Forelesninger

Anvendt metode

Motivasjon

Statistikk kan trekkes inn i en rekke sammenhenger i en bedrift:

Markedsanalyse

Produktplanlegging

Produksjonsplanlegging

Kvalitetstyring

Produktivitet

Økonomisk styring

Prosjektstyring

Personaladministrasjon

Intern kontroll

Revisjon

Egenvurdering

Fellesnevnere: En samler (kanskje) inn data, og trekker beslutninger ut fra dette.

Settifisering: ISO 9000.

Generelle problemer:

- Feil analyse

- Trekker beslutninger på for vagt grunnlag.

- Kortsiktig tenkning:

Et par gode resultater blir glorifisert, mens et par dårlige resultater gir grunn til krise.

Eksempel

En avdeling har dårligere tall enn gjennomsnittet, det er også dårligere enn gjennomsnittet året før. Dette representerer en nedgang, og oppfattes som negativt for bedriften.

Et slikt resultat kan være helt normalt!!!

En trenger kunnskap om spredningen av tallene for å kunne si at resultatet sannsynligvis ikke skyldes tilfeldigheter.

Elementær bearbeiding av data

Gjennomsnitt

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

R-kommando: mean()

Eksempel
>data=c(1,3,2,4,7)
>mean(data)
>[1] 3.4

Empirisk varians

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2$$

R-kommando: var()

Eksempel
>var(data)
>[1] 5.3

Standardavvik

$$S_X = \sqrt{S_X^2}$$

Måler det typiske avviket fra gjennomsnittet.

R-kommando:sd()

Eksempel
>sd(data)
>[1] 2.302173

Empirisk kovarians

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

Korrelasjonskoeffesient:

$$\rho_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$

Måler om tallene har en tendens til å trekke i samme retning. Dette kalles samvariasjon.

Stiger kursene på Oslo Børs når Dow Jones indeksen stiger? Har bankaksjer en like klar tendens til å øke i kurs når kursen på teknologiaksjer går ned?

Dette kan vi måle ved hjelp av korrelasjonskoeffesienten.

$$-1 \leq \rho_{XY} \leq 1$$

Verdier nær -1 er maksimalgrad av negativ lineær samvariasjon, og verdier nær 1 er maksimal grad av positiv lineær samvariasjon.

Eksempel

Samsvarende verdier av X og Y				
X_1	X_2	X_3	X_4	X_5
2	12	1	1	10
Y_1	Y_2	Y_3	Y_4	Y_5
3	11	3	1	12
X_6	X_7	X_8	X_9	X_{10}
25	3	9	27	2
Y_6	Y_7	Y_8	Y_9	Y_{10}
21	6	4	31	2

Tar vi en rask titt på tallene i tabellen, ser vi at X og Y samvarierer.

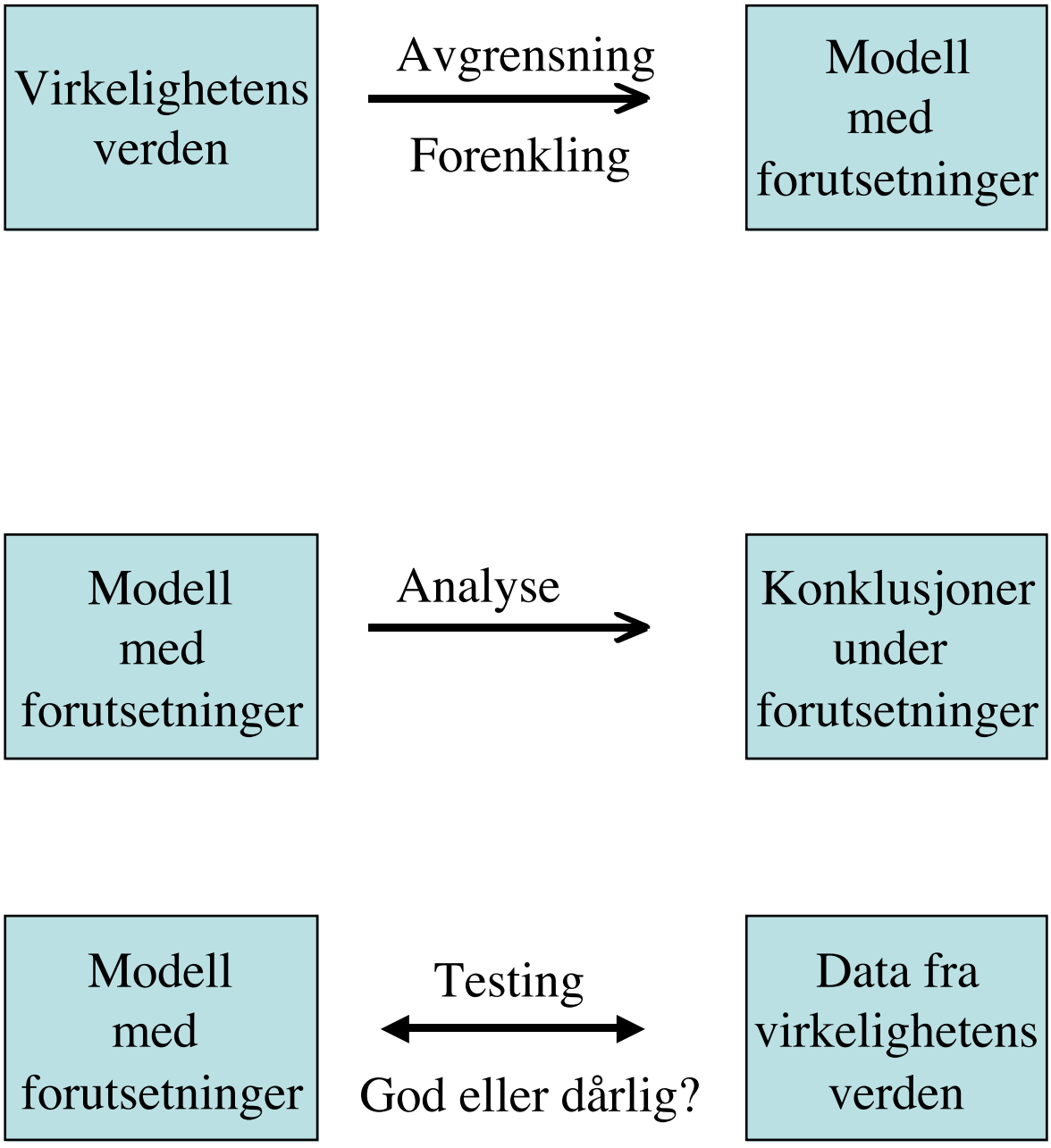
R-kommandoer: cov() og cor()

```
> x=scan()
1: 2 12 1 1 10 25 3 9 27 2
11:
Read 10 items
> y=scan()
1: 3 11 3 1 12 21 6 4 31 2
11:
Read 10 items
> cov(x,y)
[1] 90.8
> cor(x,y)
[1] 0.955797
```

Vi ser at korrelasjonskoeffesienten er så stor som 0.96, som viser nær maksimal grad av positiv lineær samvariasjon.

Modellering

Gangen i en statistisk modellering kan vi illustrere slik



Grunnleggende teori

Utfallsrom og sannsynligheter

Utfallsrommet avgrenser problemstillingen:
Kun disse utfallene kan inntreffe

Sannsynlighetene spesifiserer modellen:
Hvert utfall har en gitt sannsynlighet

Tekniske betingelser

$$0 \leq p_i \leq 1$$

$$p_1 + \cdots + p_n = 1$$

En *begivenhet* er et spesifisert sett med utfall. Vi sier at begivenheten inntreffer hvis vi får et av disse utfallene.

Viktige regneregler:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(\overline{A}) = 1 - P(A)$$

Betinget sannsynlighet

De fleste sannsynligheter er betinget i en eller annen form. Avgrensingen av modellen er i seg selv en betingelse.

Definisjon

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

Måler hvor stor "andel" $A \cap B$ utgjør av B.

Bayes lov

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

Oppsplittingsprinsippet:
Når $\Omega = B_1 \cup \cdots \cup B_n$ og $B_1, \ldots, B_n$ alle er disjunkte, er

$$P(A) = P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n)$$

To begivenheter er *uavhengige* når

$$P(A \cap B) = P(A) \cdot P(B)$$

Eksempel

Insentivordning: Alle kunder blir tilbudt et skrapelodd dersom de tar en informasjon-srunde med en kunderådgiver.

Forsøket varer i et år. 10% av kundene kjøpte bil, 30% av kundene som kjøpte bil fikk skrapelodd, og 10% av kundene som ikke kjøpte bil fikk skrapelodd.

Spørsmål: Har kunder som får skrapelodd, en høyere sannsynlighet for kjøp?

Løsning

K : Kunden kjøper bil

S : Kunden får skrapelodd.

$$P(K) = 0,1 \qquad P(K^c) = 0,9$$

$$P(S|K) = 0,3 \qquad P(S|K^c) = 0,1$$

Oppsplitting

$$\begin{aligned} P(S) &= P(S|K) \cdot P(K) + P(S|K^c) \cdot P(K^c) \\ &= 0,3 \cdot 0,1 + 0,1 \cdot 0,9 = 0,12 \end{aligned}$$

Bayes lov

$$P(K|S) = P(S|K) \cdot \frac{P(K)}{P(S)}$$
$$= 0,3 \cdot \frac{0,1}{0,12} = 0,25 = 25\%$$

Kunder som får skrapelodd har altså en høyere sannsynlighet for kjøp.

Tilfeldige variabler

Når ethvert utfall spesifiserer et bestemt tall X, kaller vi X for en tilfeldig/stokastisk variabel.

Variabelen kalles *diskret* når den bare kan ha verdiene

$$x_1, x_2, \ldots, x_n$$

Fordelingsfunksjonen

$$p(x) = P(X = x)$$

Den kumulative fordelingsfunksjonen

$$F(x) = P(X \leq x)$$

Forventningen

$$\mathrm{E}[X] = \sum_{i=1}^{n} x_i \cdot p(x_i)$$

$$\mathrm{E}[f(X)] = \sum_{i=1}^{n} f(x_i) \cdot p(x_i)$$

Setninger

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$$

$$\mathrm{E}[k \cdot X] = k \cdot \mathrm{E}[X]$$

Forventingen til en tilfeldig variabel er den verdien vi i gjennomsnitt ville observert hvis modellen var 100% korrekt, og vi hadde gjentatt eksperimentet mange ganger.

Eksempel

Et firma har 1-6 dagers leveringstid på en vare.

Leveringstider			
Dager	1	2	3
P	55%	20%	10%
Dager	4	5	6
P	5%	5%	5%

La X være antall dagers leveringstid på en tilfeldig vare.

a) Forklar at X er en tilfeldig variabel.

b) Finn den kumulative fordelingsfunksjonen $F(x)$ til X. Hva slags tolkning har $F(x)$ i dette tilfellet?

c) Finn $\mathrm{E}[X]$.

