Statistik

1. Lektion

Lærer: Michael Larsen

Definition: Fra data til information.

Når det omhandler data bruger man sandsynlighedsregning.

Sandsynlighed – tal mellem 0 og 1 – 0% og 100%. 0% er defineret som den umulige hændelse og 100% er den sikre hændelse.

Udfaldsrum: Alle mulige udfald af et eksperiment.

F.eks plat/krone

1 kast:

U = Udfald = (P, K)

2 kast:

U = (pp, kk) samme, så at få en er der 50% sandsynlighed for.

U = (0, 1, 2) antal krone – dvs. 3 udfald

U = (pp, kk, pk, kp) – 4 udfald. – 25 %

Sandsynlighed for en hændelse = summen af ssh. For de udfald hændelsen består af.

Sandsynlighedsregning

I Sandsynlighedsregning er udfaldsrummet kendt og udfaldet et ukendt.

Statistik

Population (målgruppe): Samling af elementer som ønskes undersøgt. – Ukendt.

Stikprøve – Kendt.

2 typer populationer –

Endelige: Kendt antal (personer)

Uendelige: Proces. Antal begivenheder i en tidsperiode. – F.eks. Antal kunder der kommer mellem kl. 15-18

Sandsynlighedsregning eksempel:

I en bøtte med 1000 grønne perler og 1000 røde perler. Hvis man tager en perle op, hvad er sandsynligheden for at den bliver grøn?

Ssh. For 1 grøn er $\frac{gunstige}{mulige} = \frac{1000}{2000} = 50\%$

Udtages 50 perler tilfældigt. Ssh. For 30 eller flere grønne? 10,13 %

Stikprøve på 50

30 blå

20 sorte

Estimation (konfidensintervaller)

Med 95% ssh. Vil andelen af blå ligge mellem 47% og 72%

Hypotesetest:

Vil undersøge om krukken indeholder flere blå end sorte.

Vi antager at krukken indeholder lige mange af hver.

Ssh. For stikprøven givet antagelsen er korrekt: 10,13 %

Eksempel fra valget 2011

Valgresultatet: 13,8 % DF

13. maj 2011: repræsentativt udvalg udspurgt (1024 personer): 15,7 % DF

24. Juni 2011: 973 personer: 13,4 % DF

13 maj: antag at 13,8 % i populationen vil stemme DF – Hvad er sandsynligheden for et stikprøve hvor 15,7% vil stemme DF: 3,9%

Antag at 13,8 % stemmer DF. – Hvad er sandsynligheden for en stikprøve hvor 13,4% stemmer DF: 35,9 %

Antag stikprøverne kommer fra samme population. – Ssh for disse prøver vil være 7,3 %

2. Lektion

Beskrivende Statistik

Population –> N elementer – observationer $x1 = 165,\ x2 = 177, x3 = 183$

Stikprøve –> n elementer

$$XN/Xn$$

	Population	Stikprøve
Moment – Gnmsnit/mdlv	μ = middelværdi = $\frac{x1+x2+x3}{N}$	
Centrale moment – varians. Den gennemsnitlige kvadratvigelse fra middelværdien.	$O^2 = \frac{1}{N} sum(Xi - \mu)^2$	$S^2\hat{O}^2 = \frac{1}{n-1} sum\ (X1 - X)^2$
Standardvigelsen	O=$\sqrt{O^2}$	$\frac{S}{\hat{O}} = \sqrt{S^2}$
Central standardiserede moment – skævhed.	<O Venstreskæv =O Symmetrisk >O Højreskæv	
Fraktiler: Opdeler datamaterialet – Sortere data i stigende orden – Kvartilsættet – Median: Midterste observation. Hvis 2 midterste observationer, tages gennemsnittet af disse. 50% Fraktiler. Øvre Kvartil: 75% fraktiler.		

Medianen og gennemsnittet er ikke det samme.

