



**Høgskolen i Molde**

Vitenskapelig høyskole i logistikk

## 6. Statistisk inferens

### Øving 6: Oppgaver

*“MAT110 Statistikk I”, 2021*

versjon 01

- Utlevering:
  - mandag 29. mars kl. 12:00
  - skriv ut oppgavene på papir, i farger, og ta med på øvingsdagen.
- Innlevering:
  - mandag 26. april kl. 12:00 mandag
  - dvs. man har 4 uker på å gjøre oppgavene
  - vi anbefaler at man prøver å bli ferdig etter 2 uker, siden ny øving kommer da
- **FORarbeid:**
  - Se videoene for uke 13, 14, 15 og 16
  - Les kompendiet.
  - Les gjennom øvingoppgave 6 så dere vet problemstillingene
- **ETTERarbeid:**
  - Gå gjennom løsning 6 som legges ut torsdag 8. april kl. 16:00.

**Problem 6.1 — statistisk inferens - McDonald's**

McDonald's i Molde ønsker å gjøre en analyse angående bemanning og jobbfordeling i rushperioden mellom 16:00 - 18:00 i ukedagene. De har oppdaget de smertefulle konsekvensene av å ha for få eller for mange ansatte i denne perioden:

- For få ansatte:  
gir lange køer med misfornøyde kunder som derfor heller drar til Burger King
- For mange ansatte:  
gir høye driftskostnader

For å kunne dimensjonere og fordele de ansatte i denne perioden, har de innsett at det er **to typer data** som i størst grad påvirker dimensjoneringen av bemanningen. <sup>1</sup>

- **IAT** - *Inter Arrival Time*:  
Tiden mellom to påfølgende kunder.  
"Kunde" betyr her en *gruppe* personer som kommer til McDonald's som en enhet. <sup>2</sup>
- **Størrelsen på ordrene** fra hele gruppen målt i bearbeidingstiden ordren krever fra de ansatte. <sup>3</sup>



Figure 6.1: McDonald's.

<sup>1</sup>"Dimensjonering" betyr her hvor *mange* ansatte de har.

<sup>2</sup>En familie er et typisk eksempel på en slik gruppe. McDonald's anser hele familien som "kunden". En vennegjeng er et annet eksempel på en slik gruppe.

<sup>3</sup>Dette er selve tiden det tar å lage ordren, dvs. eventuelle ventetider på grunn av køer ikke tatt med i disse tidene. Disse tidene er mao. kun beregnet fra selve ordren, ikke fra hvordan trykket var akkurat da ordren ble bestilt.

Før McDonald's starter selve analysen rundt dimensjoneringen av bemanningen, ønsker de å skaffe seg et godt statistisk utgangspunkt. De følger oppsettet:

1. Definere **populasjonen**.
2. Skaffe til veie **data** over en gitt periode.<sup>4</sup>
3. Beregne **nøkkeltall**, dvs. beskrivende statistikk for de to datasettene.
4. Formulere statistiske **modeller** for utvalgene og populasjonen.

a) Hva er populasjonsmengden i tilfellet hos McDonald's?

Du skal hjelpe McDonald's å samle inn forsøksdata. Dataene forekommer som *par*

$$z_i = (x_i, y_i) \quad (6.1)$$

hvor  $i$  står for kunde nr.  $i$ , hvor  $i = 1, 2, 3, \dots, n$ .

For hver ankommet kunde  $i$ , måles altså tallene:

$$x_i = \text{tiden mellom kunde nr. } i \text{ og forrige kunde } i - 1 \text{ (IAT) (antall minutter)} \quad (6.2)$$

$$y_i = \text{bearbeidingstiden for ordren til kunde nr. } i \text{ (antall minutter)} \quad (6.3)$$

b) Beskriv i korte trekk hvordan du vil gjennomføre datainnsamlingen slik at forsøksdataene  $z_1, z_2, \dots, z_n$  skal utgjøre et tilfeldig utvalg fra populasjonen.



Figure 6.2: Datainnsamling.

<sup>4</sup>Målet er at dataene utgjør et tilfeldig utvalg.

I tabell 6.1 har McDonald's samlet inn data fra kundene mellom 16:00 - 18:00 i hverdagene og målt  $x_i$  (IAT) og  $y_i$  (bearbeidingstiden) for  $n = 100$  kunder.