Løsning

a) X er et tall.

b)

$$F(1) = 55\% \qquad F(2) = 55\% + 20\% = 75\%$$

$$F(3) = 55\% + 20\% + 10\% = 85\% \qquad \text{osv}$$

c)

$$\begin{aligned}\mathrm{E}[X] &= 1 \cdot 0,55 + 2 \cdot 0,2 + 3 \cdot 0,1 \\ &\quad + 4 \cdot 0,05 + 5 \cdot 0,05 + 6 \cdot 0,05 = 2\end{aligned}$$

Kategorier

Eksempel

Vi kjenner inntekt og formue til en mulig kunde, men har ikke tid til å foreta en kredittvurdering. Har kunden

A: Betjent alle lån uten problemer

B: Betjent lånene, men med purringer

C: Misligholdt et lån

I hvilken kategori bør vi sette kunden?

Kategorivariabler

Variabler der verdiene faller i to eller flere grupper.

Dikotome variabler: To grupper

Eksempler

Ja/nei, mann/kvinne, godkjent/feil.

Nominale variabler: To eller flere grupper

Eksempler

Ja/nei/vet ikke, norsk/dansk/svensk/finsk.

Ordinale variabler: To eller flere grupper i en bestemt rekkefølge

Eksempler

lav/middels/høy, god/middels/svak/dårlig.

Optimal klassifikasjon

Finn en regel for klassifikasjonen slik at forventet resultat blir best mulig.

Vi kjenner inntekt (lav/middels/høy) og sivilstand (1/2) for en kunde. Skal vi gi kunden lån eller ikke?

Må kjenne konsekvensene av klassifikasjon.

Tap i 1000 kr ved klassifikasjon		
Klassifikasjon	Kunde	
	God	Dårlig
God	0	20
Dårlig	10	0

Må kjenne data for fordeling.

Sivilstand 1		
Inntekt	God	Dårlig
Lav	90	50
Middels	360	40
Høy	270	30
Sivilstand 2		
Inntekt	God	Dårlig
Lav	270	50
Middels	540	20
Høy	270	10

Det er ialt 6 ulike kategorier, og vi har 2000 observasjoner. Vi antar at andelene er representative for hele befolkningen.

Hva er forventet tap ved å gi lån til en kunde med lav inntekt i sivilstand 1?

$$0 \cdot \frac{90}{140} + 20 \cdot \frac{50}{140} = 7,14$$

Hva er forventet tap ved å ikke gi lån til en kunde med lav inntekt i sivilstand 1?

$$10 \cdot \frac{90}{140} + 0 \cdot \frac{50}{140} = 6,42$$

Handling:

Gi lån: Forventet tap 7,14

Ikke gi lån: Forventet tap 6,42

Optimalt å ikke gi lån! De 5 andre gruppene kan analyseres på samme måte. Det viser seg at alle disse bør gis lån.

Hvilke tall er det vi bruker?

$$0 \cdot \frac{90}{140} + 20 \cdot \frac{50}{140} = 7,14$$

$$10 \cdot \frac{90}{140} + 0 \cdot \frac{50}{140} = 6,42$$

Generelt

$$0 \cdot P(\text{God}|\text{Lav}/1) + L_2 \cdot P(\text{Dårlig}|\text{Lav}/1)$$

$$L_1 \cdot P(\text{God}|\text{Lav}/1) + 0 \cdot P(\text{Dårlig}|\text{Lav}/1)$$

Vi bør dermed gi lån hvis og bare hvis

$$L_2 \cdot P(\text{Dårlig}|\text{Lav}/1) \leq L_1 \cdot P(\text{God}|\text{Lav}/1)$$

Hvis vi følger den optimale handlingsregelen, hvor mange % får lån?

	God	Dårlig
Klassifisert God	1710	150
Klassifisert Dårlig	90	50

Prosenter av de 2000 personene:

	God	Dårlig
Klassifisert God	85,5%	7,5%
Klassifisert Dårlig	4,5%	2,5%

Forventet tap pr klassifikasjon blir dermed

$$\begin{aligned} & L_1 \cdot 4,5\% + L_2 \cdot 7,5\% \\ =& 10 \cdot 4,5\% + 20 \cdot 7,5\% = 1,95 \end{aligned}$$

Hva ville skjedd hvis vi bare hadde gjort bruk av en av variablene, for eksempel inntekt. Forventet tap ved lav inntekt blir:

$$0 \cdot \frac{360}{460} + 20 \cdot \frac{100}{460} = 4,35$$

$$10 \cdot \frac{360}{460} + 0 \cdot \frac{100}{460} = 7,83$$

Nå blir tapet for stort ved ikke å gi lån, så i dette tilfellet bør vi gi alle kundene lån.

Hvor mye vi kan vi forvente å tape på det? Det er ialt 200, det vil si 10% dårlige kunder. Det

vil si at vi taper

$$L_1 \cdot 0\% + L_2 \cdot 10\%$$
$$= 10 \cdot 0\% + 20 \cdot 10\% = 2,00$$

Det betyr at tapene øker med kun 2,6% ved å gi alle kundene lån. Andre forhold som vi har sett bort fra, kan kanskje gjøre at denne strategien faktisk er den beste.

Data øvelse

- Last ned filen 'bankkunder'.

Anta at potensielle tap ved er gitt ved

Tap i kr ved klassifikasjon		
Klassifikasjon	Kunde	
	God	Dårlig
God	0	40000
Dårlig	10000	0

- Finn den optimale handlingsregelen

• Finn forventet gevinst når du følger den optimale handlingsregelen

Noen kommentarer til øvingen

Det sentrale med øvingen er å bruke R til å finne antall kunder i de ulike kategoriene. Ved å bruke kommandoen

```
> inntN=cut(c(0,150000,500000,max(innt)))
```

kan vi komme frem til tabellen

$$\begin{pmatrix} 207 & 736 \\ 258 & 2230 \\ 4 & 116 \\ 227 & 185 \\ 140 & 858 \\ 8 & 31 \end{pmatrix}$$

Resten er det lettest å regne for hånd. Med disse tallen blir den optimale handlingsregelen å gi lån til gifte med middels og høy inntekt, og til ugifte med middels inntekt.

	God	Dårlig
Klassifisert God	64,1%	8,0%
Klassifisert Dårlig	19,1%	8,8%

Forventet tap pr klassifikasjon blir da kr 5120.

For å undersøke om intervallgrensene er gode, kan vi forsøke å endre grensene.

Eksempel

> inntN=cut(c(0,160000,500000,max(innt)))

gir tabellen

$$\begin{pmatrix} 225 & 866 \\ 240 & 2100 \\ 4 & 116 \\ 230 & 254 \\ 137 & 789 \\ 8 & 31 \end{pmatrix}$$

Den optimale handlingregelen blir den samme som før, men andelene som blir klassifisert som korrekte blir endret.

	God	Dårlig
Klassifisert God	60,1%	7,6%
Klassifisert Dårlig	23,0%	9,3%

Forventet tap pr klassifikasjon blir da kr 5350. Tilsvarende tester antyder at grensene ikke kan forbedres!!!

Mer om grunnleggende teori

Varians

$$\text{Var}[X] = \text{E}[(X - \text{E}[X])^2] = \text{E}[X^2] - \text{E}[X]^2$$

Standardavvik

$$\sigma[X] = \sqrt{\text{Var}[X]}$$

Eksempel

X er antall enheter selges av en bestemt vare. Fordelingen til X:

$$p(0) = 0,2 \quad p(1) = 0,2 \quad p(2) = 0,6$$

Finn forventning $\mathrm{E}[X]$, varians $\mathrm{Var}[X]$ og standardavvik $\sigma[X]$.

Løsning

$$\mathrm{E}|X] = 0 \cdot 0,2 + 1 \cdot 0,2 + 2 \cdot 0,6 = 1,4$$

$$\begin{aligned}
&\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 \\
=&0^2 \cdot 0,2 + 1^2 \cdot 0,2 + 2^2 \cdot 0,6 - 1,4^2 \\
=&0,64
\end{aligned}$$

$$\sigma[X] = \sqrt{\mathrm{Var}[X]} = \sqrt{0,64} = 0,8$$

Viktige regler:

$$\mathrm{Var}[k \cdot X] = k^2 \mathrm{Var}[X] \qquad \mathrm{Var}[X + k] = \mathrm{Var}[X]$$

Estimering og estimatorer

Inferens: å beregne verdier for ukjente størrelser ved hjelp av observerte data.

Estimator $\hat{\theta}$: en tilfeldig variabel som vi bruker til å anslå verdien på et ukjent tall θ.

Den kanskje mest brukte estimatoren er gjennomsnittet:

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$

Den brukes til å estimere verdien til forventningen μ. Hvis $X_1, \ldots, X_n$ alle har samme forventning, μ, da er også

$$\mathrm{E}[\overline{X}] = \mu$$

Når forventningen til estimatoren er lik verdien på det tallet vi skal estimere, sier vi at estimatoren er *forventningsrett*. I motsatt fall kaller vi estimatoren *forventningsskjev*.

En forventningsrett estimator med liten varians er effektiv fordi det er svært sannsynlig at verdien ligger nær det tallet som vi skal estimere.

Hvis $X_1, \ldots, X_n$ alle er uavhengige med samme forventning og samme varians, σ^2, da er

$$\mathrm{Var}[\overline{X}] = \frac{\sigma^2}{n}$$

Jo flere observasjoner, jo bedre estimator. Som regel må vi også estimere variansen σ^2, og da kan vi bruke estimatoren

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Her kan en vise at

$$\mathrm{E}[S^2] = \sigma^2$$

så S^2 er altså forventningsrett.

Målemodellen

Et oppsett der vi gjør uavhengige observasjoner $X_1, \ldots, X_n$ av samme tilfeldige variabel, og bruker

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

til å estimere μ og σ^2, kaller vi gjerne en målemodell.

Lotterimodellen

I noen sammenhenger kan vi ikke anta at observasjonene er uavhengige. I en spørreundersøkelse kan vi for eksempel ikke spørre samme person flere ganger. I en slik situasjon må vi ta hensyn til størrelsen på populasjonen N og størrelsen på utvalget n.

Anta at populasjonen består av elementene, $\nu_1, \ldots, \nu_N$, med

$$\overline{\nu} = \frac{1}{N}(\nu_1 + \cdots + \nu_N)$$

og

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(\nu_i - \overline{\nu})^2$$

Hvis vi trekker tilfeldig (uten tilbakelegging) n elementer $Y_1, \ldots, Y_n$, da er

$$\overline{Y} = \frac{1}{n}(Y_1 + \cdots + Y_n)$$

en forventningsrett estimator $\overline{\nu}$ og

$$\mathrm{Var}[\overline{Y}] = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$$

I dette tilfellet er

$$S^2 = \frac{N}{N-1} \cdot \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

en forventningsrett estimator for σ^2. Dette oppsettet kalles en lotterimodell.

Hypotesetesting

• En hypotesetest har følgende elementer

- En nullhypotese H_0
- En alternativ hypotese H_A
- En testobservator $\hat{\theta}$
- Et forkastningsområde A

Testen gjennomføres ved at en observerer $\hat{\theta}$. En skal forkaste nullhypotesen H_0 til fordel for H_A dersom $\hat{\theta}$ ligger i forkastningsområdet.

• Signifikansnivå: En hypotesetest har α% signifikansnivå når det er høyst α% sjanse for at

en forkaster H_0 når H_0 er korrekt. Dette kalles forkastningsfeil eller type I feil.

