

Skriftlig prøve: 26.06.2025

Kursus navn og nr.: **Introduktion til Statistik (02323)**

Varighed: 4 timer

Tilladte hjælpemidler: Alle hjælpemidler - ingen internetadgang

Dette sæt er besvaret af

\_\_\_\_\_  
(studienummer)

\_\_\_\_\_  
(underskrift)

\_\_\_\_\_  
(bord nr.)

Opgavesættet består af 30 spørgsmål af “multiple choice” typen, som er fordelt på 15 opgaver. For at besvare spørgsmålene skal du udfylde “multiple choice” siderne på eksamen.dtu.dk.

Der gives 5 point for et korrekt “multiple choice” svar og  $-1$  point for et forkert svar. KUN følgende 5 svarmuligheder er gyldige: 1, 2, 3, 4 eller 5. Hvis et spørgsmål efterlades blankt eller et ugyldigt svar angives, gives der 0 point for spørgsmålet. Endvidere, hvis mere end et svar angives til det samme spørgsmål, hvilket faktisk er teknisk muligt i online-systemet, gives der 0 point for spørgsmålet. Det antal point der kræves, for at opnå en bestemt karakter eller for at bestå eksamen afgøres endeligt ved censureringen.

**Den endelige besvarelse af opgaverne laves ved at udfylde og aflevere online. Skemaet her er KUN et nød-alternativ til dette. Husk at angive dit studienummer, hvis du afleverer på papir.**

| Opgave    | I.1 | I.2 | I.3 | II.1 | II.2 | III.1 | IV.1 | V.1 | V.2 | VI.1 |
|-----------|-----|-----|-----|------|------|-------|------|-----|-----|------|
| Spørgsmål | (1) | (2) | (3) | (4)  | (5)  | (6)   | (7)  | (8) | (9) | (10) |
| Svar      |     |     |     |      |      |       |      |     |     |      |

| Opgave    | VI.2 | VII.1 | VII.2 | VII.3 | VIII.1 | VIII.2 | VIII.3 | IX.1 | IX.2 | IX.3 |
|-----------|------|-------|-------|-------|--------|--------|--------|------|------|------|
| Spørgsmål | (11) | (12)  | (13)  | (14)  | (15)   | (16)   | (17)   | (18) | (19) | (20) |
| Svar      |      |       |       |       |        |        |        |      |      |      |

| Opgave    | X.1  | XI.1 | XI.2 | XII.1 | XII.2 | XIII.1 | XIV.1 | XIV.2 | XIV.3 | XV.1 |
|-----------|------|------|------|-------|-------|--------|-------|-------|-------|------|
| Spørgsmål | (21) | (22) | (23) | (24)  | (25)  | (26)   | (27)  | (28)  | (29)  | (30) |
| Svar      |      |      |      |       |       |        |       |       |       |      |

Eksamenssættet består af 28 sider.

Fortsæt på side 2

**Multiple choice opgaver:** Der gøres opmærksom på, at der i hvert spørgsmål er én og kun én svarmulighed, som er rigtig. Endvidere er det ikke givet, at alle de anførte alternative svarmuligheder er meningsfulde. Husk altid at afrunde dit eget resultat til antallet af decimaler givet i svarmulighederne før du vælger et svar. Husk også, at der kan forekomme små afvigelser mellem resultatet af bogens formler og tilsvarende indbyggede funktioner i Python.

### Opgave I

Lad  $X$  følge en normalfordeling med gennemsnit 2 og varians 16.

#### Spørgsmål I.1 (1)

Hvad er medianen for  $X$ ?

- 1 ☐  $\text{median}(X) = -4$
- 2 ☐  $\text{median}(X) = -2$
- 3 ☐  $\text{median}(X) = 0$
- 4 ☐  $\text{median}(X) = 2$
- 5 ☐  $\text{median}(X) = 4$

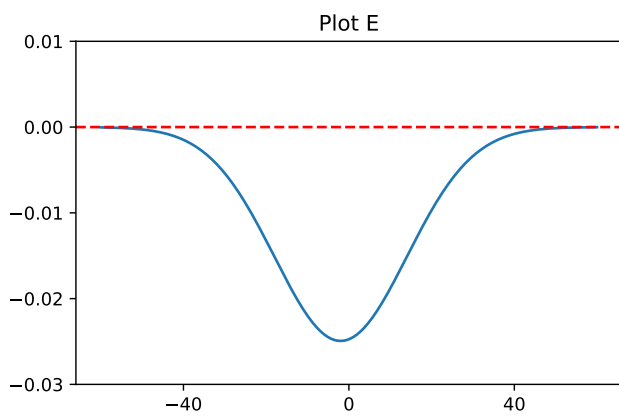
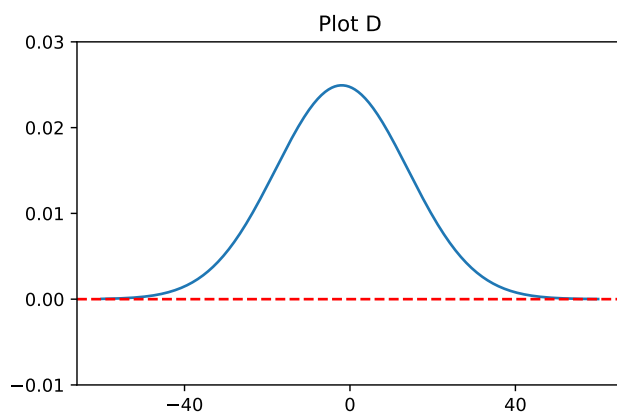
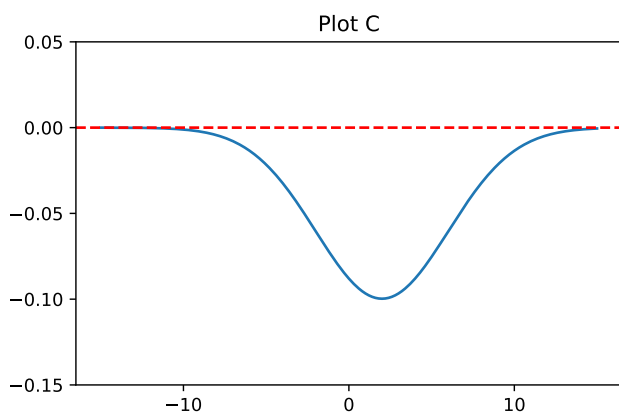
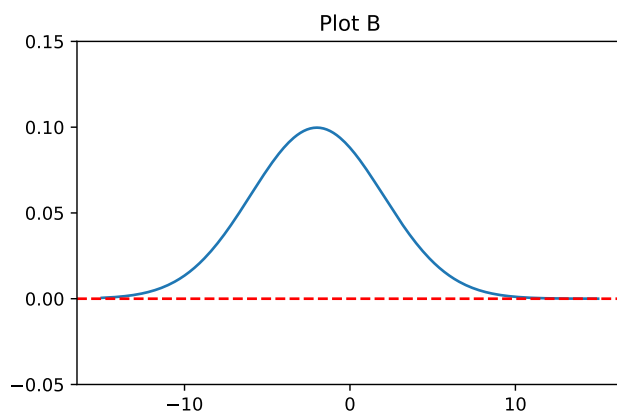
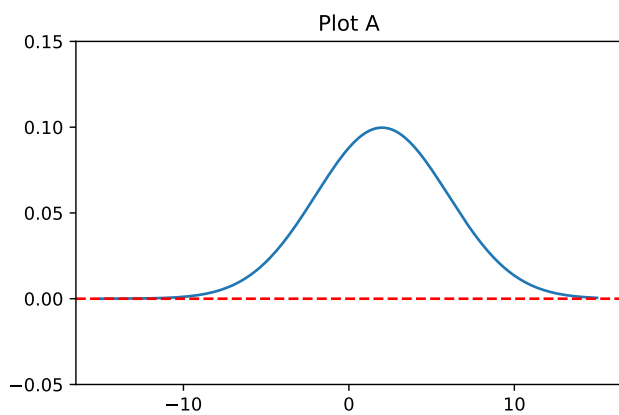
#### Spørgsmål I.2 (2)

Hvilken af følgende linjer kode udregner korrekt sandsynligheden  $P(X \leq 3)$ ?