Antall minutter:

(0.85 , 4.94)	(0.30 , 5.50)	(0.21 , 5.78)	(0.40 , 3.84)	(0.17 , 7.62)
(0.31 , 6.61)	(0.08 , 7.83)	(0.69 , 4.45)	(0.29 , 8.50)	(0.02 , 8.41)
(0.61 , 7.04)	(1.06 , 6.27)	(0.36 , 9.73)	(0.58 , 3.04)	(0.09 , 2.52)
(0.45 , 4.98)	(0.33 , 5.70)	(0.06 , 6.83)	(0.25 , 3.41)	(0.22 , 7.32)
(0.52 , 6.84)	(0.30 , 8.74)	(0.64 , 10.59)	(1.07 , 5.23)	(1.38 , 6.50)
(0.06 , 7.88)	(0.07 , 9.00)	(0.31 , 8.44)	(1.09 , 11.86)	(0.28 , 9.32)
(0.09 , 4.40)	(1.65 , 8.07)	(0.03 , 10.09)	(0.29 , 7.97)	(0.27 , 4.88)
(0.41 , 3.22)	(0.25 , 7.29)	(0.21 , 5.85)	(0.10 , 4.89)	(0.44 , 6.42)
(0.09 , 6.41)	(0.06 , 5.88)	(0.78 , 4.01)	(1.60 , 6.97)	(0.08 , 6.48)
(0.09 , 4.99)	(0.20 , 5.30)	(0.40 , 4.58)	(0.39 , 10.04)	(0.75 , 5.27)
(0.57 , 4.92)	(1.22 , 5.07)	(0.50 , 6.36)	(0.59 , 5.88)	(0.84 , 7.53)
(0.09 , 6.25)	(0.04 , 10.54)	(0.51 , 7.41)	(0.98 , 8.12)	(0.78 , 6.65)
(0.28 , 8.40)	(0.33 , 7.38)	(0.08 , 6.13)	(0.07 , 6.72)	(0.30 , 5.07)
(0.10 , 6.97)	(0.29 , 3.36)	(0.16 , 5.22)	(1.60 , 8.56)	(0.37 , 5.76)
(0.48 , 5.29)	(0.25 , 7.45)	(0.29 , 2.90)	(1.59 , 6.44)	(0.13 , 1.72)
(0.31 , 4.87)	(0.07 , 4.36)	(1.29 , 4.29)	(0.47 , 6.70)	(0.12 , 10.40)
(0.24 , 5.27)	(1.07 , 8.33)	(0.05 , 9.12)	(0.19 , 4.32)	(0.17 , 5.03)
(0.00 , 4.78)	(1.03 , 7.81)	(0.71 , 3.75)	(0.12 , 5.33)	(0.28 , 5.36)
(0.03 , 6.26)	(1.28 , 7.42)	(1.18 , 2.28)	(0.20 , 5.29)	(0.01 , 3.44)
(0.22 , 5.48)	(0.61 , 6.61)	(0.02 , 6.77)	(0.49 , 7.29)	(0.12 , 5.62)

Table 6.1: Dataene  $(x_i, y_i)$ , hvor  $i = 1, 2, 3, \dots, 100$ .

- c) Lag stolpediagrammer av relativfrekvensene  $f_r$  for  $x_i$  (IAT) og  $y_i$  (bearbeidingstiden).<sup>5</sup> Del opp verdiområdet til  $x_i$  og  $y_i$  i følgende intervaller:<sup>6</sup>

- $x_i$  - start fra 0 og lag intervaller med lengde 0.2 frem til 3:

$[0.0, 0.2)$

$[0.2, 0.4)$

$\vdots$

$[2.8, 3.0)$

- $y_i$  - start fra 0 og lag intervaller med lengde 1 frem til 12:

$[0, 1)$

$[1, 2)$

$\vdots$

$[12, 13)$

---

<sup>5</sup>Altså ett stolpediagram for  $x_i$  og ett stolpediagram for  $y_i$ .

<sup>6</sup>Beregn relativ frekvensene  $f_r$  for hvert intervall og plott dem i et stolpediagram. Se kapittel 6 i boksen. Vi regnet der ut relativfrekvenser  $f_r$  for eksemplet med legemiddel.

- d) Beregn beskrivende nøkkeltall for både IAT og bearbeidingstiden. Inkluder følgende størrelser: <sup>7</sup>

- min
- maks
- variasjonsbredde
- median
- gjennomsnitt
- typetall
- empirisk varians
- empirisk standardavvik
- 1. kvartil (25%)
- 3. kvartil (75%)
- kvartilavvik

---

<sup>7</sup>Bruk gjerne Excel for å regne ut disse nøkkeltallene. En Excel-fil med tallene fra tabell 6.1 ligger på Canvas. Da blir det ikke så mye arbeid. Oppgi tallene med 3 desimalers nøyaktighet.

