Oversikt over pensum

- Kapittel 1: Empirisk fordeling for en variabel
  - Begrepet fordeling
  - Mål for senter (gj.snitt, median) + persentiler/kvartiler
  - Mål for spredning (Standardavvik s, IQR)
  - Outliere
  - Grafiske framstillinger: Histogram etc., Boxplot, normalfordelingsplott
  - Lineærtransformasjoner, standardiserte variable
- Kapittel 2: Samvariasjon mellom to variable
  - Spredningsplott
  - Korrelasjon r
  - Enkel lineær regresjon $y_i=b_0+b_1 x_i$
  - Minste kvadraters metode
  - Sammenheng mellom r og b_1
  - Avvik fra linearitet, residualer
  - Outliere og innflytelsesrike observasjoner
  - Lurkende – konfunderende – bakenforliggende variable
  - Assosiasjon eller årsakssammenheng?

- Kapittel 3: Studieopplegg
  - Eksperimentelle versus observasjonelle studier

Eksperimentelle studier

  - Behandlingsgrupp(er) – kontrollgruppe – placebogruppe
  - Randomisering
  - Blindede og dobbeltblindede studier
  - Parvise sammenligninger og Blokkdelte eksperimenter

Observasjonelle studier

- Utvalgsundersøkelser fra endelig populasjon
  - Enkelt tilfeldig utvalg (SRS) og stratifiserte utvalg
- Utvalg fra (potensielt) uendelige populasjoner
  - Modell: Populasjonsparametre
  - Utvalg: Statistikk (observator, estimator)
  - Skjevhet (bias) versus presisjon

Etikk

  - Er det akseptabelt å utføre et eksperiment?
  - Må ofte ha tillatelse til å innhente observasjonelle data

Kapittel 4: Tilfeldighet - sannsynlighet

- Utfallsrom S og begivenheter (events)
- Relativ frekvens i det lange løp
- Uavhengige forsøk
- Regler for sannsynligheter inkludert produktregel for uavh. Begivenheter

Tilfeldige variable

- Diskrete og kontinuerlige fordelinger
- Forventning og varians
- Store talls lov: Gjennomsnittet tilnærmet lik forventningen for stor n

Betinget sannsynlighet

- Produktregel
- Bayes regel – trediagrammer

Kapittel 5: Samplingfordelinger ved SRS

- Gjennomsnitt, empirisk standardavvik og median etc. er tilfeldige var.
- Forventning og standardavvik for gjennomsnitt
- Sentralgrenseteorem, Gjennomsnitt tilnærmet normalfordelt
- Binomisk fordeling X ~ bin(n,p), med normaltilnærming

Kapittel 6: Intro til statistisk inferens

Konfidensintervall

- Intervall som inneholder interesseparameter med sannsynlighet C
- Med n uavhengige og normalfordelte $x_i \sim N(\mu, \sigma)$ der s er kjent blir et 95% konfidensintervall for μ gitt ved $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$
- Dette fordi utsagnet $\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$

  er det samme som (er "ekvivalent" med) at $-1.96 < z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < 1.96$

  samtidig som z er standardnormalfordelt og 1.96 = 97.5-persentil i N(0,1).

- Vi får konfidensintervall med annen konfidensgrad C ved å bytte ut 1.96 med andre persentiler i N(0,1), f.eks. gis konfidensgrad 90% ved å bruke 95-persentilen 1.645.

Gjennomgåelse Oblig 2

Oppgave 1: Tenk deg at du har 50 tilfeldige utvalg, hver av størrelse 25, fra en N(20,5)-fordelt populasjon. På grunnlag av hvert av de 50 utvalgene kan du danne et 90% konfidensintervall for forventnings-verdien (som vi her vet at er lik 20) ved å bruke formelen på side 359-360 i læreboka.

a) Hvor mange av de 50 konfidensintervallene venter du at vil inneholde den riktige verdien μ=20?

Svar: 50 intervaller x 0.90 = 45 intervaller
Begrunnelse:
Hvert intervall har sannsynlighet 90% = 0.9 for å inneholde μ
og det er 50 mulige intervaller.

b) La Y være en tilfeldig variabel som angir hvor mange av de 50 konfidensintervallene som vil inneholde den riktige verdien μ=20. Hvilken fordeling har Y? Svaret ditt må begrunnes!

Svar: Y er binomisk fordelt med parametre n=50 og p=0.90

Dette fordi (i) vi har en enten-eller situasjon for hvert intervall med suksessansynlighet p=0.9 for at μ er i intervallet

og (ii) at vi har n=50 uavhengige gjentak av denne situasjon.

