Samling 1: Oppfriskning i statistikk for masterstudenter

Njål Foldnes

January 23, 2024

Universitetet i Stavanger

Outline i

Beskrivende statistikk for en variabel

Data består av observasjoner (rader) og variabler (søyler)

	price	lotsize	bedrooms	airco	garagepl	prefarea
1	42000	5850	3	no	1	no
2	38500	4000	2	no	0	no
3	49500	3060	3	no	0	no
4	60500	6650	3	no	0	no
5	61000	6360	2	no	0	no
6	66000	4160	3	yes	0	no
7	66000	3880	3	no	2	no

To hovedtyper av variabler

- Kategorisk variabel: Verdiene tilsvarer grupper eller kategorier. Land, Kjønn, osv.
- Kvantitativ variabel: Måler mengden eller antallet av noe. Lønn, Alder, Arbeidsledighetsrate.

Dette er viktig!

Typen analyse du kan utføre er avhengig av variabeltypen.

Målenivå

Et mer raffinert hierarki av målepresisjon

1. Nominal. Kategorier uten rekkefølge: F.eks. Nasjonalitet.
2. Ordinal. Kategorier med rekkefølge: F.eks. Karakterer A-F.
3. Intervall. Kvantitative, men forholdet mellom verdier har ingen mening: F.eks. Temperatur.
4. Ratio. Kvantitative, hvor forholdene mellom verdier gir mening: F.eks. Lønn.

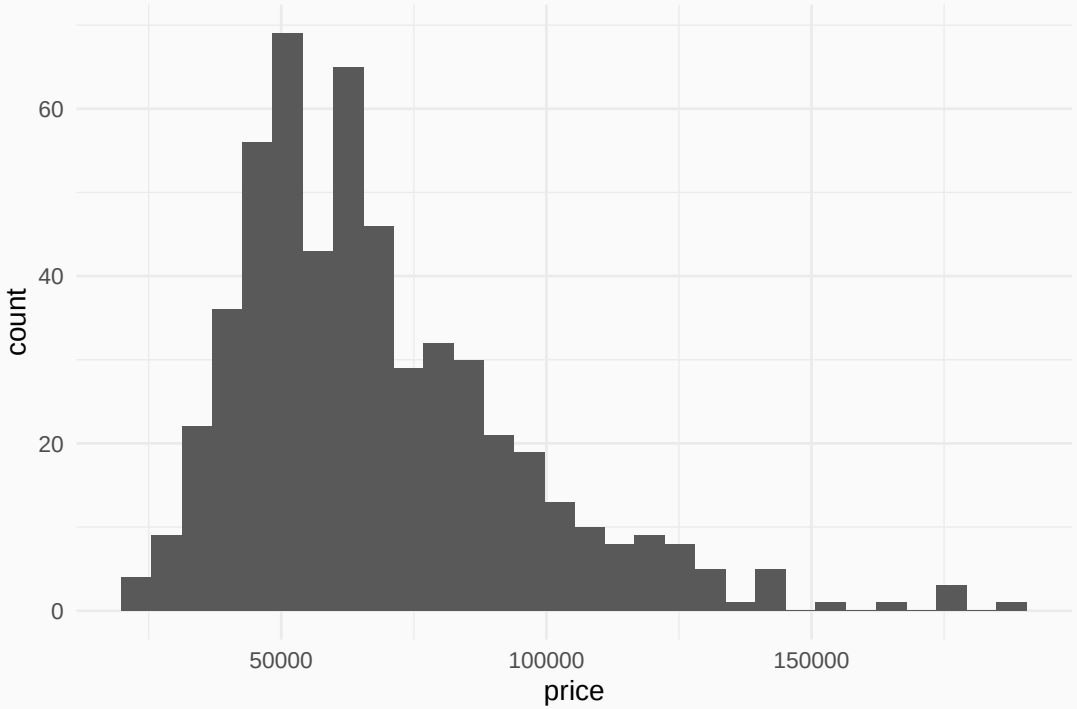
Undersøke fordelingen

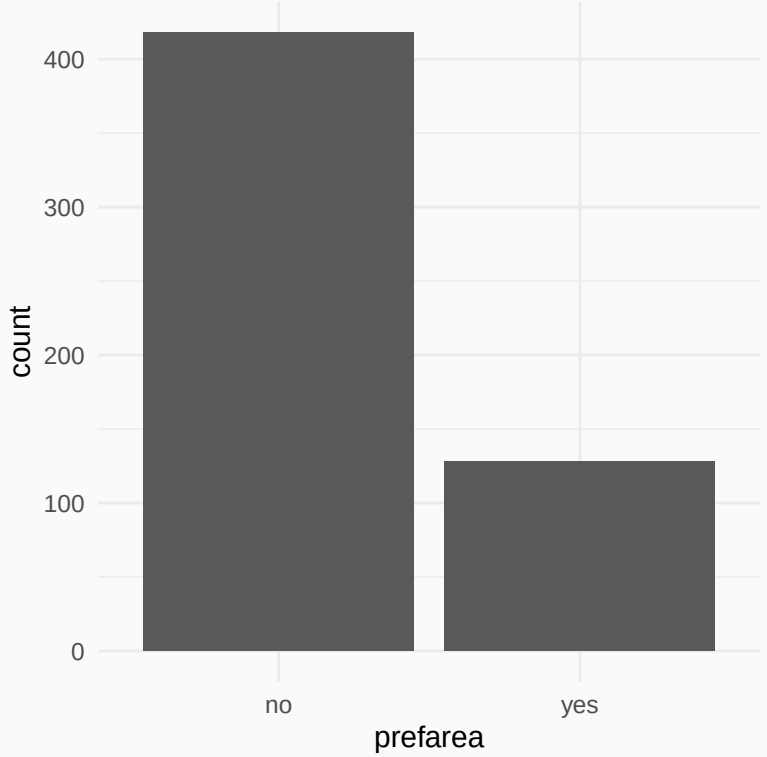
For å bli kjent med en variabel finnes det to tilnærminger:

- Grafer (Visuell informasjon)
- Beregning av nøkkeltall (Numerisk informasjon)

Undersøkelse av Pris og Foretrukket Område

Histogrammer brukes for kvantitative data, mens stolpediagrammer brukes for kategoriske data.