- 1 ☐ `stats.norm.pdf(3, loc = 2, scale = 16)`
- 2 ☐ `stats.norm.pdf(3, loc = 2, scale = 4)`
- 3 ☐ `stats.norm.cdf(3, loc = 2, scale = 16)`
- 4 ☐ `stats.norm.cdf(3, loc = 2, scale = 4)`
- 5 ☐ `1 - stats.norm.cdf(3, loc = 2, scale = 16)`

### Spørgsmål I.3 (3)

Lad  $Y = -X$ . Hvilket af følgende plot viser tæthedsfunktionen (pdf) for  $Y$ ?



- 1 ☐ Plot A
- 2 ☐ Plot B
- 3 ☐ Plot C
- 4 ☐ Plot D
- 5 ☐ Plot E

## Opgave II

En ingeniør ønsker at teste, om en ny legering har en trækstyrke med middelværdi 500 MPa. En tilfældig stikprøve på 30 prøver testes, hvilket giver et stikprøvegennemsnit på 510 MPa og en stikprøve standardafvigelse på 20 MPa. Det antages at observationerne er iid og normalfordelt.

### Spørgsmål II.1 (4)

Hvad er den tilhørende  $p$ -værdi for den relevante hypotesetest med følgende hypoteser:

$$H_0 : \mu = 500, \quad H_A : \mu \neq 500.$$

- 1 ☐  $p = 0.010$
- 2 ☐  $p = 0.621$
- 3 ☐  $p = 0.310$
- 4 ☐  $p = 0.006$
- 5 ☐  $p = 0.005$

Ingeniøren planlægger nu et nyt forsøg hvor han gerne vil opnå en “margin of error” på højst 2 MPa. Han bruger den observerede standardafvigelse som scenarie og signifikansniveau  $\alpha = 0.05$ .

### Spørgsmål II.2 (5)

Hvor stor en stikprøve skal der udtages for at opnå en “margin of error” på højst 2 MPa?

- 1 ☐ ca. 20 observationer
- 2 ☐ ca. 40 observationer
- 3 ☐ ca. 1538 observationer
- 4 ☐ ca. 16 observationer
- 5 ☐ ca. 385 observationer

Fortsæt på side 5

### Opgave III

En træner ønsker at undersøge om der er forskel på forskellige typer af målrettet træning i forhold til at forbedre den tid det tager at løbe op ad trapper. Træneren indsamler data fra 15 forsøgspersoner, der inddeles (tilfældigt) i tre lige store grupper: Gruppe A, Gruppe B og Gruppe C. Træneren sætter forsøgspersonerne til at lave nogle målrettede øvelser over de næste 4 uger. Forsøgspersoner, der er i samme gruppe, laver de samme øvelser, men træneren uddeler forskellige øvelser til de tre grupper. For hver forsøgsperson indsamles data om forbedringen i den tid det tager personen at løbe op ad en trappe i træningscenteret (tidsforbedringen måles i sekunder).

De observerede tidsforbedringer er:

| Gruppe: | Tidsforbedring (målt i sekunder): |
|---------|-----------------------------------|
| A       | 2.1, 2.5, 2.3, 2.4, 2.2           |
| B       | 2.8, 2.9, 2.7, 3.0, 2.6           |
| C       | 2.3, 2.4, 2.5, 2.2, 2.1           |

Den gennemsnitlige tidsforbedring for samtlige 15 forsøgspersoner er  $\hat{\mu} = 2.467$  og de gennemsnitlige tidsforbedringer indenfor hver gruppe er givet ved:  $\hat{\mu}_A = 2.30$ ,  $\hat{\mu}_B = 2.80$ ,  $\hat{\mu}_C = 2.30$ . Det kan antages at alle observationer er uafhængige og normalfordelt.

#### Spørgsmål III.1 (6)

Hvad er den mest passende statistiske model og analyse, når man ønsker at undersøge om der er forskel på effekten af de forskellige træningstyper?

- 1 ☐ En passende model kunne være  $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$  ( $\epsilon_{ij} \sim N(0, \sigma^2)$ ), hvor  $Y_{ij}$  er tidsforbedring af person nummer  $j$  i gruppe nr  $i$ . En relevant analyse ville da være en at udføre en t-test, der undersøger nulhypotesen  $H_0 : \mu = 0$ .
- 2 ☐ En passende model kunne være  $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$  ( $\epsilon_i \sim N(0, \sigma^2)$ ), hvor  $x_i$  er tidsforbedring af person nummer  $i$ . En relevant analyse ville da være en at udføre en t-test, der undersøger nulhypotesen  $H_0 : \beta_1 = 0$ .
- 3 ☐ En passende model kunne være  $Y_{ij} = \mu_i + \epsilon_{ij}$  ( $\epsilon_{ij} \sim N(0, \sigma^2)$ ), hvor  $Y_{ij}$  er tidsforbedring af person nummer  $j$  i gruppe nr  $i$ . En relevant analyse ville da være at udføre en variansanalyse, der undersøger nulhypotesen  $H_0 : \mu_A = \mu_B = \mu_C = 0$ .
- 4 ☐ En passende model kunne være  $Y_{ij} = \beta_0 + \beta_i x_{ij} + \epsilon_{ij}$  ( $\epsilon_{ij} \sim N(0, \sigma^2)$ ), hvor  $x_{ij}$  er tidsforbedring af person nummer  $j$  i gruppe nr  $i$ . En relevant analyse ville da være en variansanalyse, der undersøger nulhypotesen  $H_0 : \beta_i = 0$  (dvs. der udføres i alt 3 test - én for hver gruppe).
- 5 ☐ En passende model kunne være  $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$  ( $\epsilon_{ij} \sim N(0, \sigma^2)$ ), hvor  $Y_{ij}$  er tidsforbedring af person nummer  $j$  i gruppe nr  $i$ . En relevant analyse ville da være en variansanalyse, der undersøger nulhypotesen  $H_0 : \alpha_A = \alpha_B = \alpha_C = 0$ .

|                  |
|------------------|
| <b>Opgave IV</b> |
|------------------|

Antag at  $Y$  følger en eksponentialfordeling med  $E(Y) = 3$ .

**Spørgsmål IV.1 (7)**

Hvad er  $P(2 < Y < 4)$ ?