Spesielt blir da forventningen til Y: n p = 50 x 0.9 = 45 (jfr. pkt a).

Probability Density Function

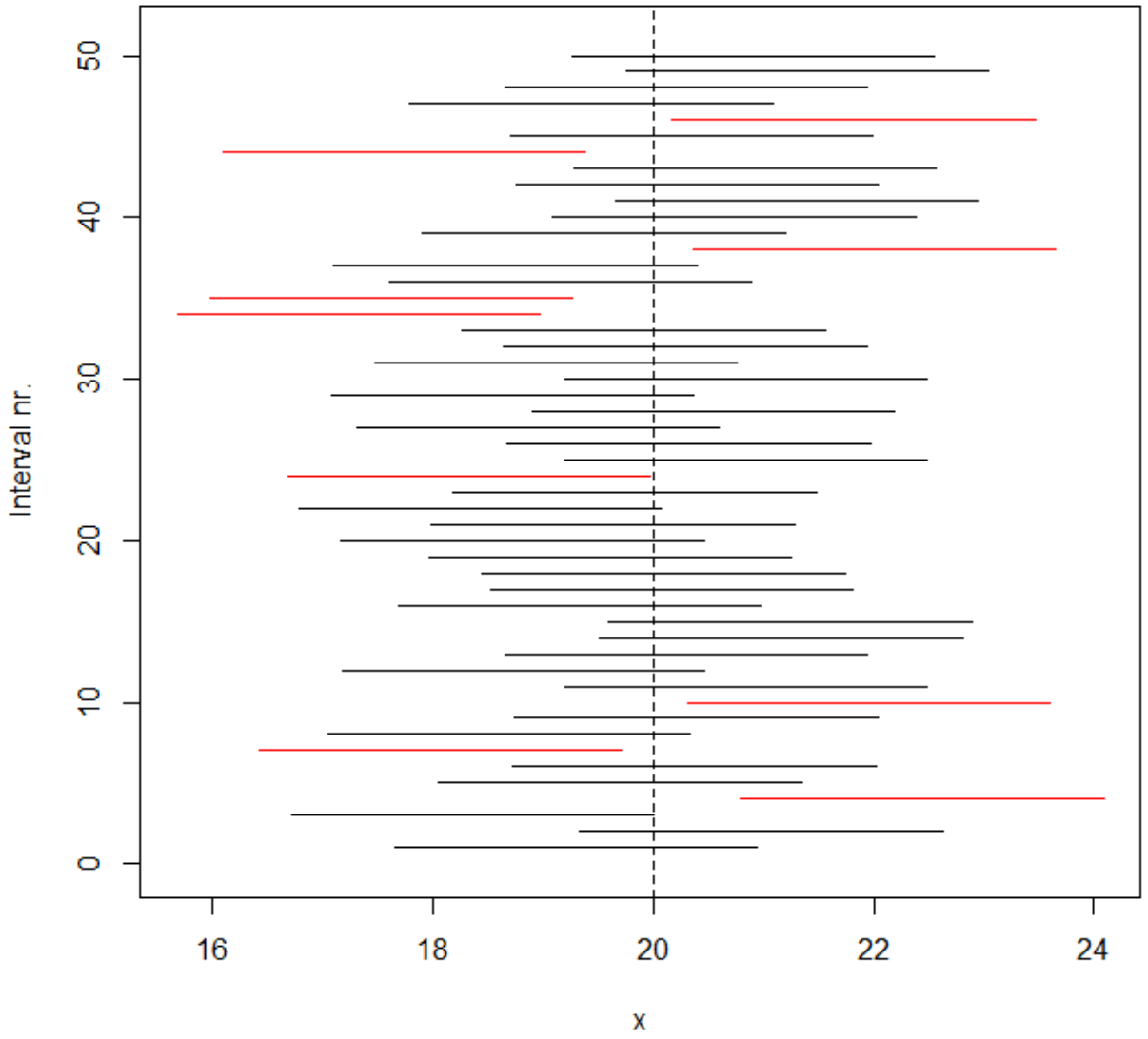
Binomial with n = 50 and p = 0,9

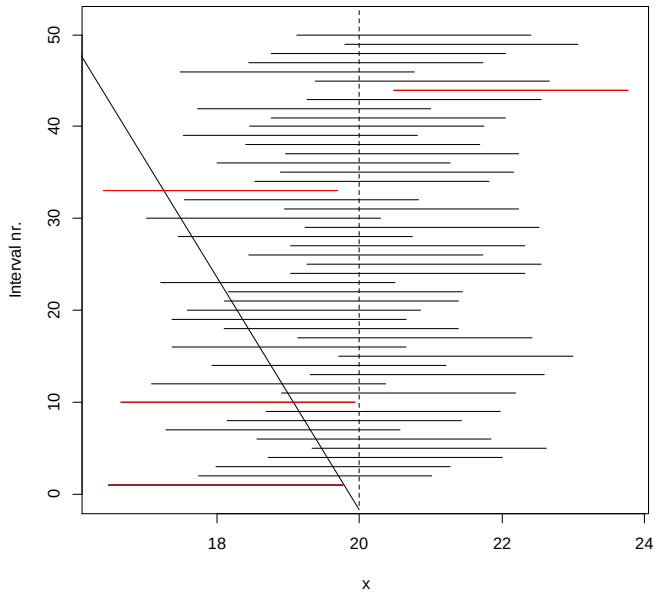
x	P(X = x)	x	P(X = x)	x	P(X = x)
40	0,0152	44	0,1541	48	0,0779
41	0,0333	45	0,1849	49	0,0286
42	0,0643	46	0,1809	50	0,0052
43	0,1076	47	0,1386		

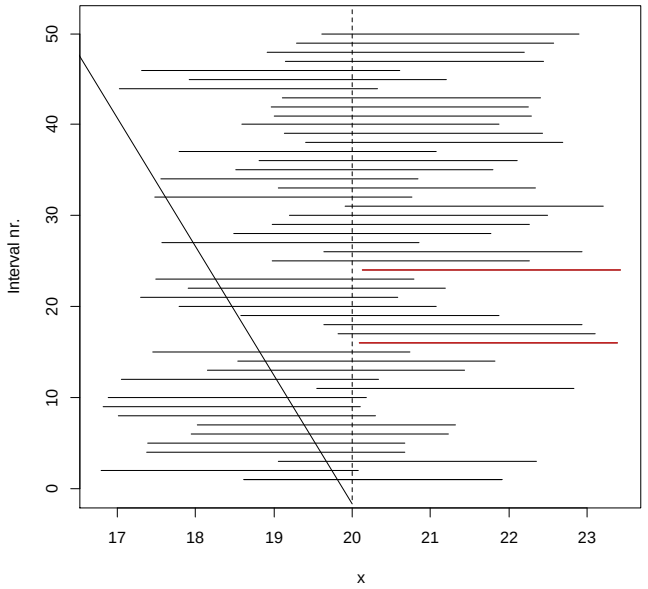