Beskrive fordelinger ved bruk av tall

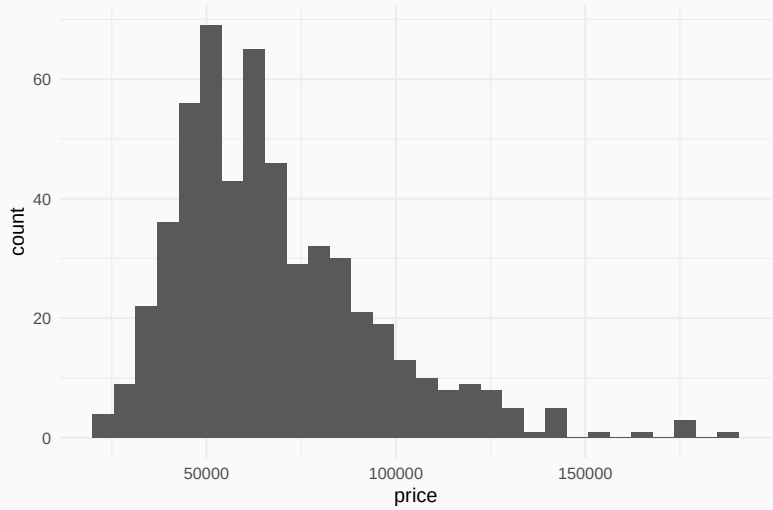
- For kategoriske data presenterer vi bare prosentandelen av hver verdi.
- For kvantitative data bruker vi oppsummeringsstatistikk som:
  - Gjennomsnitt eller Median for sentrum/lokasjonen av dataene
  - Standardavvik eller interkvartilrekkevidde (IQR) for variasjon/spredning i dataene
  - Gjennomsnittet og standardavviket er mer følsomme for *utliggere* sammenlignet med medianen og IQR
  - Uteliggere: Observasjoner som avviker betydelig fra flertallet av observasjonene

Undersøkelse av Pris med nøkkeltall

- Senter for prisene (typisk verdi) kan beregnes som gjennomsnittet $\overline{x}$ eller medianen $\tilde{x}$