1 ☐ 0.49

2 ☐ 0.25

3 ☐ 0.61

4 ☐ 0.75

5 ☐ 0.0024

Fortsæt på side 7

## Opgave V

En dyreforretning ønsker at undersøge, hvor stor en andel af danske husholdninger der har en hund. De gennemfører en stikprøveundersøgelse blandt 1000 af deres kunder, hvor de bliver spurgt, om de har en hund. Forretningen antager at de 1000 kunder repræsenterer 1000 husholdninger.

Af disse svarer 320, at de har en hund.

### Spørgsmål V.1 (8)

Hvad er den estimerede andel ( $\hat{p}$ ) af husholdninger der har en hund og hvad er usikkerheden (standardfejlen,  $s.e._{\hat{p}}$ ) på denne andel?

- 1 ☐  $\hat{p} = 0.32$  og  $s.e._{\hat{p}} = 0.00022$
- 2 ☐  $\hat{p} = 0.32$  og  $s.e._{\hat{p}} = 0.015$
- 3 ☐  $\hat{p} = 0.32$  og  $s.e._{\hat{p}} = 0.047$
- 4 ☐  $\hat{p} = 0.32$  og  $s.e._{\hat{p}} = 0.32$
- 5 ☐  $\hat{p} = 0.32$  og  $s.e._{\hat{p}} = 0.010$

### Spørgsmål V.2 (9)

Officielle tal siger at ca 20% af danskerne har en hund. Dyreforretningen havde derfor forventet at deres undersøgelse ville resultere i en andel, der var tættere på 0.20. Er det sandsynligt at deres resultat - altså at hele 32% af husholdningerne har hund - skyldes tilfældig variation? Og kan det være rigtigt at den sande andel af danske husholdninger der har hund i virkeligheden (kun) er ca 20%?

- 1 ☐ Ja, dyreforretningen har ved et tilfælde udvalgt en stikprøve, hvor flere end forventet har en hund. Dette skyldes sandsynligvis tilfældig variation og den sande andel kan godt være ca 20%.
- 2 ☐ Nej, det er ikke sandsynligt at dyreforretningens resultat skyldes tilfældig variation.  $p$ -værdien for den relevante test er 0.0015, så man vil forkaste nulhypotesen om at den sande andel er 0.20. Man må dermed konkludere at den sande andel sandsynligvis ikke er 20%.
- 3 ☐ Nej, det er ikke sandsynligt at dyreforretningens resultat skyldes tilfældig variation.  $p$ -værdien for den relevante test er 0.0015, så man vil forkaste nulhypotesen om at den sande andel er 0.20. Det er dog tvivlsomt om stikprøven er repræsentativ, så den sande andel kan godt være 20%.

- 4 ☐ Nej, det er ikke sandsynligt at dyreforretningens resultat skyldes tilfældig variation.  $p$ -værdien for den relevante test er  $2 \cdot 10^{-21}$ , så man vil forkaste nulhypotesen om at den sande andel er 0.20. Da stikprøven klart er repræsentativ må man konkludere at den sande andel af husholdninger, der har en hund, sandsynligvis ikke er 20%.
- 5 ☐ Nej, det er ikke sandsynligt at dyreforretningens resultat skyldes tilfældig variation.  $p$ -værdien for den relevante test er  $2 \cdot 10^{-21}$ , så man vil forkaste nulhypotesen om at den sande andel er 0.20. Det er dog tvivlsomt om stikprøven er repræsentativ, så den sande andel kan godt være 20%.

Fortsæt på side 9

## Opgave VI

I et studie haves data fra 4 forskellige grupper:

|           |                      |
|-----------|----------------------|
| Gruppe 1: | 89, 102, 94, 90, 100 |
| Gruppe 2: | 78, 46, 65, 72, 69   |
| Gruppe 3: | 83, 89, 81, 89, 90   |
| Gruppe 4: | 82, 101, 93, 88, 104 |

Tabellen kan indtastes i Python med følgende kode.

```
y = np.array([89, 102, 94, 90, 100,
              78, 46, 65, 72, 69,
              83, 89, 81, 89, 90,
              82, 101, 93, 88, 104])
Group = pd.Categorical([1,1,1,1,1,2,2,2,2,2,3,3,3,3,3,4,4,4,4,4])
D = pd.DataFrame({'y': y, 'Group': Group})
```

Det kan antages at data kan beskrives med følgende model:  $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ ,  $\epsilon_{ij} \sim N(0, \sigma^2)$ .

### Spørgsmål VI.1 (10)

Hvad er variationen mellem grupper,  $MS(\text{Group})$ , og variationen indenfor grupper, MSE?

- 1 ☐  $MS(\text{Group}) = 190.3$  og  $MSE = 1122$
- 2 ☐  $MS(\text{Group}) = 894.5$  og  $MSE = 70.15$
- 3 ☐  $MS(\text{Group}) = 2683$  og  $MSE = 1122$
- 4 ☐  $MS(\text{Group}) = 894.5$  og  $MSE = 2683$
- 5 ☐  $MS(\text{Group}) = 190.3$  og  $MSE = 70.15$

Fortsæt på side 10

### Spørgsmål VI.2 (11)

Hvilket udsagn om modellen ovenfor er IKKE korrekt?

- 1 ☐  $Y_{ij}$  er observation nr.  $j$  i gruppe nr.  $i$ .  $\hat{\alpha}_i$  er gruppe  $i$ 's gennemsnitlige afvigelse fra det overordnede gennemsnit  $\hat{\mu}$ .
- 2 ☐ Den samlede varians af data (dvs.  $\frac{1}{N-1}SST$ ) kan ikke være større end  $\hat{\sigma}^2$ .
- 3 ☐ MSE angiver variansen indenfor hver gruppe, og da vi antager at denne er ens i alle grupper har vi også  $MSE = \hat{\sigma}^2$ .
- 4 ☐ Hvis variansen af  $\alpha_i$ 'erne er stor sammenlignet med MSE betyder dette at der er forskel på grupperne.
- 5 ☐ I ovenstående data fås (for Gruppe 1,  $i = 1$ )  $\hat{\alpha}_1 = 9.75$ .

Fortsæt på side 11

## Opgave VII

Lad  $X \sim \mathcal{N}(50, 2^2)$  og  $Y \sim \mathcal{N}(45, 5^2)$  være uafhængige, og definer de stokastiske variable  $V = 3X + Y$  og  $U = \max(X, Y)$ , dvs.  $U$  er maksimum af  $X$  og  $Y$ . Brug simulering til at besvare de følgende tre spørgsmål.

