

# Oppgavesett nr. 5

“*MAT110 Statistikk 1*”, 2020

## Oppgave 1: ( estimering - 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 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



Figur 1: McDonald's.

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 bearbeidningstiden ordren krever fra de ansatte. <sup>3</sup>

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. Formuler statistiske **modeller** for utvalgene og populasjonen.

---

<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.

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

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

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

$$z_i = (\textcolor{red}{x}_i, \textcolor{blue}{y}_i) \quad (1)$$

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

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

$$\textcolor{red}{x}_i = \text{tiden mellom kunde nr. } i \text{ og forrige kunde } i - 1 \text{ (IAT) (antall minutter)} \quad (2)$$

$$\textcolor{blue}{y}_i = \text{bearbeidingstiden for ordren til kunde nr. } i \text{ (antall minutter)} \quad (3)$$

- 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.



Figur 2: Datainnsamling.

I tabell 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)

Tabell 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  $y_i$  bearbeidningstiden. <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:

$$\begin{aligned}[0.0, 0.2) \\ [0.2, 0.4) \\ \vdots \\ [2.8, 3.0)\end{aligned}$$

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

$$\begin{aligned}[0, 1) \\ [1, 2) \\ \vdots \\ [11, 12)\end{aligned}$$

---

<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 5 i kompendiet.  
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 1 ligger på Canvas. Da blir det ikke så mye arbeid. Oppgi tallene med 3 desimalers nøyaktighet.

Definer populasjonsvariablene:

$\textcolor{red}{X}$  = tiden mellom to påfølgende kunder i tidsrommet 16:00-18:00 på en gitt hverdag (4)

$\textcolor{blue}{Y}$  = bearbeidingstiden til en kunde i i tidsrommet 16:00-18:00 på en gitt hverdag (5)

Definer tilhørende forsøksvariabler:

$$(\textcolor{red}{X}_1, \textcolor{blue}{Y}_1), (\textcolor{red}{X}_2, \textcolor{blue}{Y}_2), \dots, (\textcolor{red}{X}_n, \textcolor{blue}{Y}_n) \quad (6)$$

e) Er

$$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \textcolor{red}{X} \quad (7)$$

$$Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \textcolor{blue}{Y} \quad (8)$$

rimelig å anta?

Gi en kort begrunnelse.

f) Er det rimelig å anta at  $\textcolor{red}{X}_i$ -ene er gjensidig uavhengige av  $\textcolor{blue}{Y}_i$  ene?  
Gi en kort begrunnelse.

- g)** På bakgrunn av kunnskapen vi har om McDonalds-eksemplet samt de beskrivende nøkkeltallene for observasjonene, formulér to statistiske modeller:

1. Statistisk modell for IAT - Inter-Arrival Times ( $X_i$ -ene)
2. 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.

## Oppgave 2: ( estimering - logistikk , McDonald's )

Vi fortsetter i denne øvingen med caset fra oppgave 1 om McDonald's.

I oppgave 1 gjorde en forsøksrekke med  $n = 100$  tilfeldig utvalgte ankommende kunder i tidsrommet mellom 16:00 og 18:00 på hverdager.

For hver ankommende kunde som ble valgt, målte vi to datastørrelser  $(x_i, y_i)$  hvor:

$$x_i \quad = \quad \text{tiden mellom kunde nr. } i \text{ og forrige kunde } i - 1 \text{ (IAT) (antall minutter)} \quad (9)$$

$$y_i \quad = \quad \text{bearbeidingstiden for ordren til kunde nr. } i \text{ (antall minutter)} \quad (10)$$

Vi husker også at for å få **identisk** og **uavhengig fordelte** forsøksvariabler (*i.i.d.*) måtte vi definere en "kunde" som en gruppe personer som ankommer McDonald's som en samlet enhet, f.eks. en familie, en vennegjeng eller en skoleklasse etc.

Vi bruker samme tabell som i oppgave 1, dvs. vi bruker tabell 1.

I tabell 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.



Foto: Andreas Witzøe

Figur 3: McDonald's.