$$\overline{x} = \frac{\sum x_i}{n} = 68121.6$$

$$\tilde{x} = 62000$$



- Variasjonsmål/spredning kan uttrykkes som

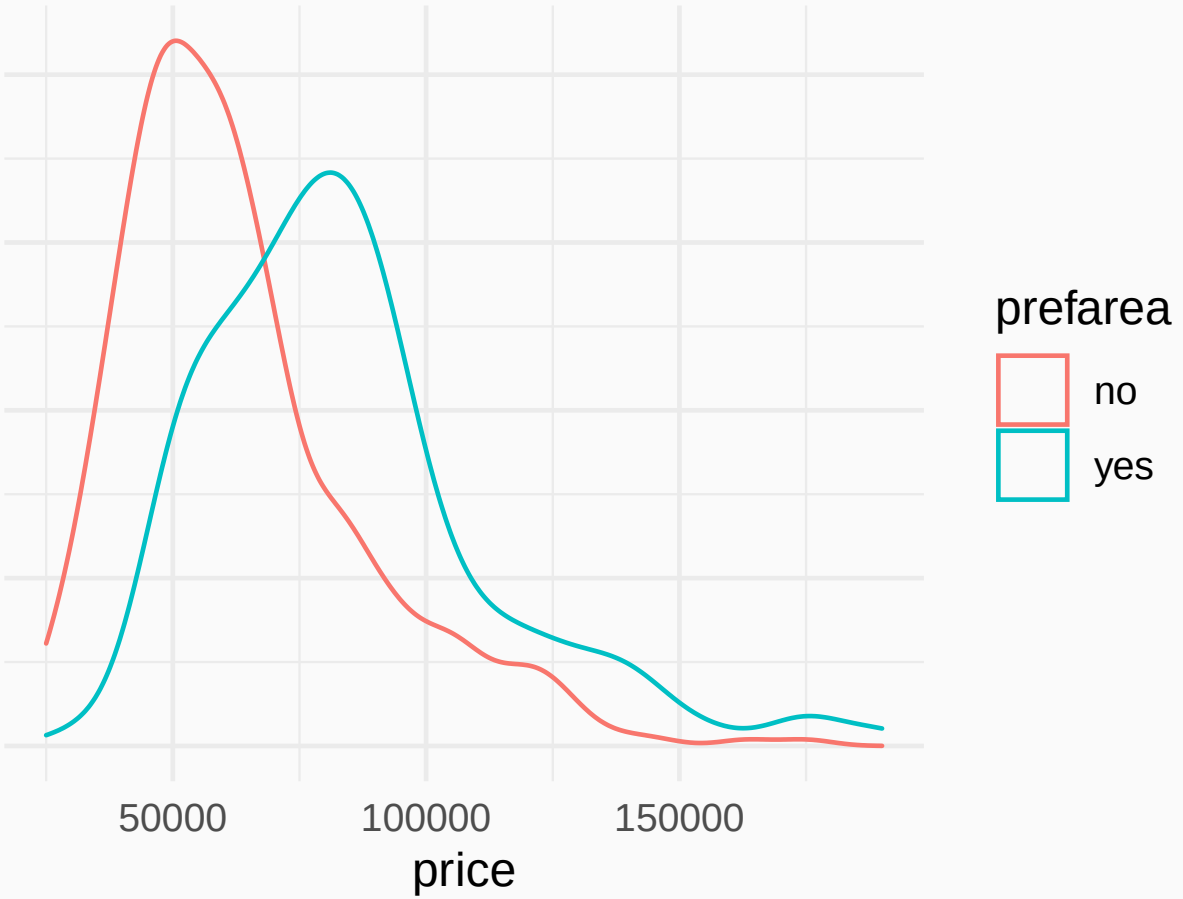
$$s = 26702.67$$

$$IQR = 32875$$

IQR er avstanden mellom første og tredje kvartil. Med andre ord, mellom 25%-percentilen og 75%-percentilen.

Standardavvik og IQR fanger opp spredning

prefarea	s	IQR
nei	24214	26375
ja	28342	30025



Et lite datasett

Example

I en gruppe er alderen til personene $16, 56, 58, 67, 62$ og 69 år. Beregn gjennomsnittet og medianen.

Løsning: $\overline{x} = \frac{328}{6} = 54.67$ og $\tilde{x} = 60$

For å beregne standardavviket må vi beregne avvikene:

x	$\overline{x}$	$x - \overline{x}$
16	54.67	-38.67
56	54.67	1.33
58	54.67	3.33
67	54.67	12.33
62	54.67	7.33
69	54.67	14.33

Variansen

Noen avvik er positive og noen er negative. Vi kvadrerer avvikene for å gjøre dem positive.

x	$x - \overline{x}$	$(x - \overline{x})^2$
16	-38.67	1495.11
56	1.33	1.78
58	3.33	11.11
67	12.33	152.11
62	7.33	53.78
69	14.33	205.44

Legg sammen kvadrerte avvik:
$1495.11 + 1.78 + 11.11 + 152.11 + 53.78 + 205.44 = 1919.33$ og del på $n - 1 = 5$ for å få *Variansen*

$$s^2 = \frac{1919.33}{6 - 1} = 383.87$$

Standardavviket er kvadratroten av variansen

Formelen for variansen er

$$s^2 = \frac{\sum(x - \overline{x})^2}{n - 1}$$

og vi får standardavviket ved å ta kvadratroten

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

Example

Standardavviket for verdiene $16, 56, 58, 67, 62$ og 69 er

$$s = \sqrt{383.87} = 19.59$$

Oppgave

Example

Fire studenter tar en test. De oppnår poengsummen $8, 9, 12$ og 16. Beregn standardavviket for testresultatene.

Løsning:

Poeng	$x - \overline{x}$	$(x - \overline{x})^2$
8	-3.25	10.56
9	-2.25	5.06
12	0.75	0.56
16	4.75	22.56

$$s^2 = \frac{10.56 + 5.06 + 0.56 + 22.56}{4 - 1} = 12.91$$

og $s = \sqrt{12.91} = 3.59$.

Standardavviket og normalfordelingen

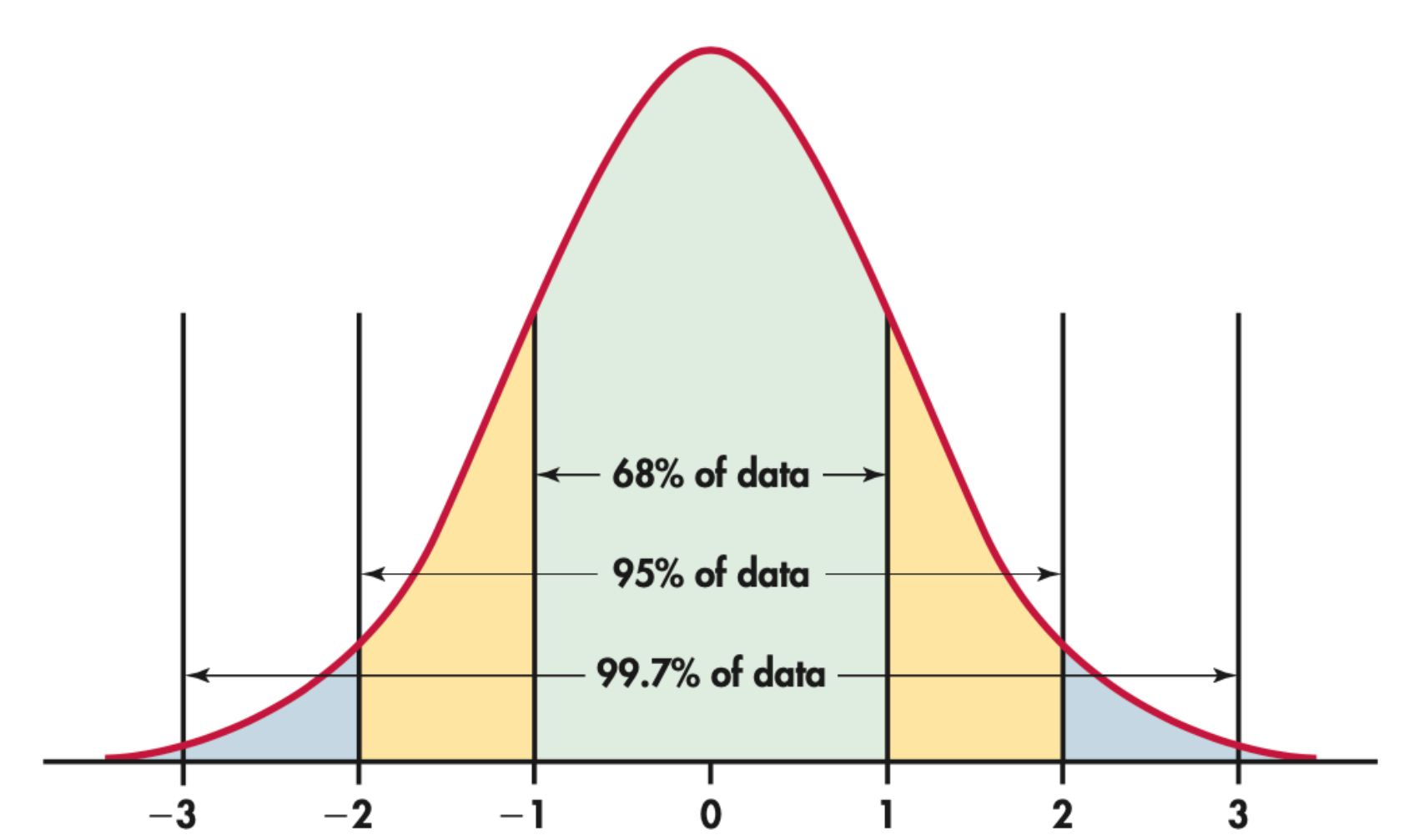
Normalfordelingen:

- Har en topp
- Er symmetrisk på begge sider av toppen
- Har en klokkeform

68-95-99.7 regelen

- 68% av observasjonene faller innenfor én standardavvik fra gjennomsnittet
- 95% av observasjonene faller innenfor to standardavvik fra gjennomsnittet
- 99.7% av observasjonene faller innenfor tre standardavvik fra gjennomsnittet

68-95-99.7 rule



Standardisering

En observasjon kan standardiseres til en z-score. z-scoren måler hvor mange standardavvik observasjonen er fra gjennomsnittsverdien.

Example

400 studenter fullfører SAT-testen. Gjennomsnittlig resultat er $\overline{x} = 504$ poeng, og standardavviket er $s = 97$ poeng.

Tore og Fanny fikk henholdsvis 410 og 730 poeng. Hva er deres z-scores?

$$z_T = \frac{410 - 504}{97} = -0.97, \qquad z_F = \frac{730 - 504}{97} = 2.33$$

Fanny er *uvanlig*, ikke Tore. z-verdier større enn 2 eller mindre enn -2 betraktes ofte som uvanlige!

Øvelse

Example

En støvsugermodell har en gjennomsnittlig levetid på $\mu = 800$ dager med et standardavvik på $\sigma = 100$ dager. En støvsuger av denne modellen varte i 1050 dager. Hva er z-verdien for denne levetiden? Varte støvsugeren uvanlig lenge?

Løsning

$$z = \frac{1050 - 800}{100} = 2.5$$

Uvanlig lenge, siden $z > 2$

Deskriptiv statistikk for to variabler

Beskrivelse av sammenhengen mellom to kategoriske variabler

Krysstabell

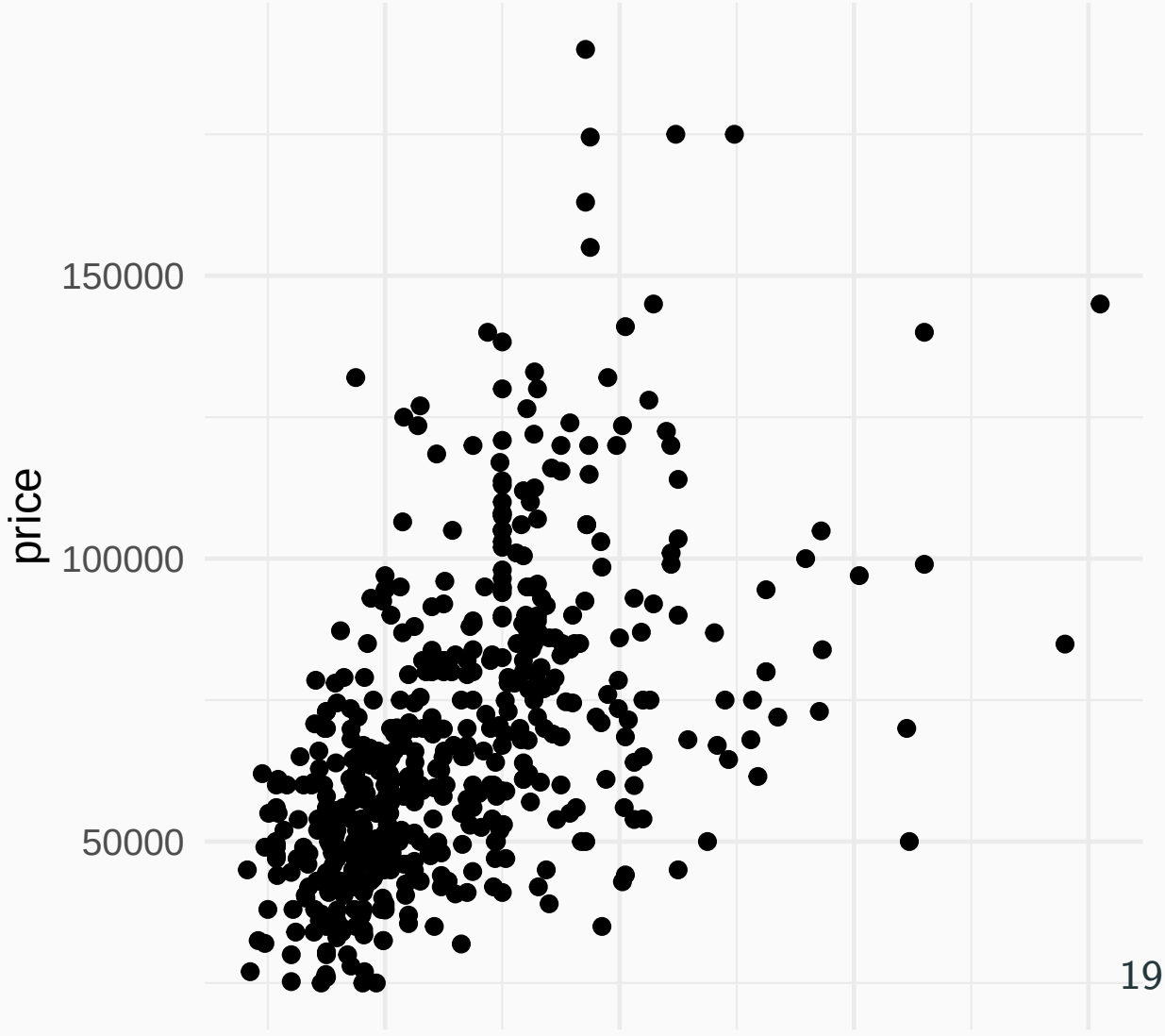
Med to kategoriske variabler kan vi oppsummere med en tabell som viser frekvenser