Definer populasjonsvariablene:

$X$  = tiden mellom to påfølgende kunder i tidsrommet 16:00-18:00 på en gitt hverdag  
(6.4)

$Y$  = bearbeidingstiden til en kunde i i tidsrommet 16:00-18:00 på en gitt hverdag  
(6.5)

Definer tilhørende forsøksvariabler:

$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$  (6.6)

e) Er

$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X$  (6.7)

$Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} Y$  (6.8)

rimelig å anta?  
Gi en kort begrunnelse.

f) Er det rimelig å anta at  $X_i$ -ene er gjensidig uavhengige av  $Y_i$ -ene?  
Gi en kort begrunnelse.

g) På bakgrunn av kunnskapen vi har om McDonald's-eksemplet samt de beskrivende nøkkeltallene for observasjonene, formuler to statistiske modeller:

- (a) Statistisk modell for IAT - Inter-Arrival Times ( $X_i$ -ene)
- (b) Statistisk modell for bearbeidingstiden ( $Y_i$ -ene)

som du mener representerer situasjonen godt. <sup>8</sup>

■

---

<sup>8</sup>Se på stolpediagrammene og se om du gjenkjenner noen kjente sannsynlighetsfordelinger. Søk også på internett for å se hva andre har gjort for IAT - Inter Arrival Times.







**Høgskolen i Molde**  
Vitenskapelig høyskole i logistikk

## 11. Regresjonsanalyse

### Problem 11.1 — regresjon

- a) Hva er regresjonsanalyse? <sup>1</sup>
- b) Hva er formålet med regresjonsanalyse? <sup>2</sup>
- c) Hvilke to typer regresjon skiller man ofte mellom?

■

---

<sup>1</sup> Her trengs kun et *kort* svar. Bruk internett, boken eller noe annet du synes er egnet å slå opp i dersom du ikke har det i hodet.

<sup>2</sup> Se boken eller andre se steder. I dette kurset ser vi kun på tilfellet når vi har én uavhengig variabel.

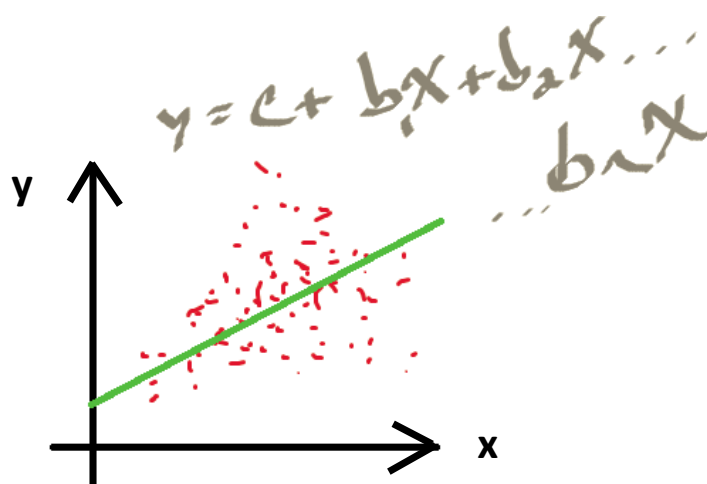


Figure 11.1: Regresjon.

### Problem 11.2 — økonomi - regresjonsanalyse

EiendomsMegler 1 Midt-Norge ønsker å se nærmere på sammenhengen mellom areal  $x$  og pris  $y$  på leiligheter. De gjør observasjoner av  $x$  og tilhørende  $y$ :

$$x = \text{areal} \quad (11.1)$$

$$y = \text{pris (i 1000 NOK)} \quad (11.2)$$

Eiendomsfirmaet solgte  $n = 6$  leiligheter i mars 2018. Areal og pris for disse leilighetene er oppsummert i følgende tabell:

$x$ ( areal i m <sup>2</sup> )	43	60	75	80	95	105
$y$ pris ( i 1000 NOK )	2100	2850	3050	3800	4525	4500

Figure 11.2: Areal  $x$  og pris  $y$ .

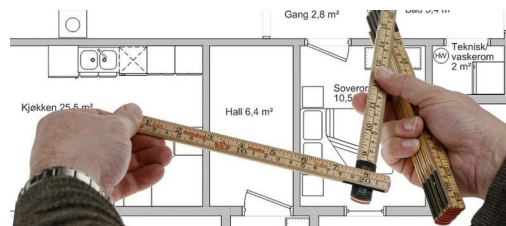


Figure 11.3: Areal og pris.

PS:

Man kan fint løse denne oppgaven kun ved bruk av vanlig kalkulator. Uten bruk av Excel. Men denne oppgaven er et eksempel på at et dataprogram som f.eks. Excel kan brukes i statistikksammenheng. Derfor er det laget en Excel-fil:

*006 Øving 6, oppgave 11.2, pris-areal, (29. mars 2021).xlsx*

Denne Excel-filen ligger på Canvas. Direktelink finner du her. Selv om det er frivillig å bruke Excel så kan det være en fin måte å sjekke dine svar på i denne oppgaven. I tillegg så illustrerer det, til en viss grad, nytten av dataprogrammer innen statistikk.