### Spørgsmål VII.1 (12)

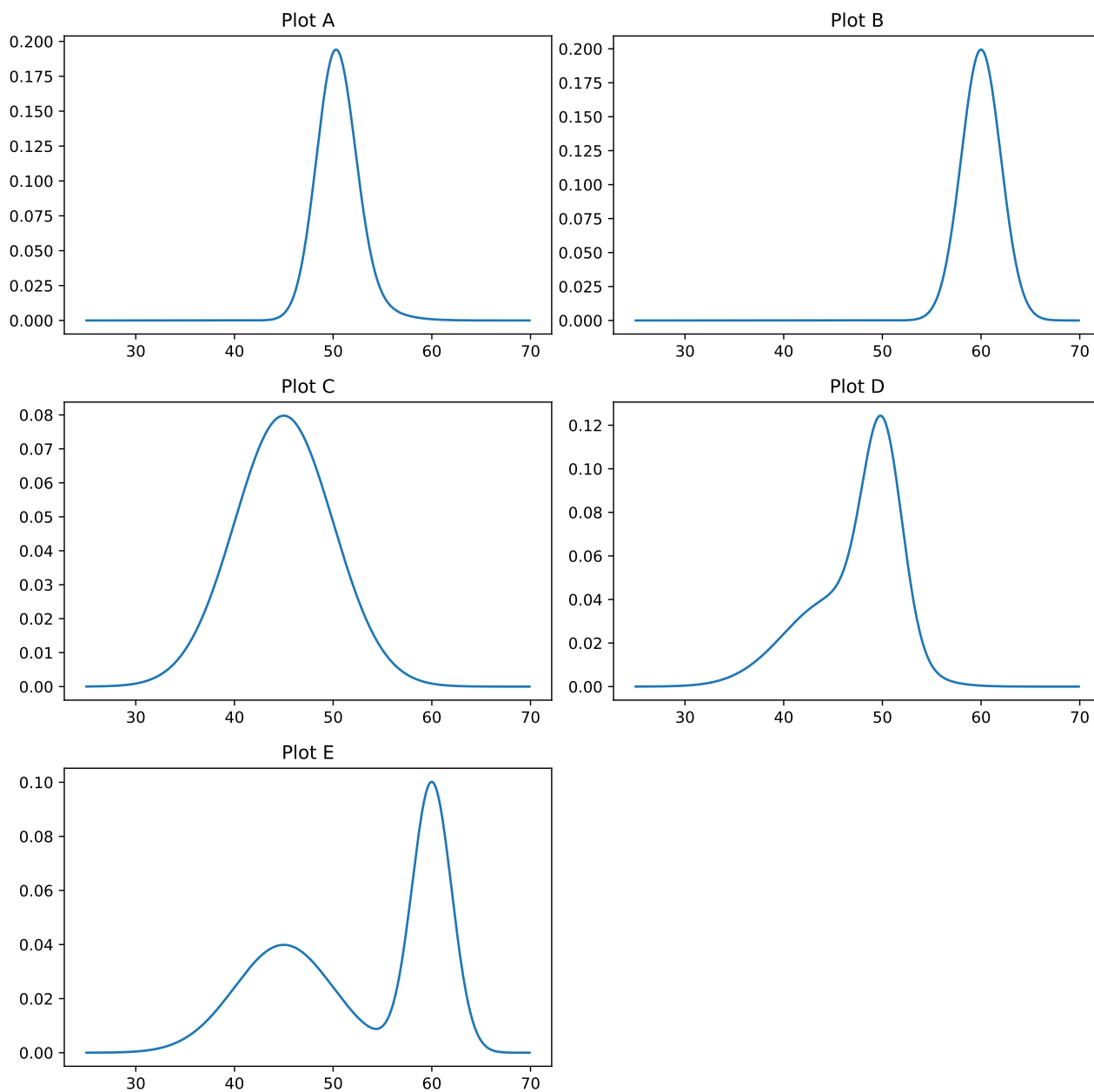
Hvad er sandsynligheden  $P(Y > X)$ ?

- 1 ☐ Cirka 17.7%
- 2 ☐ Cirka 42.2%
- 3 ☐ Cirka 50.0%
- 4 ☐ Cirka 57.8%
- 5 ☐ Cirka 82.3%

Fortsæt på side 12

### Spørgsmål VII.2 (13)

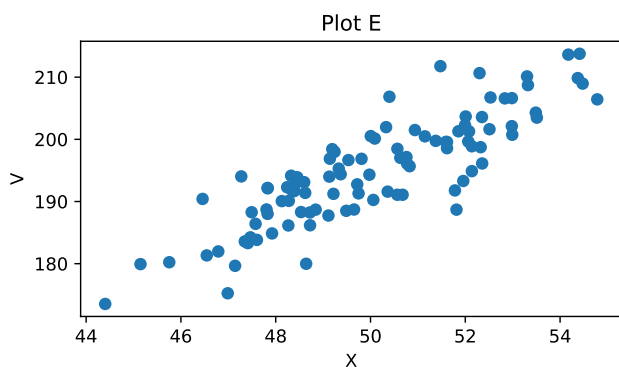
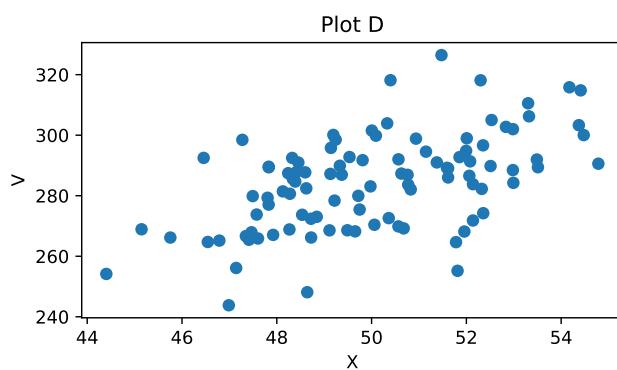
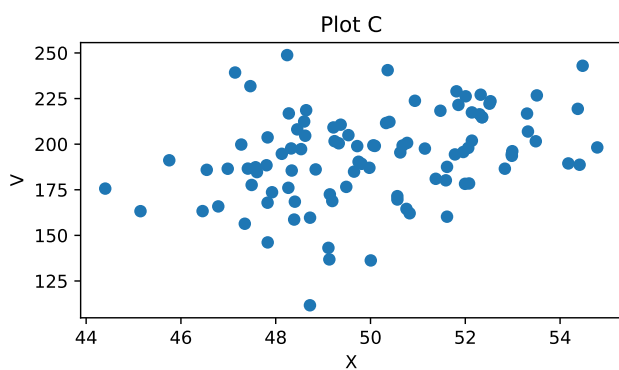
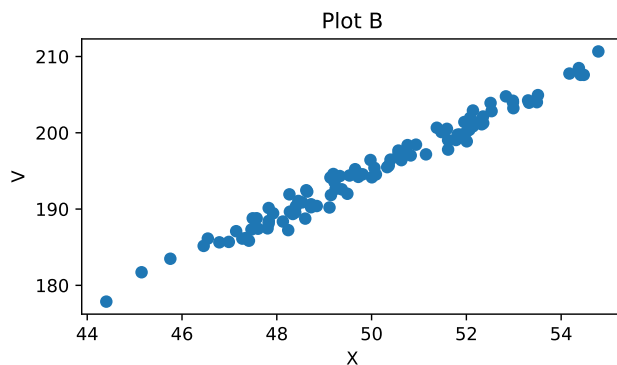
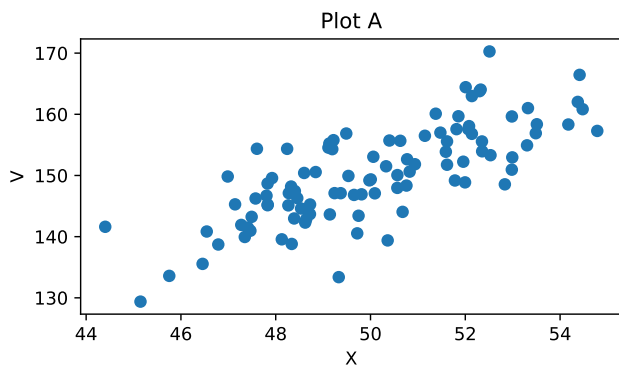
Hvilket af følgende plot viser tæthedsfunktionen (pdf) for  $U$ ?