	Foretrukket Område	
	nei	ja
ikke-aircondition	298	75
aircondition	120	53

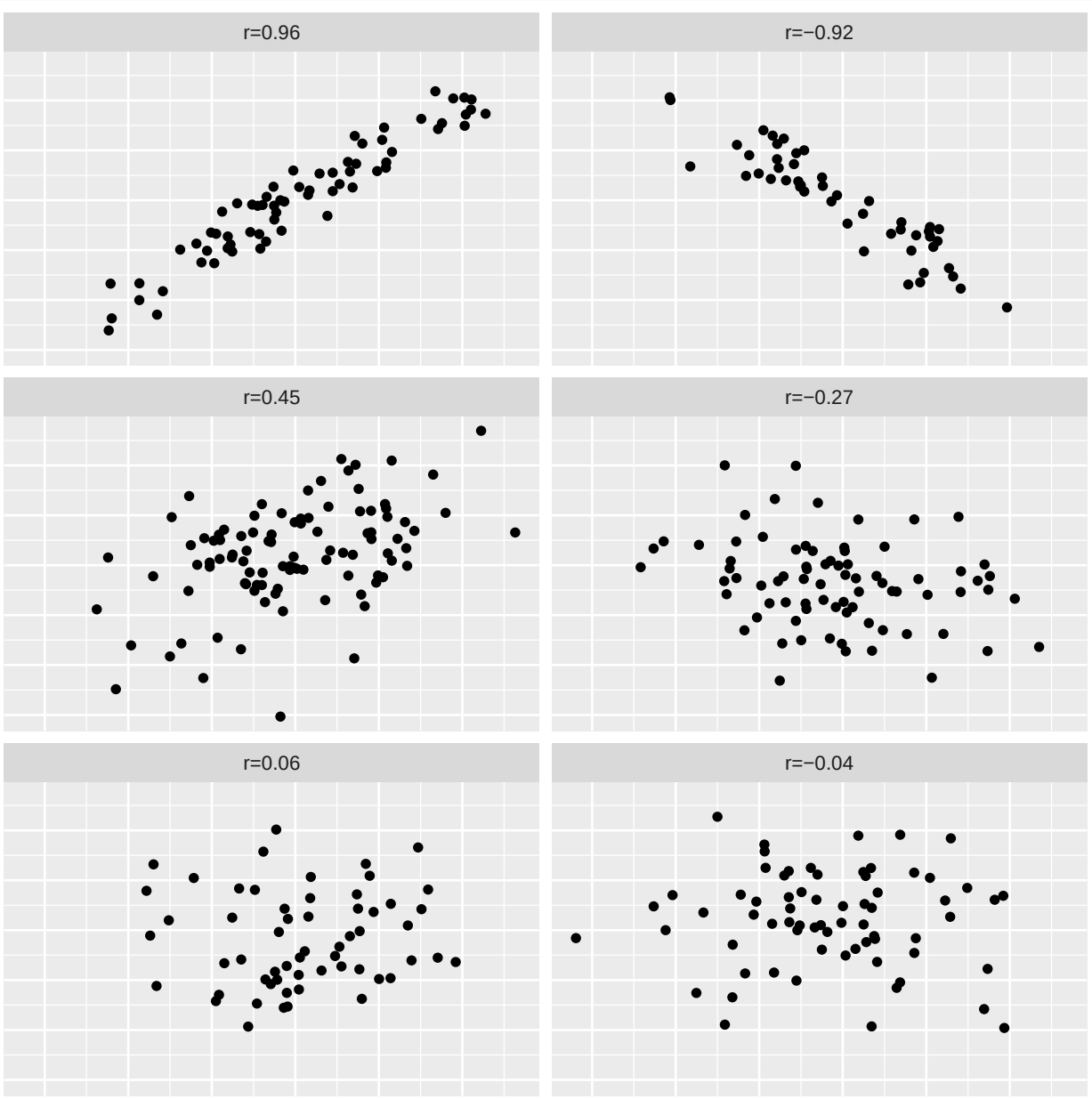
Beskrivelse av sammenhengen mellom to kvantitative variabler