• Styrke/godtakingsfeil: Et alternativ θ i H_A har styrke $\Pi(\theta) = 1 - \beta\%$ dersom det er høyst $\beta\%$ sjanse for å godta H_0 når H_0 er feil mens alternativet er korrekt. Dette kalles godtakingsfeil eller type II feil. Stor styrke gir altså liten sannsynlighet for godtakingsfeil.

Eksempel

Gebyrbruk i en bank. Vi antar at det totale gebyret X (i kr) til en tilfeldig valgt kunde har forventning μ og varians $\sigma^2 = 160000$. Ifjor var det gjennomsnittlige gebyret 1200 (kr). For å teste om gebyrbruken har gått opp, har vi sjekket forbruket hos 400 tilfeldig valgte husstander. Vi observerer $\overline{X} = 1300$ (kr). Hvilken konklusjon trekker vi?

Løsning

Her er $H_0 : \ \mu = 1200$ og $H_A : \ \mu > 1200$.

Som testestimator benytter vi

$$\overline{X} = \frac{1}{400}(X_1 + X_2 + \cdots + X_{400})$$

Vi vet at

$$\mathrm{E}[\overline{X}] = \mu \qquad \sigma[\overline{X}] = \sqrt{\frac{\sigma^2}{n}} = 20$$

Siden $n = 400$ er stor, kan vi bruke normaltilnærming. Vi vil forkaste H_0 når $\overline{X}$ er usannsynlig stor. Med 5% signifikansnivå blir grensen:

$$x_{\text{grense}} = 1200 + 1,645 \cdot 20 = 1232,90$$

Forkastningsområdet er dermed intervallet

$$A = [1232,90, \infty)$$

Siden $\overline{X} = 1300$ ligger i forkastningsområdet, forkaster vi nullhypotesen. Det er dermed svært sannsynlig at gebyrbruken har gått opp.

Motsatt betraktning

$H_0: \mu > 1200$ og $H_A: \mu = 1200$

$$\overline{X} = \frac{1}{400}(X_1 + X_2 + \cdots + X_{400})$$

Vi forkaster H_0 når $\overline{X}$ er usannsynlig liten. Med 5% signifikansnivå blir grensen:

$$x_{\text{grense}} = 1200 - 1,645 \cdot 20 = 1167,10$$

$$A = (-\infty, 1167,10]$$

I dette tilfellet vil vi beholde H_0. Vi kan ikke med rimelig sikkerhet si at gebyrbruken ikke har gått opp. Det er en helt annen og mye vagere konklusjon!

Når en velger nullhypotese, er det derfor viktig å klargjøre hvilke konklusjoner en er interessert i å trekke.

• P-verdi: P-verdien til en test måles fra den observerte verdien til estimatoren. Med utgangspunkt i at nullhypotesen er korrekt, skal

vi regne ut hvor ofte vi kan få observasjoner som avviker minst like mye fra H_0 som den verdien som vi nettopp har observert.

Eksempel

Vi gjennomfører et forsøk, og finner at P-verdien er 0.1%. Det vil si at hvis H_0 er sann, vil vi kun i et av 1000 tilfeller observere verdier som avviker like mye fra H_0 som den verdien vi nettopp har observert. Dette gjør det vanskelig å tro at avviket er tilfeldig, og vi vil stort sett alltid forkaste H_0.

Eksempel

Vi gjennomfører et forsøk, og finner at P-verdien er 90%. Det vil si at hvis H_0 er sann, vil vi i 9 av 10 tilfeller observere verdier som avviker like mye fra H_0 som den verdien vi nettopp har observert. Dette er et avvik som inntreffer så ofte at vi stort sett aldri vil forkaste H_0 på dette grunnlaget.

Eksempel

Vi gjennomfører et forsøk, og finner at P-verdien er 4%. Det vil si at hvis H_0 er sann, vil vi kun i et av 25 tilfeller observere verdier som avviker like mye fra H_0 som den verdien vi nettopp har observert.

Dette er ganske sjeldent, men kan likevel skyldes tilfeldigheter. Konklusjonen vil i dette tilfellet være avhengig av hvilket signifikansnivå vi har valgt for forsøket. Med 5% signifikansnivå vil vi forkaste H_0, mens vi beholder H_0 på 2.5% signifikansnivå.

Prinsipper

Liten P-verdi	$\leftrightarrow$	H_0 er lite sannsynlig
Stor P-verdi	$\leftrightarrow$	H_0 er for sannsynlig

Forkast H_0 når P-verdien er mindre enn signifikansnivået.

Eksempel

Et firma reklamerer med at høyst 10% av kundene er misfornøyde. Vi tror på denne

påstanden, og spør 400 tilfeldige kunder om de er misfornøyde. La X være antall i utvalget som svarer at de er misfornøyd. Vi observerer $X = 53$. Hva er P-verdien til testen?

Løsning

La p være sannsynligheten for at en tilfeldig valgt kunde er misfornøyd. Vi har

$$H_0 : \ p \leq 10\% \qquad H_A : \ p > 10\%$$

X er binomisk fordelt med parametere

$$n = 400 \qquad p = p$$

P-verdien til testen finner vi ved

$$\begin{aligned} P_{H_0}(X \geq 53) &\leq P_{p=10\%}(X \geq 53) \\ &= 1 - P_{p=10\%}(X \leq 52) \\ &\approx 1 - G\left(\frac{52-40}{6}\right) \\ &= 1 - G(2) = 1 - 0,9772 \\ &= 2,28\% \end{aligned}$$

Siden P-verdien er liten, er det usannsynlig at nullhypotesen er riktig. Det er statistisk grunnlag for å avvise at høyst 10% av kundene er misfornøyde.

Estimering av sannsynligheter

Eksempel

En bedrift vurderer å tilby en serviceavtale til et produkt.

Vurdering: Lønnsomt bare hvis mer enn 20% av kjøperne tegner en slik avtale.

Test: Spør 40 kjøpere, og 11 svarer at de ville tegnet en slik avtale.

Spørsmål: Kan vi ut fra denne testen være rimelig sikker på at minst 20% av kjøperne tegner en slik avtale?

Teoretisk analyse: Vi antar at en andel p av kjøperne vil tegne serviceavtalen, og at kundene vurderer dette uavhengig av hverandre. Vi spør n kunder, og lar

$$X = \text{Antall kunder som ønsker avtale}$$

Da er X binomisk (n, p), og

$$\hat{p} = \frac{X}{n}$$

er forventningsrett for p med

$$\text{Var}[\hat{p}] = \frac{p(1-p)}{n}$$

Hvor stor sannsynlighet er det for at $X \geq 11$ gitt at $p \leq 0,20$? Sannsynligheten blir større jo større p er, og størst når $p = 0,2$. Hvis $p = 0,2$, er

$$\begin{aligned} P(X \geq 11) &= \sum_{i=11}^{40} P(X = i) \\ &= \sum_{i=11}^{40} \binom{40}{i} 0,2^i (1-0,2)^{40-i} \\ &= 16,08\% \end{aligned}$$

Selv om $\hat{p} = 11/40 = 27,5\%$ er godt over 20%, er det likevel en ganske stor sjanse for at prosjektet vil gå med tap. Testen er ikke god nok!

Hvor stor må n være for at vi kan være rimelig sikre på konklusjonene?

• Ønsker høyst $\alpha = 5\%$ sjanse for forkastningsfeil.

I dette tilfellet vil det si at vi i høyst 5% av tilfellene tror at $p > 0,2$ når det i virkeligheten er slik at $p \leq 0,2$.

• Ønsker høyst $\beta = 5\%$ sjanse for godtakingsfeil, når alternativet er så godt som $p = 0,3$.

I dette tilfellet vil det si at vi i høyst 5% av tilfellene tror at $p \leq 0,2$ når det i virkeligheten er slik at $p = 0,3$.

For en ensidig test kan vi bruke formelen:

$$n = \left(\frac{k_\alpha \sqrt{p_0(1-p_0)} + k_\beta \sqrt{p_A(1-p_A)}}{p_A - p_0} \right)^2$$

der $G(k_\alpha) = \alpha$ og $G(k_\beta) = \beta$. Det gir

$$n = \left(\frac{1,645\sqrt{0,2 \cdot 0,8} + 1,645\sqrt{0,3 \cdot 0,7}}{0,3 - 0,2} \right)^2$$
$$= 200$$

Oppsettet må modifiseres hvis vi har en tosidig test.

Estimering av andeler

En bedrift har pr idag 210 kunder med en service avtale, og vurderer å endre denne avtalen. Avtalen bør bare endres hvis et flertall av kundene ønsker å endre avtalen.

Test: Spør 40 kunder om de foretrekker den nye avtalen, 26 svarer Ja.

Spørsmål: Kan vi ut fra denne testen være rimelig sikre på at flertallet av kundene foretrekker den nye avtalen?

Teoretisk analyse: Denne problemstillingen skiller seg fra den forrige ved at vi spør en ikke

uvesentlig andel av alle kundene. Det gir en hypergeometrisk fordeling.

Forutsetninger: N kunder, M av disse foretrekker avtalen. Velger et tilfeldig utvalg på n kunder, og observerer Y kunder som foretrekker den nye avtalen.

Y er hypergeometrisk (N, M, n). Estimatoren

$$\hat{a} = \frac{Y}{n}$$

er en forventningsrett estimator for andelen $\frac{M}{N}$, og

$$\sigma[\hat{a}] = \sqrt{\frac{N-n}{N-1} \cdot \frac{a(1-a)}{n}}$$

Når $N = 210, M = 105$ og $n = 40$, finner vi

$$P(Y \geq 26) = 2,6\%$$

Det virker da svært sannsynlig at $M > 105$.

Legg merke til faktoren $\sqrt{\frac{N-n}{N-1}}$. Vi kan tenke på denne faktoren som en korreksjon i forhold til

den binomiske situasjonen. Her er

$$\sqrt{\frac{N-n}{N-1}} = 0,9$$

Estimering av forventning

En bedrift vurderer å tilby et produkt med og uten serviceavtale.

Spørsmål: Hvor mye er kunder som kjøper en slik avtale villige til å betale i gjennomsnitt?

Test: Vi spør 40 kunder. 20 kunder ønsker avtale, og blant disse finner vi en gjennomsnittlig betalingsvillighet på 252,50 kr.

Teoretisk analyse: X er betalingsvilligheten til en tilfeldig kunde som ønsker serviceavtale. Hvilken verdi har $\mu = \mathrm{E}[X]$?

Estimatorer:

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

For å vurdere en slik problemstilling, kan t fordelingen være aktuell. Vi har ganske få observasjoner, $n = 20$, så normalfordelingen er litt unøyaktig.

Antar vi at X-ene er tilnærmet normalfordelte, kan vi regne ut nøyaktige svar. Da er nemlig

$$\frac{\overline{X} - \mu}{S[\overline{X}]}$$

t fordelt med parameter $\nu = n - 1 = 19$. Et 95% konfidensintervall finner vi da ved hjelp av t tabellen på 2,5% nivå. Konfidensintervallet er avgrenset av

$$\overline{X} \pm 2,093 \cdot S[\overline{X}]$$

For å regne ut disse verdiene, trenger vi å vite at

$$S[\overline{X}] = \sqrt{\frac{S^2}{n}}$$

Verdien til S^2 må beregnes ut fra observasjonene. Hvis vi for eksempel finner

$$S^2 = 10650$$

blir

$$S[\overline{X}] = \sqrt{\frac{10650}{20}} = 23,1$$

som gir oss konfidensintervallet

$$252,5 \pm 2,093 \cdot 23,1$$

det vil si intervallet

$$[204\ ,\ 301]$$

Dersom antagelsen om tilnærmet normalfordeling er rett, er det altså 95% sikkert at forventet betalingsvillighet ligger i dette intervallet.