Interkvartilbredden: Øvre – nedre kvartil

Vibrationsbredden: Største-mindste observation

Typetallet/middelværdi: Den observation som forekommer flest gange.

3. Lektion

Grupperede

- Intervaller
- punkter

U grupperet

-

To typer af data

- **Numeriske**
- - diskrete (hele tal)
- - kontinuerte (decimaltal, alle tal)
- **Kategoriske**
- - Ja/nej
- - Køn: M/K
- - Kommune
- - Parti

Dvs. tekstdata

Estimation af populationsandel

n = stikprøvens størrelse

x = antal mærkede i stikprøven

punktestimation for populationsandelen =

^P = x/n

Konfidensinterval for populationen – andelen.

lille alfa = α

α: Risikoen for at begå fejl i forbindelse med konfidensintervaller.

Typisk er alfa = 5 % (1-10%)

Med (1-alfa) ssh. vil intervallet

Indeholde populationsandelen med (1-alfa) ssh. vil populationsandelen ligge mellem nedre og øvre grænse.

Næste Lektion

Konfidensintervaller

Alfa: Risikoen for at begå fejl i forbindelse med ki

Med (1-alfa) ssh. vil intervallet indeholde populationsparameteren

Med (1-alfa) ssh. vil populationsparameteren ligge mellem nedre og øvre grænse.

Stikprøve på 39 perler.

Heraf er der 18 blå.

Punktestimat for andelen af blå: P = x/n = 18/39 = 0,4615= 46,15 %

Tandtråd af mærket Floss

95 % ki for markedsandel

Stikprøve for Floss og andre

Floss: 107 Andel: 33,44 %

Andre: 213 Andel: ,66,56 %

I alt: 320 Andel: 100 %

Med 95 % ssh/sandsynlighed vil Floss' markedsandel ligge mellem 28 % og 39 %.

Symboler:

Middelværdi - μ

Andel – P/pi

Varians – sigma i ^2

Standardafvigelse – sigma

Intensitet (antal forekomster pr. enhed) – Lambda

Et bryggeri eksemplet

Bryggeriet vil brygge en ny øl med en alkohol procent på 4,6 % - De laver en stikprøve på 10 øl.

1) Beregn 95 % Ki for den gennemsnitlige alkoholprocent.
2) Kan bryggeriet med god samvittighed påstå at den nye øl har en alkohol procent på 4,6 %.

- **Svar**

1. Med 95% ssh vil den gennemsnitlige alkoholprocent ligge mellem 4,5 % og 5,2 %
2. Da 4,6 % ligger i ki kan det ikke afvises at den gennemsnitlige alkoholprocent er 4,6 %

Hypotesetest

Fx: Vi vil gerne påvise at andelen som stemmer på rød blok er større end 50 %.

- Vi opstiller 2 hypoteser som udelukker hinanden.

H_0 – Nulhypotesen. Erfaringerne.

H_1 eller H_A: Alternativhypotesen: Udtrykker det som skal påvises.

	2-sidet/skarpt	1-sidet/retningsbestemt	
H0	=	<	>
H1	≠	>	<

H0: P< 0,5 (P=0,5)

H1: P > 0,5

Testniveau:

Alfa = 5 % (1 % - 10 %)

Hvis p-værdien/signifikans-ssh er større end testniveauet – så accepteres H0.

Konklusion: Det kan ikke afvises, at (H0 i ord)

eller: Det kan ikke påvises, at (H1 i ord)

Modsatte konklusion: Hvis p-værdien/signifikans ssh er mindre end testniveauet – så forkastes H0.

Konklusion: Stikprøven tyder på, at (H1 i ord)

- Det kan ikke afvises at, (H1 i ord)
- Det kan ikke påvises at (H0 i ord)

Supermarked: Har erfaring for at det gennemsnitlige salg pr. kunde er 300 kr, kampagne med at indsamle point. Stikprøve på 12 kunder tyder stikprøven på at det gennemsnitlige salg er steget?

H0: M < 300

H: M > 300