- 1 ☐ Plot A
- 2 ☐ Plot B
- 3 ☐ Plot C
- 4 ☐ Plot D
- 5 ☐ Plot E

### Spørgsmål VII.3 (14)

Hvilket af de følgende plot viser med størst sandsynlighed et punktdiagram (scatter plot) af 100 observationer af  $(X, V)$ ?



- 1 ☐ Plot A
- 2 ☐ Plot B
- 3 ☐ Plot C
- 4 ☐ Plot D
- 5 ☐ Plot E

Fortsæt på side 14

### Opgave VIII

Politiet foretager en trafikkontrol, hvor de tilfældigt udvælger biler til inspektion. I gennemsnit kontrollerer de 10 biler i timen, og antallet af biler, der udvælges til inspektion i løbet af en time, følger en Poissonfordeling. Lad  $X$  betegne det samlede antal biler, der udvælges i løbet af en periode på fem timer.

#### Spørgsmål VIII.1 (15)

Hvad er den mest passende model for  $X$ ?

- 1 ☐  $X$  følger en binomialfordeling med parametrene  $n = 5$  og  $p = 0.1$ .
- 2 ☐  $X$  følger en binomialfordeling med parametrene  $n = 5$  og  $p = 0.5$ .
- 3 ☐  $X$  følger en Poissonfordeling med raten  $\lambda = 2$ .
- 4 ☐  $X$  følger en Poissonfordeling med raten  $\lambda = 10$ .
- 5 ☐  $X$  følger en Poissonfordeling med raten  $\lambda = 50$ .

#### Spørgsmål VIII.2 (16)

Hvad er sandsynligheden for, at ingen biler udvælges til inspektion i løbet af en given time?

- 1 ☐ Cirka 0%
- 2 ☐ Cirka 10%
- 3 ☐ Cirka 20%
- 4 ☐ Cirka 80%
- 5 ☐ Cirka 90%

Fortsæt på side 15

### Spørgsmål VIII.3 (17)

Historiske data viser, at 40% af inspektionerne resulterer i en bøde. Hvis politiet i løbet af en given time tilfældigt udvælger tre biler til inspektion, hvad er da sandsynligheden for, at præcis to af inspektionerne resulterer i en bøde?

1 ☐ 0.096

2 ☐ 0.144

3 ☐ 0.288

4 ☐ 0.432

5 ☐ 0.720

Fortsæt på side 16

## Opgave IX

Et par planlægger at købe et hus og ønsker at estimere den forventede pris. De indsamler data om nylige boligsalg i området og bruger Python til at fitte en multipel lineær regressionsmodel med boligprisen som responsvariabel og størrelsen af huset og grunden (i kvadratmeter) som forklarende variable.

De er særligt interesserede i to meget lignende ejendomme. Ejendom A har et hus på 175 kvadratmeter og en grund på 800 kvadratmeter, mens Ejendom B har samme grundstørrelse men et hus, der er 10 kvadratmeter mindre, dvs. 165 kvadratmeter.

Parret får følgende output fra deres model i Python, hvor prisen er angivet i tkr. (DKK tusinde kroner).

```
##                                OLS Regression Results
## =====
## Dep. Variable:                  Pris    R-squared:                  0.892
## Model:                        OLS      Adj. R-squared:            0.887
## No. Observations:              45      F-statistic:              173.8
## Covariance Type:              nonrobust Prob (F-statistic):        4.84e-21
## =====
##                                coef      std err          t      P>|t|      [0.025      0.975]
## -----
## Intercept      585.8165      613.968        0.954      0.345      -653.220      1824.853
## Hus            43.9742       2.757       15.951      0.000       38.411       49.538
## Grund          3.1214       0.275       11.334      0.000        2.566        3.677
## =====
```

### Spørgsmål IX.1 (18)

Hvad er den forventede pris på Ejendom A ifølge modellen?

- 1 ☐ Cirka DKK 10.193 M
- 2 ☐ Cirka DKK 10.778 M
- 3 ☐ Cirka DKK 20.200 M
- 4 ☐ Cirka DKK 35.726 M
- 5 ☐ Cirka DKK 36.311 M

Fortsæt på side 17

En ejendomsmægler påstår, at Ejendom B bør koste 400.000 kroner mindre end Ejendom A, hvilket svarer til, at hver kvadratmeter hus er 40.000 kroner værd. Parret ønsker at vurdere, om dataene understøtter denne påstand (nulhypotese) baseret på den fittede model.

**Spørgsmål IX.2 (19)**

Den sædvanlige test af ejendomsmæglerens påstand (nulhypotese) giver anledning til hvilken  $p$ -værdi?