- Visuell: Spredningsdiagram
- Numerisk: Korrelasjon

Korrelasjonen mellom tomtestørrelse og pris er $r = +.55$



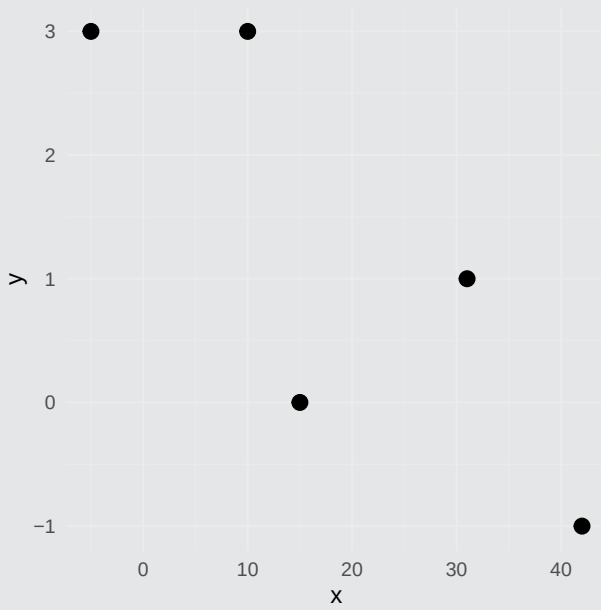
Korrelasjon kan være positiv eller negativ



Formel for korrelasjon

$$r = \frac{\sum(x - \bar{x}) \cdot (y - \bar{y})}{(n-1)s_x s_y}$$

Example



Disse to har en negative assosiasjon

Example

x	-5	10	15	31	42
y	3	3	0	1	-1

1. $\bar{x} = 18.6,\ s_x = 18.339,\ \bar{y} = 1.2,\ s_y = 1.789$
2.

$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x}) \cdot (y - \bar{y})$
-5-18.6=-23.60	3-1.2=1.80	-23.60 · 1.80= -42.48
-8.60	1.80	-15.48
-3.60	-1.20	4.32
12.40	-0.20	-2.48
23.40	-2.20	-51.48

3. $\sum(x - \bar{x}) \cdot (y - \bar{y}) =$
   $-42.48 - 15.48 + 4.32 - 2.48 - 51.48 = -107.6$
4. $r = \frac{-107.6}{(5-1)\cdot 18.339 \cdot 1.789} = -0.82$

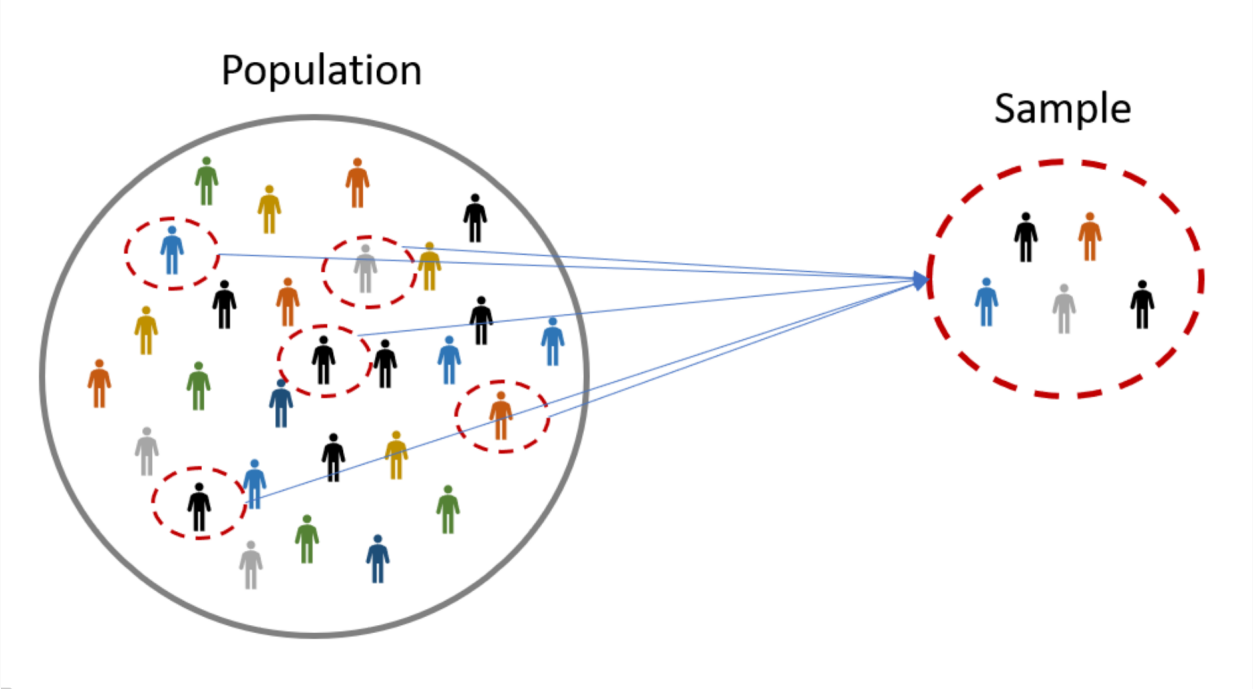
Viktige fakta om korrelasjon

- Måler bare styrken i en *lineær* sammenheng
- Positiv sammenheng gir en positiv r. Negativ sammenheng gir en negativ r
- Hvis punktene ligger nær en linje, vil r være nær $+1$ eller -1.
- Ingen sammenheng betyr $r \approx 0$
- Skala-uavhengig, slik at r ikke vil endres hvis vi multipliserer alle verdiene i en vektor med en konstant

Inferens for en variabel

Inferens

Utvalget er et *utvalg* fra en mye større *populasjon*



Inferens er å generalisere fra utvalget til populasjonen. Estimering og hypotesetesting er sentrale inferenstemaer.