Vi merker oss at svaret ikke blir veldig forskjellig dersom vi istedet bruker normaltilnærming

med faktoren 1,96. I mange situasjoner kan det være slik at

$$P\left(\frac{X-\mu}{S[X]} \leq z\right) \neq G(z)$$

Likevel kan konfidensintervallene bli ganske like fordi begge disse funksjonene raskt nærmer seg 0, og dermed bryter igjennom verdien som svarer til signifikansnivået omtrent samtidig.

Estimering av gjennomsnitt

En reklamebedrift har 210 kunder, og vi spør 20 tilfeldig valgte kunder om hvor mye de har satt av til innkjøp av reklame. Gjennomsnittet blant de 20 kundene var 1,2 millioner kroner med et empirisk standardavvik på 0,4 millioner kr.

Problemstilling: Hva er gjennomsnittet i hele populasjonen?

Problemstillingen er svært lik den forrige problemstillingen, men her må vi igjen ta hensyn til

at vi har en hypergeometrisk situasjon. Konfidensintervallet blir estimert til

$$\overline{Y} \pm k \cdot \sqrt{\frac{N-n}{N-1}} \cdot \frac{S}{\sqrt{n}}$$

der k er antall standardavvik som kreves for et angitt sikkerhetsnivå. Selv om normaltilnærmingen er ganske grov her, så gir den likevel omtrent riktig størrelse på k. Uttrykket

$$\overline{Y} \pm 1,96 \cdot \sqrt{\frac{N-n}{N-1}} \cdot \frac{S}{\sqrt{n}}$$

vil derfor ikke være vesentlig forskjellig fra et 95% konfidensintervall. I denne situasjonen kan vi ikke bruke t fordelingen.

U-testen

U-testen brukes til å til å teste om det er forskjell på sannsynligheten for suksess i to uavhengige binomiske forsøksrekker.

Eksempel

Vi spør 40 menn og 60 kvinner om de synes statistikk er et viktig fag.

Sannsynligheten p_1 for svaret "Ja" blant menn er den ene sannsynligheten og sannsynligheten p_2 for svaret "Ja" blant kvinner er den andre.

Vi har en rekke med 40 binomiske forsøk for mennene og en rekke med 60 binomiske forsøk for kvinnene.

Generell diskusjon

$$H_0\colon\ p_1 = p_2$$
$$H_A\colon\ p_1 \neq p_2$$

X_1 = Antall suksesser i den første rekken

X_2 = Antall suksesser i den andre rekken

Vi definerer

$$\hat{p}_1 = \frac{X_1}{n_1} \qquad \hat{p}_2 = \frac{X_1}{n_2}$$

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

der n_1 og n_2 er antall forsøk i de to rekkene. Siden X_1 og X_2 er uavhengige binomiske vari-

abler, er

$$\begin{aligned}\mathrm{Var}[\hat{p}_1 - \hat{p}_2] &= \mathrm{Var}[\hat{p}_1] + \mathrm{Var}[\hat{p}_2] \\ &= \frac{1}{n_1^2}\mathrm{Var}[X_1] + \frac{1}{n_2^2}\mathrm{Var}[X_2] \\ &= \frac{1}{n_1} p_1(1-p_1) \\ &+ \frac{1}{n_2} p_2(1-p_2)\end{aligned}$$

Antar vi at

$$p_1 = p_2 = p$$

er

$$\mathrm{Var}[\hat{p}_1 - \hat{p}_2] = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) p(1-p)$$

Når n_1 og n_2 er store, er X_1 og X_2 tilnærmet normalfordelte, og

$$U = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}}$$

vil være tilnærmet standard normalfordelt.

Vi forkaster H_0 hvis U ligger for langt vekk fra null. Med 5% signifikansnivå bruker vi normalfordelingstabellen med 2, 5% til å finne forkastningsgrensene

$$U \leq -1,96 \qquad U \geq 1,96$$

Eksempel

Vi spurte 40 menn og 60 kvinner om de syntes statistikk var et viktig fag. Svarene var som i tabellen under

En spørreundersøkelse			
Menn		Kvinner	
40		60	
Nei	Ja	Nei	Ja
11	29	14	46

Vi har

$$\hat{p}_1 = \frac{11}{40} \qquad \hat{p}_2 = \frac{14}{60}$$

og

$$\hat{p} = \frac{11 + 14}{40 + 60} = \frac{25}{100}$$

Disse størrelsene setter vi inn, og regner ut

$$U = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\hat{p}(1 - \hat{p})}} = 0,47$$

Siden verdien ligger utenfor forkastningsområdet, beholder vi H_0. Vi kan ikke påstå at det høyst sannsynlig er forskjell mellom kjønnene.

Hva er P-verdien i dette tilfellet? Det er en tosidig test, så P-verdien bestemmes ved

$$\begin{aligned} P\text{-verdi} &= P(U \geq 0,47) + P(U \leq -0,47) \\ &= 2P(U \geq 0,47) \\ &= 2 - 2P(U \leq 0,47) = 2 - 2G(0,47) \\ &= 2 - 2 \cdot 0,6808 = 63,8\% \end{aligned}$$

Kjikvadrat-tester

Denne testen finnes i mange varianter. I den enkleste versjonen har vi et utfallsrom med m mulige utfall, og vil teste om gitt en modell med spesifiserte sannsynligheter $p_1, \ldots, p_m$ er OK eller ikke.

H_0 : Sannsynlighetene er $p_1, \ldots, p_m$

H_A : Sannsynlighetene er ikke $p_1, \ldots, p_m$

Testestimator:

$$Q = \sum_{i=1}^{m} \frac{(X_i - np_i)^2}{np_i}$$

der X_i er antall utfall i kategori i i løpet av n uavhengige forsøk.

Nullhypotesen forkastes når $Q \geq Q_{\text{grense}}$ der Q_{grense} finnes fra signifikansnivået i kjikvadrat-tabellen med parameter $\nu = m - 1$.

Eksempel

I en markedsundersøkelse har vi bedt 500 personer om å velge hvilket av fire mulige produkter som de liker best. Vi fikk følgende fordeling:

Hvilket produkt likte du best?	
Produkt	Antall
Produkt A	129
Produkt B	140
Produkt C	113
Produkt D	118

Bruk en Q-test til å teste nullhypotesen om at alle de fire produktene er like godt likt. Bruk 5% signifikansnivå.

Løsning

$$H_0 : p_1 = p_2 = p_3 = p_4 = 25\%$$

Forkastningsgrense: Kjikvadrattabellen med 3 frihetsgrader på 5% nivå gir

$$Q_{\text{grense}} = 7,81$$

Vi regner ut

$$Q = \frac{(129-125)^2}{125} + \frac{(140-125)^2}{125} + \frac{(113-125)^2}{125} + \frac{(118-125)^2}{125} = 3,472$$

Siden $Q < Q_{\text{grense}} = 7,81$, gir ikke testen grunnlag for å påstå at noen av produktene er bedre likt enn de andre.

Slik utføres testen i R:

```
> data=c(129,140,113,118)
> sannsynligheter=c(25,25,25,25)/100
> chisq.test(data,p=sannsynligheter)

Chi-squared test for given probabilities

data: data
X-squared = 3.472, df = 3, p-value = 0.3244
```

Toveistabeller og kjikvadrattest for uavhengighet

I en toveistabell skal hvert objekt klassifiseres ved to ulike kjennetegn, og hvert kjennetegn finnes blant to eller flere kategorier.

Eksempel

Hver kunde klassifiseres etter sivilstand og kredittvurdering. Sivilstand er i kategoriene 1 eller 2, mens kredittvurderingen er i kategoriene A, B eller C. Det gir ialt 6 kombinasjoner.

En toveistabell lages ved at vi gjør n uavhengige observasjoner, og observerer hvor mange observasjoner som faller innenfor de ulike kombinasjonene.

Eksempel

Vi undersøker $n = 1000$ kunder, og finner

Sivilstand/kredittvurdering		
	1	2
A	200	300
B	140	260
C	40	60

Vi ønsker å teste om kjennetegnene er uavhengige. Det kan for eksempel tenkes at kunder i sivilstand 1 er systematisk forskjellige fra kunder i sivilstand 2, eller at kunder med kredittvurdering A er systematisk forskjellige fra kunder med kredittvurdering B eller C.

Hvis kjennetegnene er uavhengige, fins det sannsynligheter q_1, q_2 og q_A, q_B, q_C slik at

$$p_{ij} = q_i \cdot q_j \qquad i = 1, 2,\ j = A, B, C$$

Her kan vi tolke q_1 som sannsynligheten for at en kunde er i sivilstand 1, q_2 som sannsynligheten for at en kunde er i sivilstand 2, q_A som sannsynligheten for at en kunde har kredittvurdering A, osv.

Test for uavhengighet i R:

```
> A=c(200,300)
> B=c(140,260)
> C=c(40,60)
> chisq.test(data.frame(A,B,C))

Pearson's Chi-squared test

data: data.frame(A, B, C)
X-squared = 2.5467, df = 2, p-value = 0.2799
```

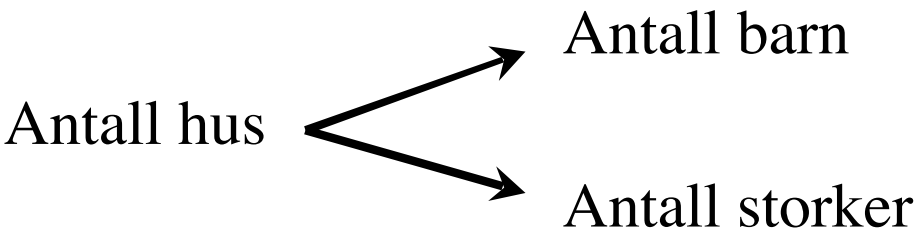
Årsaksforhold

En test av en toveistabell kan brukes til å avgjøre om det finnes samvariasjon mellom to kjennetegn. Hvis vi kan forkaste nullhypotesen om uavhengighet, betyr det at det høyst sannsynlig finnes samvariasjon mellom de to kjennetegnene. Det behøver *ikke* bety at endringer i hyppighet av observasjoner kan forklares ved endringer i kjennetegn!

Eksempel

Ved en undersøkelse av danske småbyer ble det brukt to kjennetegn: antall barn og antall storker. Det viste seg at det bodde flere barn i byer med mange storker. Betyr det at antall storker forklarer hyppigheten av barn?

I dette tilfellet finnes et tredje kjennetegn; antall hus.



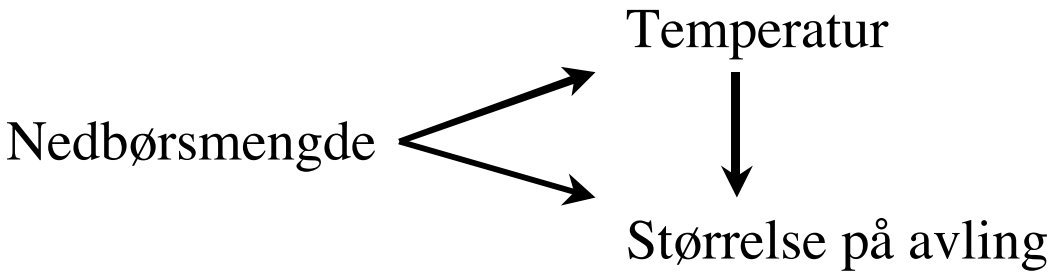
Antall hus forklarer antall barn og antall storker.