1 ☐  $p = 0.922$

2 ☐  $p = 0.692$

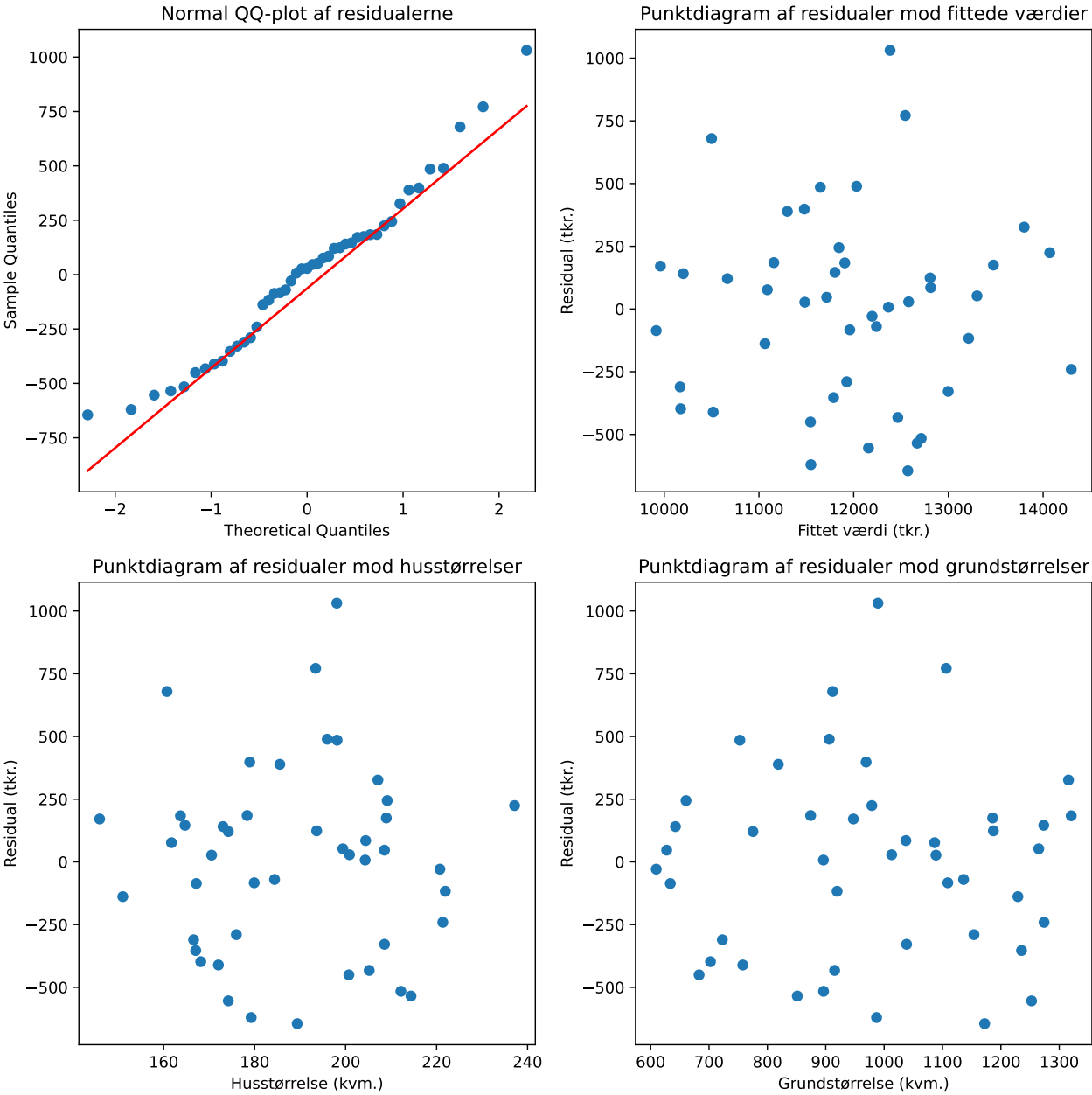
3 ☐  $p = 0.345$

4 ☐  $p = 0.157$

5 ☐  $p < 0.001$

Fortsæt på side 18

Parret validerer modellen og genererer følgende diagnostiske plot:



Fortsæt på side 19

### Spørgsmål IX.3 (20)

Hvilken af følgende udsagn er falsk?

- 1 ☐ Normal QQ-plottet af residualerne antyder, at residualerne kan være en smule højreskæve, men viser ingen åbenlyse brud på normalitetsantagelsen.
- 2 ☐ Normal QQ-plottet af residualerne antyder, at residualerne muligvis kan være afhængige, men viser ingen åbenlyse brud på antagelsen om uafhængighed.
- 3 ☐ Punktdiagrammet af residualer mod fittede værdier antyder, at variansen af residualerne muligvis øges med de fittede værdier, men viser ingen åbenlyse brud på antagelsen om homoskedasticitet (varianshomogenitet).
- 4 ☐ Punktdiagrammet af residualer mod husstørrelser antyder, at variansen af residualerne muligvis øges med husstørrelsen, men viser ingen åbenlyse brud på antagelsen om homoskedasticitet (varianshomogenitet).
- 5 ☐ Punktdiagrammet af residualer mod grundstørrelser antyder, at variansen af residualerne muligvis øges med grundstørrelsen, men viser ingen åbenlyse brud på antagelsen om homoskedasticitet (varianshomogenitet).

Fortsæt på side 20

### Opgave X

Man ønsker at sammenligne gennemsnittet i to stikprøver, A og B. Begge stikprøver indeholder 50 uafhængige målinger der antages normalfordelte. Det oplyses at 95% konfidensintervallet for middelværdien i hver gruppe er:

95% CI for  $\hat{\mu}_A = [15.2, 17.8]$

95% CI for  $\hat{\mu}_B = [13.0, 15.5]$ .

#### Spørgsmål X.1 (21)

Hvilket af følgende udsagn er korrekt?

- 1 ☐ Da konfidensintervallerne overlapper kan de to underliggende populationer godt have samme gennemsnit. Derfor ser vi nemt at forskellen mellem de to stikprøvegennemsnit ikke er signifikant forskellig fra nul (ved signifikansniveau på 5%).
- 2 ☐ Da konfidensintervallerne overlapper, er forskellen mellem de to stikprøvegennemsnit ikke statistisk signifikant forskellig fra nul (ved signifikansniveau på 5%). Dvs. de to underliggende populationer har samme fordeling.
- 3 ☐ Stikprøvegennemsnittet i hhv. stikprøve A og B er 16.50 og 14.25, og der er signifikant forskel på disse gennemsnit ved signifikansniveau på 5% (men ikke ved signifikansniveau 1%).
- 4 ☐ Stikprøvegennemsnittet i hhv. stikprøve A og B er 16.50 og 14.25, og der er signifikant forskel på disse gennemsnit (ved signifikansniveau på 1%).
- 5 ☐ Stikprøvegennemsnittet i hhv. stikprøve A og B er 16.50 og 14.25, men der er ikke signifikant forskel på disse gennemsnit (ved signifikansniveau på 5%).

Fortsæt på side 21

## Opgave XI

Fangst genfangst er en metode hvor et antal individer (dyr) indfanges, mærkes, og sættes ud. Efter en periode indfanges et antal individer og det undersøges hvor mange individer der er mærket. Metoden kan bruges til at estimere populationsstørrelser.

En biolog har indfanget  $n_1 = 150$  fisk i en sø, mærket dem og sat dem ud igen. Biologen planlægger nu at vende tilbage og indfange  $n_2 = 200$  fisk fra samme sø.

### Spørgsmål XI.1 (22)

Hvis vi benævner det totale antal fisk i søen ved  $N$  (og antager at  $N$  er det samme ved udsætning og genfangst), hvilken fordeling følger da antallet af mærkede fisk ( $Y$ ) ved genfangsten (det antages at alle mærkede fisk overlever og at det er helt tilfældigt hvilke af de  $N$  fisk der indfanges)?

- 1 ☐ En binomial fordeling med  $p = \frac{150}{N}$ , og  $n = 200$ , dvs.  $Y \sim B(200, \frac{150}{N})$ .
- 2 ☐ En normalfordelig med parametre  $\mu = \frac{200 \cdot 150}{N}$ , og  $\sigma^2 = \frac{200 \cdot 150}{N} (1 - \frac{150}{N})$ .
- 3 ☐ En hypergeometrisk fordeling med parametre  $n = 200$ ,  $a = 150$ , og  $N$ , dvs.  $Y \sim H(200, 150, N)$ .
- 4 ☐ En poisson fordeling med parameter  $\lambda = \frac{150 \cdot 200}{N}$ , dvs.  $Y \sim Pois(\frac{150 \cdot 200}{N})$ .
- 5 ☐ En exponential fordeling med parameter  $\lambda = \frac{N}{150 \cdot 200}$ , dvs.  $Y \sim Exp(\frac{N}{150 \cdot 200})$ .

Længden på de fangede fisk måles med henblik på at give et estimat på deres alder. Således klassificeres fisk på mellem 6 og 10 cm. som 1-årige, mens fisk over 10 cm. klassificeres som ældre. Det antages at længden på en et-årig fisk følger en normal fordeling med middelværdi  $\mu = 8$  cm. og standardafvigelse  $\sigma = 1$  cm.