Estimering

Populasjonsparametere

I populasjonen er det parametere som vi er interessert i å lære mer om.

- μ: Gjennomsnittlig lønn for ansatte i Norge?
- p: Andelen studenter i Norge som noen ganger føler seg ensomme?

Estimering

Vi vet ikke disse parameterne, men vi kan forsøke å estimere dem

- I et stort tilfeldig utvalg av norske ansatte var gjennomsnittslønnen $\overline{x} = 452300$ NOK
- I en stor klasse på UiS rapporterte $\hat{p} = 22\%$ av studentene at de føler seg ensomme noen ganger.

Tilfeldig utvalg er nøkkelen til gyldig inferens

Hvilke estimater kan vi stole på?

Gjennomsnittslønnen og andelen deprimerte på forrige side er estimater for populasjonsparametrene μ og p, henholdsvis. Estimatet for μ er mer pålitelig enn estimatet for p. Hvorfor?

Fordi et tilfeldig utvalg bedre vil representere populasjonen enn en enkelt klasse ved en enkelt institusjon. (det siste utvalget vil sannsynligvis produsere *skjeve* resultater)

Tilfeldig utvalg

Når utvalget produseres ved å tilfeldig velge observasjoner, vil det tendere til å representere sin populasjon, spesielt hvis utvalgsstørrelsen n er stor. (loven om store tall) Alle inferensprosedyrer krever et tilfeldig utvalg, og de vil fungere bedre jo større utvalgsstørrelsen er.

konfidensintervaller for gjennomsnittet

Generell form for konfidensintervall

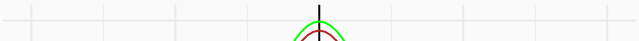
$$\text{Estimat} \pm \text{Feilmargin}$$

Vanligvis er tillitsnivået 95%. Vi sier da at vi er 95% sikre på at intervallet inneholder parameteren.

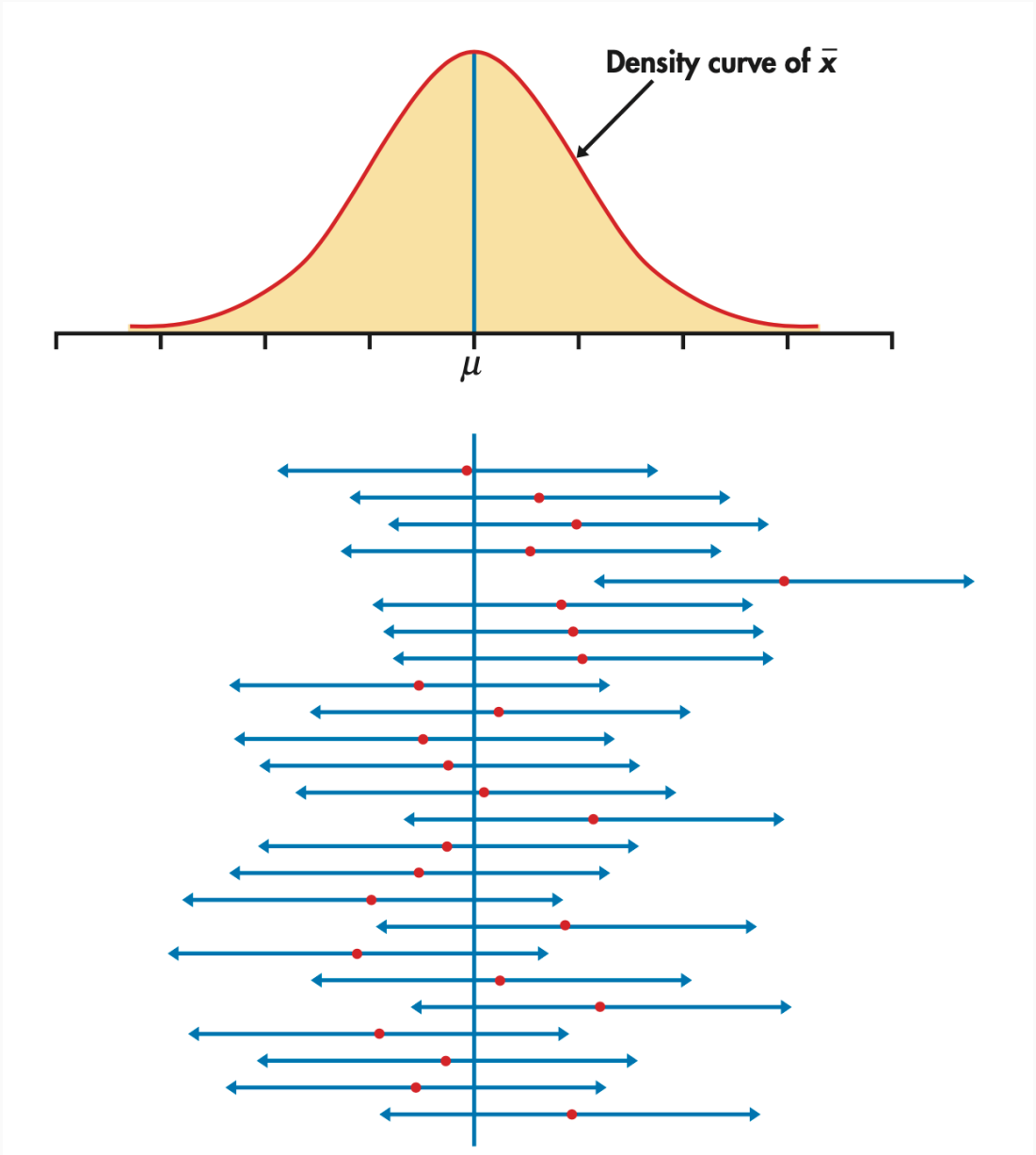
95% konfidensintervall for gjennomsnittet μ