Populasjonsvariabler: ( blant alle **potensielle** kunde i det aktuelle tidsrommet ) <sup>9</sup>

$X$  = inter-arrival time - IAT (11)

dvs. tiden mellom når en **vilkårlig kunde** ankommer  
og når forrige kunde ankom i tidsrommet 16:00 - 18:00 på en vilkårlig hverdag

$Y$  = en **vilkårlig kundes** ordrestørrelse (dvs. bearbeidstid) (12)  
på en vilkårlig hverdag mellom 16:00 - 18:00

Tilhørende forsøksvariabler:

$X_1, X_2, \dots, X_{100}$  og  $Y_1, Y_2, \dots, Y_{100}$  (13)

---

<sup>9</sup>Vilkårlig kunde - av alle **potensielle** kunder hos McDonald's i Molde i hverdagene fra 16:00 til 18:00.

Fra beskrivende statistikk og kjent forskning rundt inter-arrival times, konkluderte vi i øving 5 med at vi har følgende statistiske modeller for dataene:

Den statistiske modellen for de stokastiske forsøksvariablene  $X_1, X_2, \dots, X_{100}$  og populasjonsvariabelen  $X$  er gitt ved:

$$X_1, X_2, X_3, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X \sim \text{Exp}[\lambda]_{\lambda > 0} \quad (14)$$

hvor

$$(0, \infty) = \text{verdimengden } V \text{ til de stokastiske variablene} \quad (15)$$

$$\text{Exp}[\lambda]_{\lambda > 0} = \text{familien av eksponential-fordelinger med parameter } \theta = \lambda \quad (16)$$

$$(0, \infty) = \text{parametermengden } \Theta, \quad (17)$$

dvs. de mulige verdiene for parameteren  $\theta = \lambda$  som er  $\lambda > 0$

Den statistiske modellen for de stokastiske forsøksvariablene  $Y_1, Y_2, \dots, Y_{100}$  og populasjonsvariabelen  $Y$  er gitt ved:

$$Y_1, Y_2, Y_3, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} Y \sim N[\mu, \sigma]_{(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+} \quad (18)$$

hvor <sup>10</sup>

$$\overbrace{\mathbb{R}}^{\text{reelle tall}} = \text{verdimengden } V \text{ til de stokastiske variablene} \quad (19)$$

$$N[\mu, \sigma]_{(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+} = \text{familien av normalfordelinger med parametre } \mu \text{ og } \sigma \quad (20)$$

$$\mathbb{R} \times \mathbb{R}_+ = \text{parametermengden } \Theta, \quad (21)$$

dvs. de mulige verdiene for parameteren  $\mu$  og  $\sigma$

---

<sup>10</sup> $\mathbb{R}_+$  = alle **positive** reelle tall. Elementene i  $\mathbb{R} \times \mathbb{R}_+$  er alle par  $(\mu, \sigma)$ , hvor  $-\infty < \mu < \infty$  og  $\sigma > 0$ .

De tre parametrene

$$\underbrace{\lambda}_{\text{ukjent}}, \underbrace{\mu}, \underbrace{\sigma} \quad (22)$$

er alle ukjente størrelser som vi i denne øvingen skal estimere.  
Vi foreslår følgende estimatorer for  $\lambda$ ,  $\mu$  og  $\sigma$ :

$$\hat{\lambda} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (23)$$

$$\hat{\mu} = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad (24)$$

$$\hat{\sigma}^2 = S_y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (25)$$

Vi vet fra kapittel 6 i kompendiet at  $\hat{\mu}$  og  $\hat{\sigma}^2$  er forventningsrette estimatorer.

## Forventningsrett og realiseringer

- a) Vis at  $\hat{\lambda}$  er en forventningsrett estimator for  $\lambda$ . <sup>11</sup>
- b) Bestem realiseringer for  $\lambda$ ,  $\mu$  og  $\sigma$  (punkttestimater) basert på dataene i tabell 1. <sup>12</sup> <sup>13</sup>

---

<sup>11</sup>Bruk at  $E[\mathbf{X}_i] = \lambda$ .