### Spørgsmål XI.2 (23)

Hvad er sandsynligheden for at en et-årig fisk klassificeres som ældre end et år?

- 1 ☐ 0.159
- 2 ☐ 0.5
- 3 ☐ 0.841
- 4 ☐ 0.0228
- 5 ☐ 0.977

Fortsæt på side 22

## Opgave XII

En forbrugerorganisation ønsker at undersøge, hvor ofte et pakkeleverings firma leverer pakker til den nærmeste pakkeshop. De indsamler data fra 750 pakkeleveringer, fordelt på de 5 regioner. Resultaterne er opsummeret i følgende tabel:

| Region      | Leveret til nærmeste pakkeshop | Leveret et andet sted | Total |
|-------------|--------------------------------|-----------------------|-------|
| Hovedstaden | 40                             | 110                   | 150   |
| Midtjylland | 95                             | 55                    | 150   |
| Syddanmark  | 80                             | 70                    | 150   |
| Nordjylland | 85                             | 65                    | 150   |
| Sjælland    | 70                             | 80                    | 150   |
| Total       | 370                            | 380                   | 750   |

Der udføres nu en  $\chi^2$ -test, for at undersøge om andelen af pakker leveret til nærmeste pakkeshop er den samme i alle 5 regioner.

### Spørgsmål XII.1 (24)

Hvad er de - under nulhypotesen - forventede værdier i hver celle i tabellen?

1 ☐

| Region      | Leveret til nærmeste pakkeshop | Leveret et andet sted |
|-------------|--------------------------------|-----------------------|
| Hovedstaden | 75                             | 75                    |
| Midtjylland | 75                             | 75                    |
| Syddanmark  | 75                             | 75                    |
| Nordjylland | 75                             | 75                    |
| Sjælland    | 75                             | 75                    |

2 ☐

| Region      | Leveret til nærmeste pakkeshop | Leveret et andet sted |
|-------------|--------------------------------|-----------------------|
| Hovedstaden | 80                             | 70                    |
| Midtjylland | 70                             | 80                    |
| Syddanmark  | 60                             | 90                    |
| Nordjylland | 50                             | 100                   |
| Sjælland    | 40                             | 110                   |

3 ☐

| Region      | Leveret til nærmeste pakkeshop | Leveret et andet sted |
|-------------|--------------------------------|-----------------------|
| Hovedstaden | 100                            | 0                     |
| Midtjylland | 100                            | 0                     |
| Syddanmark  | 100                            | 0                     |
| Nordjylland | 100                            | 0                     |
| Sjælland    | 100                            | 0                     |

4 ☐

| Region      | Leveret til nærmeste pakkeshop | Leveret et andet sted |
|-------------|--------------------------------|-----------------------|
| Hovedstaden | 74                             | 76                    |
| Midtjylland | 74                             | 76                    |
| Syddanmark  | 74                             | 76                    |
| Nordjylland | 74                             | 76                    |
| Sjælland    | 74                             | 76                    |

5 ☐

| Region      | Leveret til nærmeste pakkeshop | Leveret et andet sted |
|-------------|--------------------------------|-----------------------|
| Hovedstaden | 40                             | 110                   |
| Midtjylland | 95                             | 55                    |
| Syddanmark  | 80                             | 70                    |
| Nordjylland | 85                             | 65                    |
| Sjælland    | 70                             | 80                    |

### Spørgsmål XII.2 (25)

Der udføres en  $\chi^2$ -test, for at undersøge om andelen af pakker leveret til nærmeste pakkeshop er den samme i alle 5 regioner, den relevante teststørrelse er udregnet til 47.21. Hvad er  $p$ -værdien for den relevante test og hvad er den tilsvarende konklusion (brug signifikansniveau  $\alpha = 0.05$ )?

- 1 ☐  $p$ -værdien bliver 0.10 og konklusionen er dermed at der er forskel i andelen af pakker leveret til nærmeste pakkeshop i de forskellige regioner
- 2 ☐  $p$ -værdien bliver 0.10 og konklusionen er dermed at der ikke er forskel i andelen af pakker leveret til nærmeste pakkeshop i de forskellige regioner
- 3 ☐  $p$ -værdien bliver 0.05 og konklusionen er dermed at der ikke er forskel i andelen af pakker leveret til nærmeste pakkeshop i de forskellige regioner
- 4 ☐  $p$ -værdien bliver  $1.4 \cdot 10^{-9}$  og konklusionen er dermed at der ikke er forskel i andelen af pakker leveret til nærmeste pakkeshop i de forskellige regioner
- 5 ☐  $p$ -værdien bliver  $1.4 \cdot 10^{-9}$  og konklusionen er dermed at der er forskel i andelen af pakker leveret til nærmeste pakkeshop i de forskellige regioner

Fortsæt på side 24

### Opgave XIII

En sensor måler temperaturen på en maskine, der helst ikke må blive varmere end  $80^{\circ}\text{C}$ . En stikprøve på 40 temperaturmålinger giver et stikprøvegennemsnit på  $75.2^{\circ}\text{C}$  og en stikprøve standardafvigelse på  $2.5^{\circ}\text{C}$ . Antag, at temperaturmålingerne er uafhængige af hinanden, samt at de følger en normalfordeling.

#### Spørgsmål XIII.1 (26)

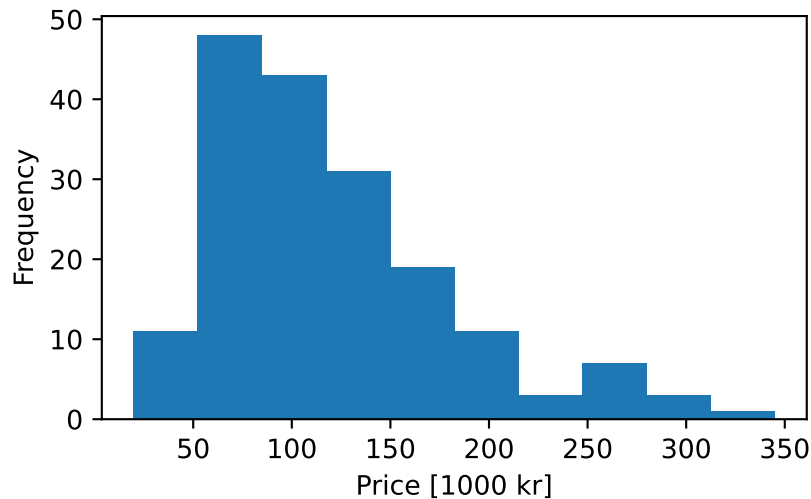
Hvad er et 95% konfidensinterval for den sande middelværdi af temperaturen?

- 1 ☐  $[72.7, 77.7]^{\circ}\text{C}$
- 2 ☐  $[70.1, 80.3]^{\circ}\text{C}$
- 3 ☐  $[74.4, 76.0]^{\circ}\text{C}$
- 4 ☐  $[75.1, 75.3]^{\circ}\text{C}$
- 5 ☐  $[71.0, 79.4]^{\circ}\text{C}$

Fortsæt på side 25

## Opgave XIV

En bilejer ønsker at købe en ny (brugt) bil, for at undersøge hvad prisen bør være har hun indsamlet priser på den bil (mærke og model) hun ønsker. Histogrammet herunder viser fordelinger af priser på den bil hun ønsker sig.