Eksempel

I en undersøkelse av avlinger ble det brukt to kjennetegn: temperatur og størrelse på avling. Det viste seg at det ble større avlinger ved lav temperatur. Betyr det at lav temperatur forklarer gode avlinger?

I dette tilfellet finnes et tredje kjennetegn, nedbørsmengde. Nedbørsmengden forklarer

både temperaturen og størrelsen på avlingene. Ved samme nedbørsmengde, får vi større avlinger ved høy temperatur.



Paradokser

En observert samvariasjon kan forsvinne ved innføringen av et nytt kjennetegn.

En observert uavhengighet kan forsvinne ved innføringen av et nytt kjennetegn.

En observert samvariasjon kan bytte retning ved innføringen av et nytt kjennetegn.

Tester for målevariabler

En kjikvadrattest er hensiktsmessig kun når antall kombinasjoner er relativt få. Variabler som måles på en skala, kaller vi målevariabler. For slike variabler trenger vi andre tester.

For å avgjøre om det finnes samvariasjon mellom to målevariabler X og Y, kan vi gjøre n uavhengige observasjoner

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

og regne ut den empiriske korrelasjonskoeffesienten

$$\rho_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$

der S_{XY} er empirisk kovarians, mens S_X og S_Y er empirisk standardavvik.

$$S_{XY} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

$$S_X = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2} \quad S_Y = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \overline{Y})^2}$$

For å teste for samvariasjon vil vi teste

$$H_0 : \rho_{XY} = 0 \text{ mot } H_A : \rho_{XY} \neq 0$$

Det gjelder å bestemme en forkastningsgrense ρ_{grense} slik at

$$P_{H_0}(|\rho_{XY}| \geq \rho_{\text{grense}}) = \text{Signifikansnivå}$$

Hvis paret (X, Y) er en såkalt binormalfordeling, er

$$T = \frac{\rho_{XY}}{\sqrt{1 - \rho_{XY}^2}} \cdot \sqrt{n-2}$$

t-fordelt med $\nu = n - 2$.

Eksempel

Anta at vi har gjort 30 uavhengige observasjoner av en binormalfordeling (X, Y), og har beregnet at korrelasjonskoeffesienten $\rho_{XY} = 0,3$. Kan vi påstå at det høyst sannsynlig er samvariasjon mellom X og Y?

Løsning

Vi regner ut

$$T = \frac{0,3}{\sqrt{1-0,3^2}} \cdot \sqrt{30-2} = 1,66$$

Vi vet at T er t-fordelt med parameter

$$\nu = 30 - 2 = 28$$

Bruker vi 5% signifikansnivå, blir forkastningsgrensen $\rho_{\text{grense}} = 2.048$. Vi kan dermed ikke forkaste H_0, og kan ikke påstå at det høyst sannsynlig er samvariasjon.

Tilfeldige utvalg

Når en gjennomfører tester, er det viktig at objektene som testes representerer et tilfeldig utvalg. Testen bør om mulig gjennomføres på en slik måte at andre forhold ikke kan påvirke resultatet.

Ofte vil vi studere effekten av et nytt tiltak ved å sammenlikne resultatetene for to tilfeldig valgte grupper, en testgruppe som prøver det nye tiltaket og en kontrollgruppe.

En slik randomisering kan skje på to ulike måter:

• Testgruppen velges tilfeldig innenfor en fastlagt populasjon, der kontrollgruppen er resten av populasjonen. Dette kalles komplett randomisering.

• Populasjonen deles inn i blokker, og innenfor hver blokk velges et tilfeldig utvalg til testgruppen, mens resten danner kontrollgruppen. Dette kalles randomisering innen blokker.

Eksempel

Med randomisering: Blant 21 arbeidstagere velger vi 11 tilfeldige arbeidere til testgruppen, mens resten danner kontrollgruppen.

Uten randomisering: 11 arbeidere på arbeidsplass A benyttes til testgruppen, mens 10 arbeidere på arbeidsplass B benyttes til kontrollgruppen.

• Hvis testen med randomisering viser at testgruppen gjør det signifikant bedre enn kontrollgruppen, kan vi konkludere at det nye tiltaket har positiv effekt.

• Hvis testen uten randomisering viser at testgruppen gjør det signifikant bedre enn kontrollgruppen, kan vi istedet tenke oss at effekten skyldes systematiske forskjeller mellom de to arbeidsplassene.

Variansanalyse

Variansanalyse er en samlebetegnelse for teknikker der størrelsen på variansen brukes til å avgjøre om forskjeller i gjennomsnitt er signifikante eller ikke. Variansanalyse finnes i mange ulike versjoner. Variansanalyse kalles også for ANOVA (ANalysis Of VAriance).

Variansanalyse: To-utvalgsmodeller

For å studere forskjeller mellom to randomiserte utvalg kan vi gjøre variansanalyse på gjennomsnittsverdiene:

Teori

La $X_1, X_2, \ldots, X_r$ være de observerte verdiene i det ene utvalget, mens $Y_1, Y_2, \ldots, Y_s$ er de observerte verdiene i det andre utvalget. Anta

$$\mathrm{E}[X_i] = \mu_1 \quad \mathrm{Var}[X_i] = \sigma_1^2 \quad i = 1, \ldots, r$$

$$\mathrm{E}[Y_i] = \mu_2 \quad \mathrm{Var}[Y_i] = \sigma_2^2 \quad i = 1, \ldots, s$$

Da er

$$\hat{\delta} = \overline{X} - \overline{Y}$$

en forventningsrett estimator for $\delta = \mu_1 - \mu_2$, og

$$\mathrm{Var}[\hat{\delta}] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Som oftest må variansene estimeres av programmet. Antar vi at

$$\sigma_1 = \sigma_2 = \sigma$$

kan en vise at

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{r}(X_i - \overline{X})^2 + \sum_{i=1}^{s}(Y_i - \overline{Y})^2\right)$$

er forventningsrett, og vi har

$$\sigma[\hat{\delta}] \approx S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \stackrel{\mathrm{def}}{=} S[\hat{\delta}]$$

En test for

$$H_0 : \mu_1 = \mu_2 \text{ mot } H_A : \mu_1 \neq \mu_2$$

kan da utføres ved hjelp av testestimatoren

$$T = \frac{\hat{\delta}}{S[\hat{\delta}]}$$

som vil være t-fordelt med $n_1 + n_2 - 2$ frihetsgrader. Programvare kan imidlertid enkelt behandle den generelle situasjonen der $\sigma_1 \neq \sigma_2$.

Merk imidlertid at programvaren forutsetter at observasjonene er normalfordelte.

Eksempel

Metode 1 gav resultatene:

$$4,7 \quad 3,5 \quad 3,3 \quad 4,2 \quad 3,6$$

Metode 2 gav resultatene:

$$3,2 \quad 4,2 \quad 3,3 \quad 3,9 \quad 3,0$$

Antar vi normalfordeling og $\sigma_1 = \sigma_2 = \sigma$, kan vi regne ut

$$\overline{X} = 3,86 \quad \overline{Y} = 3,52 \quad \hat{\delta} = 0,34$$

$$S = 0,543 \quad S[\hat{\delta}] = 0,344$$

$$T = \frac{\hat{\delta}}{S[\hat{\delta}]} = \frac{0,34}{0,344} = 0,99$$

Her er T t-fordelt med 8 frihetsgrader, så forkastningsgrensen på 5% nivå for en tosidig test er 2,306. Vi beholder derfor H_0.

Slik utføres testen i R:

```
> x=c(4.7,3.5,3.3,4.2,3.6)
> y=c(3.2,4.2,3.3,3.9,3.0)
> t.test(x,y,alternative="two.sided", var.equal = TRUE)

Two Sample t-test

data: x and y
t = 0.9898, df = 8, p-value = 0.3513
```

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.4521384 1.1321384

sample estimates:

mean of x mean of y

3.86 3.52

Variansanalyse: Parrede observasjoner

Når vi har parrede observasjoner, kan vi utføre testene på en litt annen måte.

Teori

La $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ være de observerte verdiene. Anta

$$\mathrm{E}[X_i] = \alpha_i + \delta \quad \mathrm{Var}[X_i] = \sigma_1^2$$

$$\mathrm{E}[Y_i] = \alpha_i \quad \mathrm{Var}[Y_i] = \sigma_2^2$$

Da er $\hat{\delta} = \overline{X} - \overline{Y}$ en forventningsrett estimator for δ, og $\mathrm{Var}[\hat{\delta}] = \frac{1}{n}(\sigma_1^2 + \sigma_2^2)$. Her kan $\sigma^2 = \sigma_1^2 + \sigma_2^2$ estimeres ved

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - Y_i - \hat{\delta})^2$$

og tester gjennomføres ved at

$$T = \frac{\hat{\delta} - \delta}{S/\sqrt{n}}$$

er t-fordelt med $\nu = n - 1$ frihetsgrader. Teorien forutsetter normalfordelte observasjoner.

Slik kan testen utføres i R:

```
> x=c(4.7,3.5,3.3,4.2,3.6)
> y=c(3.2,4.2,3.3,3.9,3.0)
> t.test(x,y,paired=TRUE)
```

Paired t-test

data: x and y

t = 0.9408, df = 4, p-value = 0.4001

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.663369 1.343369

sample estimates:

mean of the differences

0.34

Variansanalyse: Modeller med en faktor

I modeller med en faktor ønsker vi å vite om observasjonene blir systematisk endret når vi endrer faktoren. For å kunne avgjøre dette, gjør vi et sett med observasjoner for hver mulig variant av faktoren.

Faktor: Variant 1 Variant 2 Variant 3...

Observ.: Sett 1 Sett 2 Sett 3...

De ulike settene trenger ikke å inneholde samme antall observasjoner.

I denne situasjonen ønsker vi å teste om forventningen er den samme i alle gruppene.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots$$

$$H_A : \text{Ikke alle er like}$$

Hvis vi kan forkaste H_A, har vi sannsynliggjort at faktoren påvirker resultatene.

Analysen utføres ved at en først finner gjennomsnittet for hver enkelt gruppe for seg:

$$\overline{Y}_1, \overline{Y}_2, \overline{Y_3}, \ldots$$

Disse verdiene ønsker vi å sammenlikne med gjennomsnittet av alle observasjonene $\overline{Y}$. Som mål på forskjellen benyttes

$$Q_A = \sum_{i=1}^{r} n_i (\overline{Y}_i - \overline{Y})^2$$

der r er antall grupper, og n_i antall observasjoner innenfor hver gruppe. Denne størrelsen ønsker vi å sammenlikne med

$$Q_0 = \sum_{i=1}^{r} (n_i - 1) S_i^2$$

der S_i^2 er den empiriske variansen innenfor hver gruppe. Testestimatoren er

$$F = \frac{Q_A/(r-1)}{Q_0/(n-r)}$$

der n er antall observasjoner totalt.