$$\overline{x} \pm t^* \frac{s}{\sqrt{n}}$$

- $\frac{s}{\sqrt{n}}$ er *standardfeilen* til utvalgsgjennomsnittet $\overline{x}$, dvs. standardavviket til $\overline{x}$
- Tallet t^* vil omtrent være lik 2, men avhenger av utvalgsstørrelsen og bør slås opp i en tabell/kalkulator



95% konfidensintervall bommer 1 av 20 ganger (5%)



t-fordeling ser ut som en normalfordeling, men med tykkere haler

Det er en t-fordeling for hver frihetsgrad $df = 1, 2, \ldots$. Når $df > 30$ er den nesten ikke til å skille fra en standard normalfordeling.

Når er t-metoden gyldig?

- Hvis $n \geq 30$: t-metoder er gyldige
- Hvis $n < 30$: t-metoder er gyldige hvis dataene er omtrent normalfordelte. Hvis dataene er skjeve, er ikke t-metoder egnet

Hypotesetesting

Regelen om sjeldne hendelser

- Anta at vi har en nullhypotese om verdien til en populasjonsparameter.
- I utvalget vårt er estimert verdi ikke i nærheten av H_0-verdien.
- Konklusjon: Det må være noe galt med H_0. Vi forkaster den!

Hypotesetesting

Hypotesetest = Beslutningsregel

- Det er to konkurrerende hypoteser, H_0 og den alternative H_A
  - Nullhypotesen H_0 inneholder $=$
  - Den alternative hypotesen H_A inneholder $<$, $>$ eller $\neq$
- Du må velge enten H_0 eller H_A. Men utgangspunktet er at H_0 er sann
- Beslutningen tas basert på utvalget
- Hvis utvalget er veldig annerledes enn det vi forventer hvis H_0 er sann, forkaster vi H_0. Hvis ikke, beholder vi H_0
- Bevisbyrden ligger dermed på H_A.
- Hvis du vil finne støtte for H_A, er det best å anta at H_0 er sann, og deretter vise at noe veldig usannsynlig har skjedd.

Hypotesetesting Feil

To måter ting kan gå galt på

- Type I-feil: Forkaste H_0 når den er sann
- Type II-feil: Ikke forkaste H_0 når den er usann

Viktige begreper

Teststatistikk

For å beregne avviket mellom H_0 og dataene beregner vi en standardisert teststatistikk

$$t = \frac{\text{utvalgs-estimat} - \text{hypotetisk verdi}}{\text{standardavviket til estimatet}}$$

- Hvis H_0 er sann, bør teststatistikken ikke være stor
- Hvis teststatistikken er for langt fra null, forkaster vi H_0

p-verdien

Sannsynligheten (under forutsetning av at H_0 er sann) for at teststatistikken skal ta en verdi som er like ekstrem eller mer ekstrem enn den som faktisk ble observert. Jo mindre p-verdien er, jo sterkere er bevisene mot H_0.

Feil konklusjon på hypotesetesting

To måter ting kan gå galt på

- Type I-feil: Forkaste H_0 når den er sann
- Type II-feil: Ikke forkaste H_0 når den er usann

Signifikansnivå

Signifikansnivå

- Signifikansnivået α bestemmer hvor sterke bevisene mot H_0 må være for å forkaste den
- Hvis p-verdien er under .05, forkaster vi H_0
- Dette kalles statistisk signifikans

Hypotesetesting for gjennomsnittet μ

$H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$

Teststatistikk

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Hvis H_0 er sann, er denne statistikken t-fordelt med $n - 1$ frihetsgrader.

Example

Anta at vekten på en eske med cornflakes er normalfordelt. Vi har et tilfeldig utvalg på syv esker. Eskenes vekt skal være 500 g. Vektene er $502, 498, 479, 492, 488, 494, 494$

Vi ønsker å teste om gjennomsnittsvekten av cornflakes-esker faktisk er 500 g. Vi setter signifikansnivået til $\alpha = 0.05$.

a. Formuler H_0 og H_A.
b. Beregn teststatistikken t.
c. Bruk en tabell for å finne kritisk verdi, eller beregn p-verdien på en kalkulator.
d. Skal vi forkaste H_0?

Løsning

a. $H_0 : \mu = 500$ vs $H_A : \mu \neq 500$

b. $t = (492.429 - 500)/(7.390/\sqrt{7}) = -2.711$

c. $t^* = 2.447, \quad p = 2 \cdot 0.0175 = 0.035$

d. Ja, vi forkaster H_0.