I tillæg til histogrammet har hun udregnet gennemsnit og empiriske varians for de observerede priser (`price`) [1000 kr.].

```
np.mean(price)
np.float64(118.94622598870056)

np.var(price, ddof=1)
np.float64(3633.1131006077294)
```

### Spørgsmål XIV.1 (27)

Baseret på ovenstående, hvilken af følgende antagelser om fordelingen af prisen ( $Y$ ) er så mest rimelig?

- 1 ☐ En log-normalfordeling med parametre  $\alpha = 4.66$  og  $\beta^2 = 0.478$ , dvs.  $Y \sim LN(4.66, 0.478^2)$ .
- 2 ☐ En normalfordeling med parametre  $\mu = 118.94$  og  $\sigma^2 = 3633.1^2$ , dvs.  $Y \sim N(118.9, 3633.1^2)$ .
- 3 ☐ En exponentialfordeling med parameter  $\lambda = 118.9$ , dvs.  $Y \sim Exp(118.9)$ .
- 4 ☐ En log-normal fordeling med parametre  $\alpha = 118.9$  og  $\beta^2 = 60.28^2$ , dvs.  $Y \sim LN(118.9, 60.28^2)$ .
- 5 ☐ En normalfordeling med parametre  $\mu = 118.94$  og  $\sigma^2 = 60.28^2$ , dvs.  $Y \sim N(118.9, 60.28^2)$ .

Bilejeren har også indsamlet data om alder og kilometerstand på bilerne. For at undersøge sammenhængen mellem pris, alder og kilometerstand på bilen har bilejeren kørt følgende kode (`price` [1000 kr.] er prisen, `age` [år] er alderen på bilen, `dist` [1000 km.] er kilometerstanden på bilen og `cars` er datasættet med de indsamlede tal),

```
fit = smf.ols('price ~ age + dist', data = cars).fit()
```

svarende til modellen

$$Y_i = \mu_i + \epsilon_i,$$

hvor formelen for  $\mu_i$  er bestemt fra inputtet til `smf.ols` og  $E(\epsilon_i) = 0$ .

### Spørgsmål XIV.2 (28)

Hvilket af følgende udsagn om modellen eller model antagelserne er korrekt?

- 1 ☐  $\epsilon_i$ 'erne er normalfordelt og iid. (uafhængige og identisk fordelte).
- 2 ☐ Det antages at den observerede korrelation mellem `age` og `dist` er lig nul.
- 3 ☐  $Y_i$ 'erne er normalfordelt og iid. (uafhængige og identisk fordelte).
- 4 ☐  $\mu_i$ 'erne følger en normalfordeling og er iid.
- 5 ☐  $\mu_i = \beta_1 x_{1i} + \beta_2 x_{2i}$ , hvor  $x_{1i}$  og  $x_{2i}$  er henholdsvis alderen og kilometerstanden på bil  $i$ .

Resultaterne af estimationen ovenfor er angivet nedenfor (idet en del tal er erstattet med symboler)

```
fit.summary(slim = True)
```

| OLS Regression Results |           |         |                     |          |        |        |
|------------------------|-----------|---------|---------------------|----------|--------|--------|
| =====                  |           |         |                     |          |        |        |
| Dep. Variable:         | price     |         | R-squared:          | 0.793    |        |        |
| Model:                 | OLS       |         | Adj. R-squared:     | 0.791    |        |        |
| No. Observations:      | 177       |         | F-statistic:        | 334.0    |        |        |
| Covariance Type:       | nonrobust |         | Prob (F-statistic): | 2.65e-60 |        |        |
| =====                  |           |         |                     |          |        |        |
|                        | coef      | std err | t                   | P> t     | [0.025 | 0.975] |
| -----                  |           |         |                     |          |        |        |
| Intercept              | 255.8098  | 5.707   | T1                  | P1       | Q1_1   | Qu_1   |
| age                    | -9.6077   | 0.963   | T2                  | P2       | Q1_2   | Qu_2   |
| dist                   | -0.4756   | 0.055   | T3                  | P3       | Q1_3   | Qu_3   |
| =====                  |           |         |                     |          |        |        |

### Spørgsmål XIV.3 (29)

Hvilket af følgende udsagn er korrekt ved brug af signifikans niveau  $\alpha = 0.05$ ?

- 1 ☐ Begge effekter (alder og kilometerstand) er signifikant forskellig fra nul og den forventede pris falder med alder og kilometerstand.
- 2 ☐ Ingen af effekterne (alder og kilometerstand) er signifikant forskellig fra nul.
- 3 ☐ Alder på bilen har en signifikant effekt på prisen, mens der ikke kan påvises en effekt af kilometerstand. Prisen stiger når alderen stiger.
- 4 ☐ Kilometerstand har en signifikant effekt på prisen, mens der ikke kan påvises en effekt af alder. Prisen falder når kilometerstand stiger.
- 5 ☐ Alder på bilen har en signifikant effekt på prisen, mens der ikke kan påvises en effekt af kilometerstand. Prisen falder når alderen stiger.

Fortsæt på side 28

## Opgave XV

En konsulent har modtaget data for komme- og gå-tider for 35 medarbejdere på en given arbejdsplads. Komme- og gå-tider er registreret for de samme 35 medarbejdere på to forskellige dage - én dag om sommeren og én dag om vinteren. Konsulenten ønsker nu at vurdere om den gennemsnitlige arbejdstid er lige lang på de to dage.

### Spørgsmål XV.1 (30)

Hvilken analyse er relevant at udføre?

- 1 ☐ For hver komme- og gå-tid udregnes arbejdstiden. Man har nu to uafhængige stikprøver (én for dagen om sommeren og én for dagen om vinteren) med 35 målinger i hver. Gennemsnittene af disse stikprøver sammenlignes med en t-test med nulhypotesen  $H_0 : \mu_1 = \mu_2$ .
- 2 ☐ For hver af de to dage beregnes den gennemsnitlige ankomsttid og den gennemsnitlige gå-tid. Nu udfører man to t-test: én t-test undersøger om der er signifikant forskel på ankomst-tiden og den anden t-test undersøger om der er signifikant forskel på gå-tiden.
- 3 ☐ For hver komme- og gå-tid udregnes arbejdstiden. Nu har man to parrede stikprøver (én for dagen om sommeren og én for dagen om vinteren) med 35 målinger i hver. Med en parret t-test undersøges det om den gennemsnitlige forskel i arbejdstid er signifikant forskellig fra nul.
- 4 ☐ For hver af de to dage beregnes et 95% konfidens interval for den gennemsnitlige ankomsttid  $CI_{\bar{x}_{arrive}}$  og for den gennemsnitlige gå-tid  $CI_{\bar{x}_{leave}}$ . Hvis de to konfidensintervaller ikke overlapper er der signifikant forskel på den gennemsnitlige arbejdstid på de to dage.
- 5 ☐ For hver komme- og gå-tid udregnes en samlet arbejdstid. Nu har man to stikprøver med 35 målinger. Men en one-way ANOVA model undersøges det om der er forskel på den gennemsnitlige arbejdstid melle på de to dage.