Når nullhypotesen gjelder, er F er såkalt Fisher-fordeling eller F-fordeling med frihetsgrader $(r-1, n-r)$. Forkastningsgrensene finnes derfor ved å benytte F-tabellen med disse parameterne.

Normalfordelte observasjoner er igjen en forutsetning for gyldigheten av analysen.

Eksempel Metode 1 gav resultatene:

$$4,7 \quad 3,5 \quad 3,3 \quad 4,2 \quad 3,6$$

Metode 2 gav resultatene:

$$3,2 \quad 4,2 \quad 3,3 \quad 3,9 \quad 3,0$$

Metode 3 gav resultatene:

$$3,1 \quad 2,9 \quad 2,2 \quad 3,0 \quad 2,8$$

Her er

$$\overline{Y}_1 = 3,86 \quad \overline{Y}_2 = 3,52 \quad \overline{Y_3} = 2,80$$

$$\overline{Y} = 3,39 \quad Q_A = 2,93 \quad Q_0 = 2,86$$

$$F = \frac{Q_A/(r-1)}{Q_0/(n-r)} = \frac{2,93/2}{2,86/12} = 6,15$$

Antar vi normalfordelte observasjoner, er F Fisherfordelt med frihetsgrader $(2,12)$. F-tabellen gir en forkastningsgrense på $3,89$, så vi kan forkaste H_0 og påstå at metoden høyst sannsynlig påvirker resultatene.

Slik kan testen utføres i R:

```
> verdier=scan()
1: 4.7 3.5 3.3 4.2 3.6
6: 3.2 4.2 3.3 3.9 3.0
11: 3.1 2.9 2.2 3.0 2.8
16:
Read 15 items
> metode=c(1,1,1,1,1,1,2,2,2,2,2,3,3,3,3,3)
> tabell=data.frame(metode=factor(metode), verdier)
> summary(aov(verdier~metode,data=tabell))
```

Df Sum Sq Mean Sq F value Pr(>F)

metode 2 2.92933 1.46467 6.1455 0.01454 *

Residuals 12 2.86000 0.23833

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Variansanalyse: Modeller med to faktorer

I modeller med to faktorer ønsker vi å vite om observasjonene blir systematisk endret når vi endrer faktorene. Her finnes det flere varianter, alt ettersom hvordan vi tror faktorene virker inn.

Additive effekter

I denne versjonen lar vi Y_{ij} være en tilfeldig variabel som angir resultatet når faktor 1 har verdien i og faktor 2 har verdien j. Vi trenger kun en observasjon for hver kombinasjon av faktorene. Tanken er at

$$\mathrm{E}[Y_{ij}] = \mu + \alpha_i + \beta_j$$

Det vil si at vi har et bidrag μ som ikke skyldes noen av faktorene, et bidrag α_i som skyldes faktor 1 og et bidrag β_j som skyldes faktor 2.

Mulige nullhypoteser kan for eksempel være:

$$H_0^{(1)} : \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$$

$H_0^{(1)}$ sier at faktor 1 ikke har noen betydning.

$$H_0^{(2)} : \beta_1 = \beta_2 = \cdots = \beta_s = 0$$

$H_0^{(2)}$ sier at faktor 2 ikke har noen betydning.

Teorien er svært lik teorien for en-faktor modellen.

Eksempel

En to-faktor modell				
Y_{ij}	$j = 1$	$j = 2$	$j = 3$	$j = 4$
$i = 1$	5,4	4,3	3,7	3,2
$i = 2$	6,3	6,0	5,7	4,5
$i = 3$	5,3	5,8	4,8	4,0

Slik kan testen utføres i R:

```
> verdier=scan()
1: 5.4 4.3 3.7 3.2 6.3 6.0 5.7 4.5 5.3 5.8 4.8 4.0
13:
Read 12 items
> faktor1=c(1,1,1,1,2,2,2,2,3,3,3,3)
> faktor2=c(1,2,3,4,1,2,3,4,1,2,3,4)
> tabell=data.frame(faktor1 =factor(faktor1),faktor2 =factor(faktor2),verdier)
> summary(aov(verdier~faktor1+faktor2,data =tabell))
            Df Sum Sq Mean Sq F value Pr(>F)
faktor1 2 4.3717 2.1858 15.833 0.004042 **
faktor2 3 5.4967 1.8322 13.272 0.004665 **
Residuals 6 0.8283 0.1381
—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Mer avanserte versjoner - samspill

Hvis vi har mere enn en observasjon for hvert par av de to faktorene, kan vi bestemme mer avanserte hypoteser.

$$Y_{ijk}$$

betyr at faktor 1 har verdi i, faktor 2 har verdi j, og vi har observasjon nr k av denne varianten. Vi kan nå studere effekter på formen

$$\mathrm{E}[Y_{ijk}] = \mu + \alpha_i + \beta_j + \phi_{ij}$$

Vi har et bidrag μ som ikke skyldes noen av faktorene, et bidrag α_i som skyldes faktor 1 og et bidrag β_j som skyldes faktor 2. I denne versjonen har vi i tillegg et ledd ϕ_{ij} som åpner for at faktorene gjensidig kan påvirke hverandre. Dette kalles samspill. Den naturlige nullhypotesen er

$$H_0^{(1,2)} : \text{Alle } \phi_{ij} = 0$$

Hvis denne nullhypotesen kan forkastes, kan vi påstå at vi høyst sannsynlig har samspill mellom faktorene.

Wilcoxons rang test for to utvalg

Blant 21 arbeidstagere velger vi 11 tilfeldige arbeidere til testgruppen, mens resten danner kontrollgruppen. Arbeidstidene til de to gruppene er vist i tabellen under:

Kontrollgruppe (arbeidstid i sekunder)

45 48 61 52 48 63 52 54 50 58

Testgruppe (arbeidstid i sekunder)

47 59 43 50 45 45 49 41 47 95 50

H_0: Arbeidstidene er like

H_A: Arbeidstiden er kortere for testgruppen

Rang: Sett alle observasjonene i stigende rekkefølge, og finn posisjonene til testgruppen.

<u>41</u> <u>43</u> <u>45</u> 45 <u>45</u> <u>47</u> <u>47</u> 48 48 <u>49</u> <u>50</u>

50 <u>50</u> 52 52 54 58 <u>59</u> 61 63 <u>95</u>

Wilcoxon estimatoren W finnes ved å summere posisjonene til testgruppen:

$$W = 1 + 2 + 3 + 5 + 6 + 7 + 10 + 11 + 13 + 18 + 21 = 97$$

Ideen er at hvis arbeidstiden er kortere for kontrollgruppen, så finnes disse observasjonene tidlig i rekken og verdien på W blir unormalt lav.

Merk at når enkelte verdier faller sammen, er det ikke klart hvem som skal settes først. Dataprogrammer justerer for dette ved å ta gjennomsnittet av alle mulige rekkefølger.

For å beregne utfallet av testen, må vi regne ut P-verdien

$$P_{H_0}(W \leq 97)$$

Teori: Når N er rimelig stor og gruppene ikke altfor ulike i størrelse, er W tilnærmet normalfordelt med

$$\mathrm{E}[W] = \frac{n(N+1)}{2}$$

$$\mathrm{Var}[W] = \frac{n(N-n)(N+1)}{12}$$

I vårt tilfelle blir

$$\mathrm{E}[W] = \frac{11 \cdot 22}{2} = 121$$

$$\sigma[W] = \sqrt{\frac{11 \cdot 10 \cdot 22}{12}} = 14,20$$

og vi finner

$$\begin{aligned} P_{H_0}(W \leq 97) &= G\left[\frac{97-121}{14,2}\right] \\ &= G[-1,69] = 0,0455 \end{aligned}$$

Verdien på W er derfor urimelig liten, og vi forkaster H_0 på 5% signifikansnivå.

Testen omtales i litteraturen som Wilcoxons test for to utvalg eller Mann-Whitneys test.

Slik utføres testen i R:

```
> x=c(45,48,61,52,48,63,52,54,50,58)
```

```
> y=c(47,59,43,50,45,45,49,41,47,95,50)
> wilcox.test(x,y,alternative="greater")

Wilcoxon rank sum test with continuity correction

data: x and y
W = 79, p-value = 0.04838
alternative hypothesis: true mu is greater than 0

Warning message:
Cannot compute exact p-value with ties in: wilcox.test.default(x, y, alternative =
"greater")
```

Merk at R regner ut W på en litt annen måte.

Wilcoxons test for parrede observasjoner

Alternativt kan vi teste en ny metode ved å la 10 arbeidere gjennomføre testen to ganger, en gang med ny metode og en gang med gammel metode. For å unngå at rekkefølgen påvirker

resultatet, velger hver arbeider tilfeldig hvilken metode som skal brukes først. I dette tilfellet ble resultatene slik:

Gammel metode (arbeidstid i sekunder):

48 53 52 57 43 83 59 71 40 61

Ny metode (arbeidstid i sekunder):

45 42 58 50 41 47 53 66 45 53

Ideen er nå at hvis den nye metoden er raskere, bør differensene

Tid gammel metode – Tid ny metode

jevnt over være positive. Differensene er

3 11 – 6 7 2 36 6 5 – 5 8

Rangen for denne testen bestemmer vi slik: Sett observasjonene i stigende tallverdi, og finn posisjonene til de negative leddene.

2 3 5 $\underline{-5}$ $\underline{-6}$ 6 7 8 11 36

Wilcoxon estimatoren W finnes ved å summere posisjonene til de negative leddene:

$$W = 4 + 5 = 9$$

Ideen er at at hvis den nye metoden er raskere, bør det være relativt få negative ledd, og de bør komme tidlig i rekken.

For å beregne utfallet av testen, må vi regne ut P-verdien

$$P_{H_0}(W \leq 9)$$

Testen omtales i litteraturen som Wilcoxons test for parrede utvalg.

Slik utføres testen i R:

```
> x=c(48,53,52,57,43,83,59,71,40,61)
> y=c(45,42,58,50,41,47,53,66,45,53)
> wilcox.test(x,y,alternative="greater", paired =TRUE)
```

Wilcoxon signed rank test with continuity correction

data: x and y

V = 46, p-value = 0.03310

alternative hypothesis: true mu is greater than 0

Warning message:

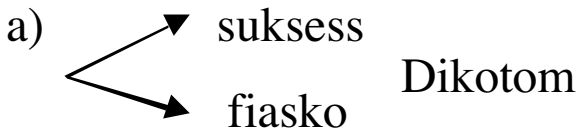
Cannot compute exact p-value with ties in: wilcox.test.default(x, y,

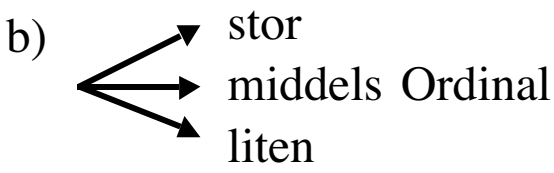
alternative = ”greater”, paired = TRUE)

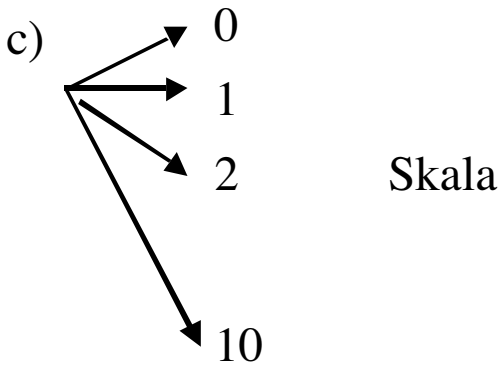
Merk at R regner ut W på en litt annen måte.

Stokastiske modeller

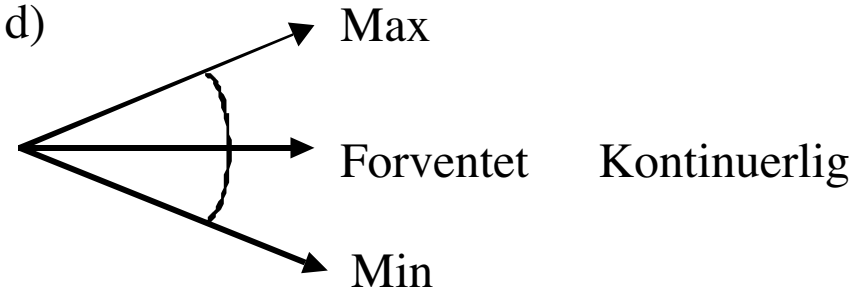
Diskret







Kontinuerlig



Klassiske diskrete fordelinger og deres egenskaper

Binær/Bernoulli/Indikator

Har utfall 0 og 1; $P(X = 1) = p$.

$$\mathrm{E}[X] = p \quad \mathrm{Var}[X] = p(1-p)$$

R-kommando: rbinom$(N, 1, p)$; generer N indikatorverdier

Eksempel

>rbinom(10,1,0.5)

[1] 0 1 0 0 1 0 0 0 1 1

Binomisk fordeling

Antall suksesser i n uavhengige forsøk der sannsynligheten for suksess er p.

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\mathrm{E}[X] = np \quad \mathrm{Var}[X] = np(1-p)$$

R-kommando: rbinom(N, n, p); generer N binomiske verdier

Eksempel

>rbinom(10,5,0.5)

[1] 4 3 3 4 1 2 2 2 2 2

Hypergeometrisk

Antall spesielle elementer når vi velger n tilfeldige elementer fra en populasjon på N objekter der M er spesielle.

$$p(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

$$\mathrm{E}[X] = n\frac{M}{N} \qquad \mathrm{Var}[X] = \frac{N-n}{N-1} \cdot n\frac{M}{N}\left(1 - \frac{M}{N}\right)$$

R-kommando: rhyper$(K, M, N-M, n)$; genererer K hypergeometriske verdier

Eksempel $(K = 10, M = 10, N = 15, n = 3)$

>rhyper$(10, 10, 5, 3)$

[1] 2 3 1 2 1 2 1 3 3 2

Poissonfordeling

Antall suksesser for en sjelden hendelse gjentatt mange ganger.

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

$$\mathrm{E}[X] = \lambda \quad \mathrm{Var}[X] = \lambda$$

R-kommando: rpois(K, λ); genererer K hypergeometriske verdier

Eksempel

>rpois$(10, 2)$

[1] 3 3 4 1 2 0 2 1 4 2

Uniform

Når alle utfall er like sannsynlige.

$$p(x) = \frac{1}{\text{Antall mulige utfall}}$$

Forventning og varians kan variere med sammensetningen av utfallene.

R-kommando: sample$(a : b, N, \text{replace} = T)$; generer N tall som er uniforme mellom a og b

Eksempel

```
>sample(20:50,10,replace=T)
[1] 49 32 24 42 49 29 50 22 50 23
```

Geometrisk: Antall fiaskoer før første suksess.

R-kommando: rgeom(N, p); generer N tall gir antall forsøk før første suksess, der p er sannsynligheten for suksess i hvert enkelt forsøk.

Eksempel

```
>rgeom(10,0.5)
[1] 0 1 0 3 4 1 2 1 0 0
```

Negativ binomisk: Antall fiaskoer før suksess nummer r. R-kommando: rnbinom(N, r, p); generer N tall gir antall forsøk før suksess nr r, der p er sannsynligheten for suksess i hvert enkelt forsøk.

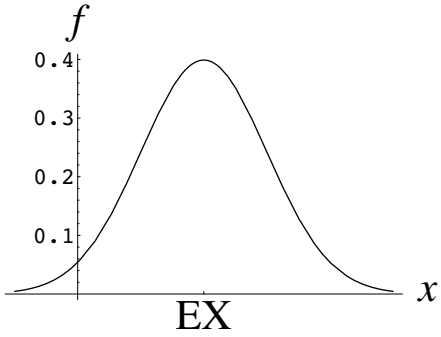
Eksempel