- a) Hva er gjennomsnittlig areal  $\bar{x}$  av leilighetene? Og gjennomsnittspris  $\bar{y}$ ?
- b) Hva er den empiriske variansen for arealet  $x$ , dvs. hva er  $S_x^2$ ? Og for prisen,  $S_y^2$ ?
- c) Hva er den empiriske kovariansen mellom  $x$  og  $y$ , dvs. hva er  $S_{xy}$ ?  
samvariasjon
- d) Finn minste kvadraters **regresjonslinje** for  $x$  og  $y$ .<sup>3</sup>
- e) Den største leiligheten i tabellen i figur 11.2 er “bare” 105 m<sup>2</sup>. Dersom eiendomsmeglerfirmaet ønsker å *estimere* hvor mye en leilighet på f.eks. 140 m<sup>2</sup> vil koste basert på slagstallene fra mars så kan de bruke den estimerte modellen fra oppgave 11.2 d), altså regresjonslinjen.
- Hvor mye predikerer regresjonslinjen at en leilighet på 140 m<sup>2</sup> vil koste?
- f) Finn forklaringsstyrken  $R^2$  uten å gjøre noe regning “for hånd”. Bare les av fra Excel-utskriften i figur 11.4.<sup>4</sup>

A	B	C	D	E	F	G	H	I
SUMMARY OUTPUT								
<b>Regression Statistics</b>								
Multiple R	0,970746809							
R Square	0,942349368							
Adjusted R Square	-1,5							
Standard Error	260,8356815							
Observations	1							
<b>ANOVA</b>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	6	4448379,822	741396,637	65,38345404	#NUM!			
Residual	4	272141,0111	68035,25276					
Total	10	4720520,833						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95,0%</i>	<i>Upper 95,0%</i>
X Variable 5	290,9395319	407,4209876	0,714100502	0,514621858	-840,2424749	1422,121539	-840,2424749	1422,121539
X Variable 6	41,6579974	5,151864817	8,086003589	0,001271107	27,35412755	55,96186725	27,35412755	55,96186725

Figure 11.4: Utskrift fra Excel.

<sup>3</sup>Bruk boken.<sup>4</sup>Hint: Se i boken.

g) Kommenter svaret i oppgave 11.2 f).<sup>5</sup>

h) Bruk Excel til å plote **regresjonslinjen** fra oppgave 11.2 d).

Skriv ut ditt Excel-plott og legg ved i din innlevering. Din Excel-utskrift skal se omtrent ut som figur 11.5, men med regresjonslinjen fra oppgave 11.2 d) i tillegg.<sup>6</sup>

PS:

Usikker på hvordan man lager regresjonsplott i Excel? Se video HER.

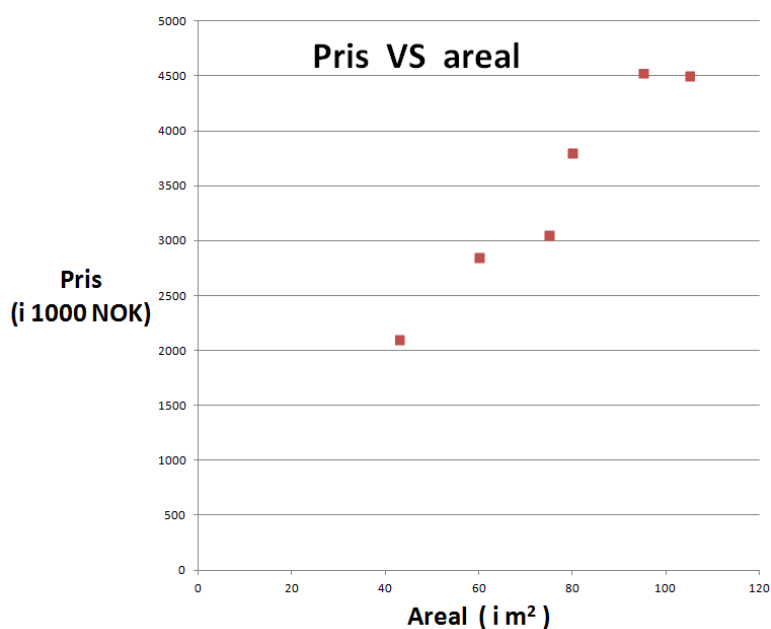


Figure 11.5: Excel.

■

<sup>5</sup>For et gitt areal, vi du si at regresjonslinjen predikerer prisen i stor eller liten grad? Med stort eller lite presisjonsnivå?

<sup>6</sup>Tittel på x-aksen kan f.eks. være “Areal (i m²)”. Tittel på y-aksen kan f.eks. være “Pris (i 1000 NOK)”.