<sup>12</sup>Bruk resultatene du fikk fra oppgave 1.

<sup>13</sup>Eksponentialfordelingen har to versjoner, og vi tar utgangspunkt i tilfellet hvor tetthetsfunksjonen er gitt ved  $f(x) = \frac{1}{\lambda}e^{-\frac{1}{\lambda}x}$ . Husk også at:

- 1) Parameteren  $\lambda$  tilhører  $\mathbf{X}_i$ -variablene, dvs. IAT-tiden.
- 2) Parametrene  $\mu$  og  $\sigma$  tilhører  $\mathbf{Y}_i$ -variablene, dvs. bearbeidingsstiden.

## Konfidensintervall for $\mu$

- c) Konstruer et asymptotisk  $(1 - \alpha) 100\%$  konfidensintervall for  $\mu$  ( $= E[Y]$ ) ved hjelp av *sentralgrensesetningen*.<sup>14</sup>

Regn ut realiseringen av konfidensintervallet for  $\alpha = 0.01$  basert på dataene i tabell 1.

- d) Skriv ned et asymptotisk  $(1 - \alpha) 100\%$  konfidensintervall for  $\mu$  ( $= E[Y]$ ) ved hjelp av *Student's t-fordelingen*.<sup>15</sup>

Regn ut en realisering av konfidensintervallet for  $\alpha = 0.01$  basert på dataene i tabell 1.

Forventer du at intervallet blir større eller mindre sammenlignet med intervallet i oppgave 2c? Gi en kort begrunnelse.

PS:

- 1) Via kompendiet ser vi at vi kan bruke samme formelen som vi konstruerte i oppgave 1c, men bytter ut kvantilene  $z_{\alpha/2}$  og  $z_{1-\alpha/2}$  til normalfordelingen med kvantilene  $q_{\alpha/2}$  og  $q_{1-\alpha/2}$  til Student's t-fordelingen.

Grunnen at at vi kan gjøre dette er at Student's t-fordelingen og normalfordelingen er svært like. Begge er symmetriske fordelinger, men for små  $n$  er halene mye "tykkere" enn normalfordelingen.

- 2) For å finne tabellen for kvantilene til Student's t-fordelingen, bruk gjerne internett.

---

<sup>14</sup>Se hvordan dette gjøres i kompendiet. Kom på øvingstimene å få hjelp.

<sup>15</sup>Her, i oppgave 2d, skal du bare skrive ned konfidensintervallet uten utregninger. Se kompendiet.

## Konfidensintervall for $\sigma$

- e) Skriv ned et asymptotisk  $(1 - \alpha) 100\%$  konfidensintervall for  $\sigma$  ( $= \sigma[Y]$ ) ved hjelp av sentralgrensesetningen.<sup>16</sup>

Regn ut realiseringen av konfidensintervallet for  $\alpha = 0.01$  basert på dataene i tabell 1.

- f) Konstruer et eksakt  $(1 - \alpha) 100\%$  konfidensintervall for  $\sigma$  ( $= \sigma[Y]$ ) ved hjelp av  $\chi^2$ -fordelingen.<sup>17</sup>

Regn ut en realisering av konfidensintervallet for  $\alpha = 0.01$  basert på dataene i tabell 1.

---

<sup>16</sup>Her, i oppgave 2e, skal du bare skrive ned konfidensintervallet uten utregninger. Se kompendiet.

<sup>17</sup>Se hvordan dette gjøres i kompendiet. Kom på øvingstimene å få hjelp.

## Konfidensintervall for $\lambda$

- g) Konstruer et asymptotisk  $(1 - \alpha)$  100 % konfidensintervall for  $\lambda$  ved hjelp av sentralgrensesetningen.<sup>18</sup> <sup>19</sup>

Regn ut en realisering av konfidensintervallet for  $\alpha = 0.01$  basert på dataene i tabell 1.

---

<sup>18</sup>Bruk at  $Var[\mathbf{X}_i] = \lambda^2$

<sup>19</sup>Dere skal *ikke* bytte  $\lambda$  med  $\hat{\lambda}$  i denne oppgaven.