller med observationer. Typisk for at skabe et bedre overblik, hvis der er mange observationer. Dog har det den pris, at jo større intervaller, desto flere nuancerne forsvinder der. Hvis man anvender intervaller er det meget vigtigt at man anvender lige store intervaller, for ikke at fordreje observationerne.
2. Forskellen er her at tabellen over observationerne skal indeholdende intervallerne og angives ved hjælp af kantede parenteser: F.eks. [16-20[ betyder at 16 er med, og at 20 ikke er med.
    a. Intervalhyppighed h(a;b) viser antallet af observation i det givne interval.
    b. Summeret intervalhyppighed H(x) viser hvor mange observationer der er i et givet interval og intervallerne derunder.
    c. Intervalfrekvens f(a;b) viser intervalhyppigheden i procent.
    d. Summeret intervalfrekvens F(x) viser hvor mange procent af det samlede observationssæt, der ligger i det givne interval og derunder.

**Scan QR-koden for at se en videovejledning til de grupperede observationers diagrammer i regneark.**

| | Intervalhyppighed | Summeret intervalhyppighed | Intervalfrekvens | Summeret intervalfrekvens |
|---|---|---|---|---|
| **0** | | | | 0% |
| **[16-20[** | 4 | 4 | 9% | 9% |
| **[20-24[** | 18 | 22 | 40% | 49% |
| **[24-28[** | 15 | 37 | 33% | 82% |
| **[28-32[** | 5 | 42 | 11% | 93% |
| **[32-36[** | 3 | 45 | 7% | 100% |

3. Intervalmidtpunktet er det midterste tal i observationsintervallet.
4. Typeinterval er det interval som indeholder flest observationer.
5. Middeltal/gennemsnit findes ved at gange intervalmidtpunktet med intervalhyppigheden og lægge dem sammen. Derefter divideres der med det samlede antal observationer.
6. Der findes mange forskellige diagrammer. Herunder er der angivet hvad man kan vise med hvilke diagrammer. (Scan QR-koden)
    a. Søjlediagram kan vise intervallerne for hyppighed.
    b. Kurvediagram kan vise summeret hyppighed eller summeret frekvens.
    c. Blokdiagram eller cirkeldiagram kan vise frekvens.
7. Kvartiler findes ud fra den summerede intervalfrekvens F(x). 1. Kvartil er ved 25 %, 2. Kvartil er ved 50 % og 3. Kvartil er ved 75 %.

**4) Forudsigelser og fortolkning**

Hypoteserne af- eller bekræftes, forstået på den måde, at man foregriber, hvad man vil se i undersøgelser, der endnu ikke er foretaget. Hvor stor en tillid skal man have til konklusionerne og hvilke fejlmuligheder må der regnes med.

Når man skal beskrive data, er det vigtigt, at man forstår at tolke korrekt. Derved undgår man fejlslutninger. F.eks. er middeltallet særligt følsom over for værdier der ligger langt fra de andre. Det er medianen derimod ikke, derfor kan man overveje om denne burde bruges i stedet.

Det er vigtigt, at man ikke fordrejer data, som f.eks. ved følgende udsagn:
- Du kan godt bunde i åen, gennemsnitsdybden er kun 50 cm.
- Et menneske kan blive over 110 år.
- Hvert 4. sekund dør et mennesker af sult.