```
>rnbinom(10,4,2)
[1] 3 3 0 3 4 4 3 2 3 1
```

Klassiske kontinuerlige fordelinger

Normalfordeling

Symmetrisk med senter i forventningsverdien μ med standardavvik σ.



Standardmodellen for summen av mange uavhengige observasjoner.

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\mathrm{d}x$$

$$= \text{Areal under kurven til venstre for } x$$

$$\mathrm{E}[X] = \mu \quad \mathrm{Var}[X] = \sigma^2$$

Når $\mu = 0$ og $\sigma^2 = 1$, kalles den en standard normalfordeling, og verdiene finnes i tabell.

R-kommando: rnorm(N, μ, σ); generer N tall som er standard normalfordelte med forventning μ og varians σ^2

Eksempel

> rnorm(10,0,1)

[1] 1.09875708 -0.26633639 1.07915084 0.70834396 -1.01058764 -0.03791374

[7] 1.88705310 0.93547034 -0.96933994 -1.96797844

> rnorm(10,3,5)

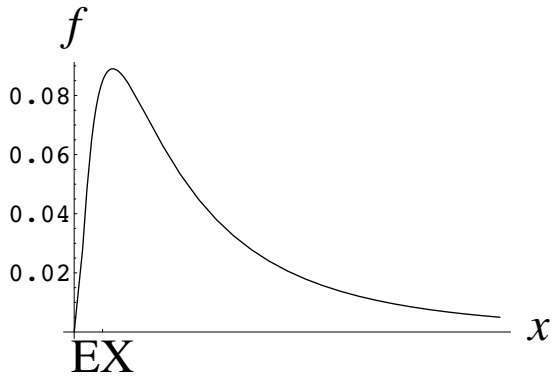
[1] -0.6896694 11.8922070 -12.8978063 1.1224679 11.7106385 -6.1709764

[7] -3.9625837 7.3821440 -2.6890942 5.5670479

I det siste eksempelet har vi generert 10 normalfordelte tall med $\sigma = 5$ og $\mu = 3$.

Lognormalfordeling

Lang høyrehale, normalt kun definert for positive tall. Standardmodell for aksjekurs.



R-kommando: rlnorm(N, r, s); generer N tall som er lognormalfordelte med

$$\mu = e^{r+s^2/2} \qquad \sigma^2 = \mu^2 \cdot (e^{s^2} - 1)$$

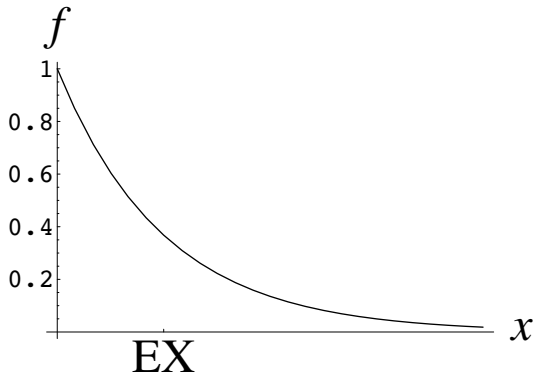
Eksempel

>rlnorm(10,0,1)

[1] 1.0908980 1.1214807 2.8838957 2.3878792 3.5962904 0.6990530 0.3493614 [8] 0.2847458 0.9810324 0.8029445

Eksponensialfordeling

Avtagende, lang høyrehale, kun definert for positive tall. Standardmodell for levetider.



R-kommando: rexp(N, λ); generer N tall som er eksponensialfordelte med parameter λ

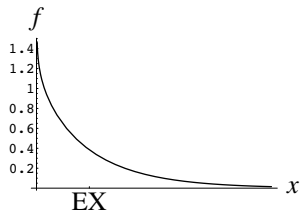
Eksempel

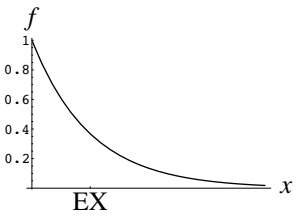
>rexp(10,0,1)

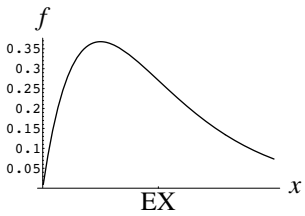
[1] 0.92939294 0.59548280 0.08927572 0.72086834 0.74884623 0.78800433
[7] 2.63746392 0.50356218 0.71913007 0.25707634

Gammafordeling

Lang høyrehale, kun definert for positive tall. Standardmodell for summer av levetider.

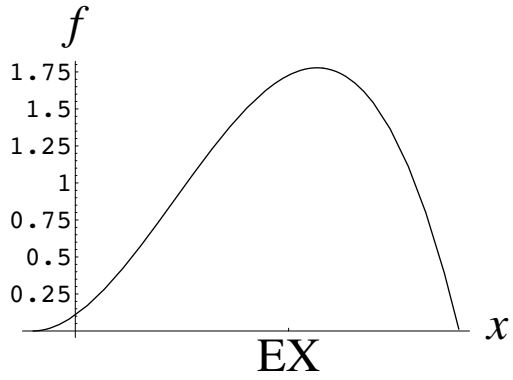






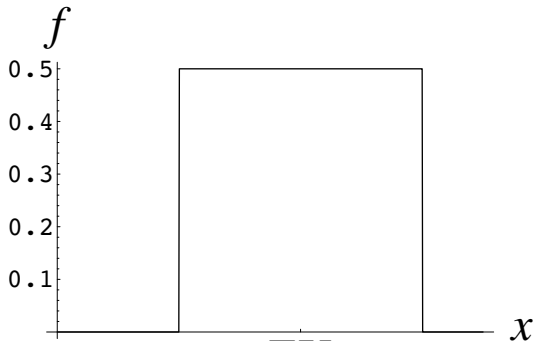
Betafordeling

Avgrenset av min og max. Kan være skjev i begge retninger.



Uniformfordeling/rektangulær

Avgrenset av min og max. Alle utfall er like sannsynlige.



EX

R-kommando: runif(N, a, b); genererer N tall som er uniforme mellom a og b

Eksempel

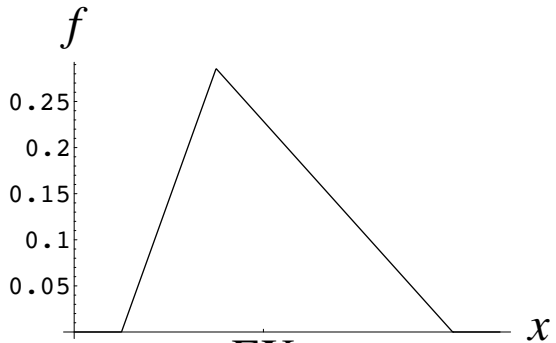
> runif(10,15,20)

[1] 19.99682 17.19472 15.68107 19.82319 19.31255 18.21970 17.20695 15.70275

[9] 17.51317 17.58581

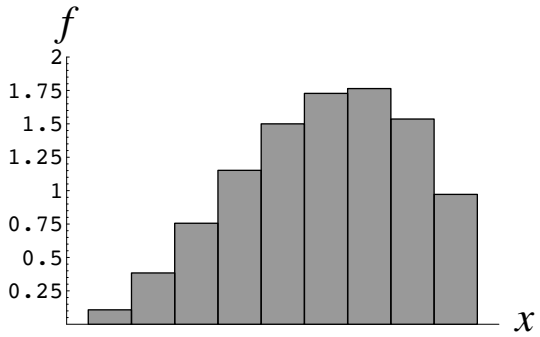
Triangelfordeling

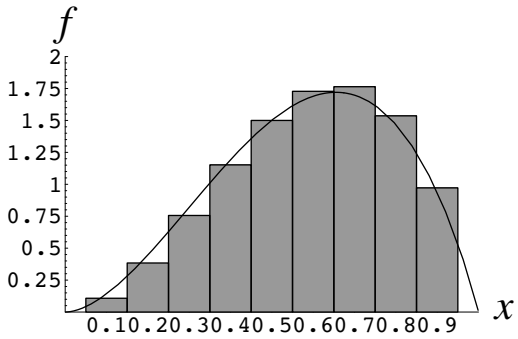
Avgrenset av min og max. Kan være skjev i begge retninger.



EX

Når en skal velge en modell, tar en utgangspunkt i endel observasjoner, og forsøker å finne en profiltype som passer. Deretter finner en verdier på parameterne slik at tilpasningen blir best mulig.





Beta-fordeling?

Sannsynligheter og forventning til kontinuerlige fordelinger

Vanligvis oppgir vi tettheten $f_X(x)$ til en kon-

tinuerlig variabel X. Da er

$$P(a \leq X \leq b) = \int_a^b f_X(x)\mathrm{d}x$$

Kumulativ fordeling

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(u)\mathrm{d}u$$

Merk at

$$f_X(x) = F'(x)$$

Generelt har vi

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\mathrm{d}x$$

$$\mathrm{Var}[X] = \int_{-\infty}^{\infty} (x - \mathrm{E}[X])^2 \cdot f_X(x)\mathrm{d}x$$

Forventning og varians for de klassiske fordelingene er imidlertid ferdig utregnet i tabeller. Det er derfor sjelden vi har behov for integrasjon!

Simulering

Eksempel

K : Utviklingskostnader

Q : Kvantum som kan selges

P : Pris som kan oppnås for varen

Alle disse kan generelt være tilfeldige variabler, og det er for eksempel rimelig å anta at fordelingen til Q er avhengig av verdien til P.

Problem: Hva er fordelingen til

$$R = P \cdot Q - K?$$

Selv om vi setter opp enkle regler for P, Q og K, kan fordelingen til R bli vanvittig komplisert!

Eksempel

K : Uniform på intervallet $[100000, 200000]$.

P : Uniform på intervallet $[100, 300]$.

Q : Uniform på intervallet

$$[3000 - 8P, 5000 - 11P]$$

Ellers uavhengige.

Slike elementer kan trekkes i R. Etter å ha trukket si 10000 tall av disse tre typene, kan en lage en ny kolonne med

$$R = P \cdot Q - K$$

og se hvordan den er fordelt.

Triks: For å lage Q, må vi først lage X uniform på $[0, 1]$.

$$Q = (3000 - 8P) \cdot (1 - X) + (5000 - 11P) \cdot X$$

Slik blir det gjort i R:

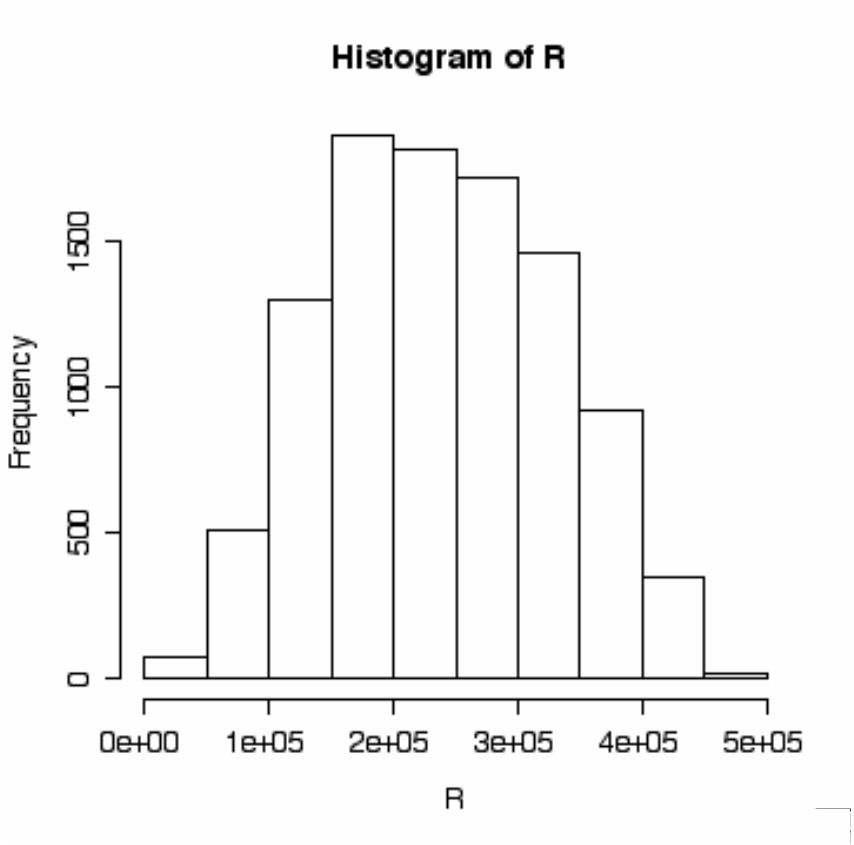
```
> K=runif(10000,100000,200000)
> P=runif(10000,100,300)
> X=runif(10000)
> Q=(3000-8*P)*(1-X)+(5000-11*P)*X
> R=P*Q-K
> mean(R)
```

[1] 237989.3

> sd(R)

[1] 90677.23

> hist(R)



Histogrammet viser fordelingen til R, og vi har beregnet

$$\mathrm{E}[R] = 237989.3 \qquad \mathrm{SD}[R] = 90677.23$$

Summer av tilfeldige variabler

Nøkkelord:

Samlet resultat etter n gjentagelser.

$$S_m = X_1 + X_2 + \cdots + X_m$$

NB! Noen ganger kan antall gjentagelser være en tilfeldig variabel M. Det kan for eksempel tenkes at vi stopper når vi har oppnådd et bestemt resultat. Antall ganger vi må prøve, kan variere fra gang til gang. NB!

$$S_M = X_1 + X_2 + \cdots + X_M$$

Oppførselen til slike variabler er vesentlig mere komplisert!

Sentralgrensesetningen

Hvis $X_1, \ldots, X_m$ alle er uavhengige med samme fordeling og n er stor, er S_m tilnærmet normalfordelt.

Endel resultater gjelder helt generelt:

Det er alltid slik at

$$\mathrm{E}[S_m] = \mathrm{E}[X_1] + \cdots + \mathrm{E}[X_m]$$

Hvis variablene er uavhengige, så er

$$\mathrm{Var}[S_m] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_m]$$

Hvis $X_1, \ldots, X_m$ er uavhengige med samme fordeling X, så gjelder:

- X indikator $p \Rightarrow S_m$ binomisk (n, p)
- X binomisk $(n, p) \Rightarrow S_m$ binomisk $(n \cdot m, p)$
- X Poisson $\lambda \Rightarrow S_m$ Poisson $m \cdot \lambda$
- X normal $N[\mu, \sigma^2] \Rightarrow S_m$ $N[m \cdot \mu, m \cdot \sigma^2]$
- X Eksponensial $\Rightarrow S_m$ Gammafordelt
- X Gammafordelt $\Rightarrow S_m$ Gammafordelt

I anvendelser er det viktig å vurdere om det er nødvendig eller rimelig å bruke sentralgrense-setningen.

Eksempel

To investeringsobjekter med avkastninger

R_1 med $\mathrm{E}[R_1] = 0,1, \mathrm{Var}[R_1] = 0,08^2$

R_2 med $\mathrm{E}[R_2] = 0,15, \mathrm{Var}[R_2] = 0,12^2$

Investerer 50% i hvert objekt, avkastningen blir da bestemt ved

$$R = \frac{1}{2}(R_1 + R_2)$$

Bruke sentralgrensesetningen på R???

Ikke samme fordeling, ikke uavhengige, ikke mange ledd, dårlig idé!!!

$$\begin{aligned}\mathrm{E}[R] &= \frac{1}{2}(\mathrm{E}[R_1] + \mathrm{E}[R_2]) \\ &= \frac{1}{2}(0,1 + 0,15) = 12,5\%\end{aligned}$$

$$\mathrm{Var}[R] = \frac{1}{2^2}\mathrm{Var}[R_1 + R_2]$$

Har *ikke* antatt uavhengighet!

$$\begin{aligned}\text{Var}[R_1 + R_2] &= \text{Var}[R_1] + \text{Var}[R_2] \\ &\quad + 2\text{Cov}[R_1, R_2]\end{aligned}$$

Hvis korrelasjonskoeffesienten $\rho[R_1, R_2]$ er kjent, kan vi finne $\text{Cov}[R_1, R_2]$.

$$\text{Cov}[R_1, R_2] = \rho[R_1, R_2] \cdot \sigma[R_1] \cdot \sigma[R_2]$$

Tilfelle I: $\rho = 0$

$$\begin{aligned}\text{Var}[R_1 + R_2] &= \text{Var}[R_1] + \text{Var}[R_2] \\ &= 0,08^2 + 0,12^2 = 0,0208\end{aligned}$$

$$\begin{aligned}\text{Var}[R] &= \frac{1}{2^2}\text{Var}[R_1 + R_2] \\ &= 0,0052 \approx 0,07^2\end{aligned}$$

Tilfelle II: $\rho = 0,5$

$$\text{Cov}[R_1, R_2] = 0,5 \cdot 0,08 \cdot 0,12 = 0,0048$$

$$\begin{aligned}\text{Var}[R_1 + R_2] &= 0,0208 \\ &\quad + 2 \cdot 0,0048 = 0,0304\end{aligned}$$

$$\text{Var}[R] = \frac{1}{2^2}\text{Var}[R_1 + R_2] = 0,0076 \approx 0.09^2$$

Tilfelle III: $\rho = -0,5$

$$\text{Cov}[R_1, R_2] = -0,5 \cdot 0,08 \cdot 0,12 = -0,0048$$

$$\begin{aligned}\text{Var}[R_1 + R_2] &= 0,0208 \\ &\quad - 2 \cdot 0,0048 = 0,0112\end{aligned}$$

$$\text{Var}[R] = \frac{1}{2^2}\text{Var}[R_1 + R_2] = 0,0028 \approx 0.05^2$$

Vi ser at risikoen blir redusert ved å spre investeringene. Dette gjelder generelt. De fleste typer av porteføljer kan behandles på denne måten.