c) Utfør kommandoene ovenfor (for å genere 50 konfidensintervall) og angi de 50 kondensintervallene du får.
Hvor mange av dem inneholder den riktige forventningen 20? Svar: 6
Kommenter resultatet i lys av punkt a). Svar: forventning 5

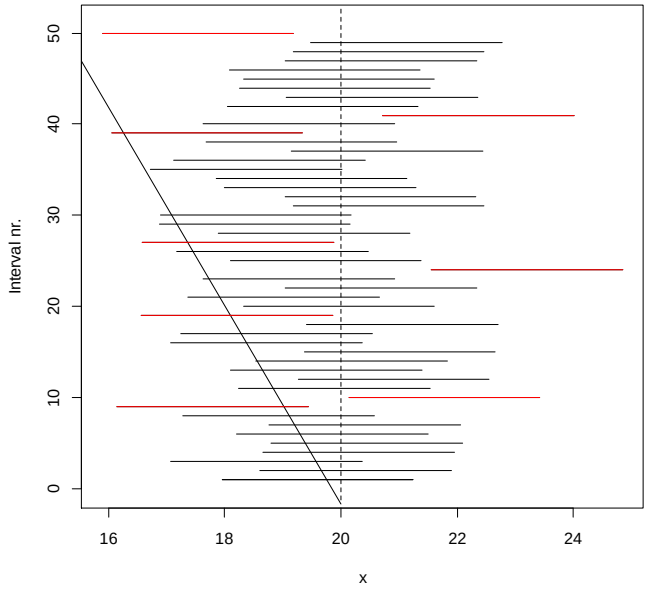
Variable	N	Mean	StDev	SE Mean	90% CI
C1	25	19,06	4,14	1,00	(17,42; 20,71)
C2	25	19,65	3,89	1,00	(18,01; 21,29)
C3	25	19,85	5,44	1,00	(18,20; 21,49)
C4	25	20,32	5,66	1,00	(18,67; 21,96)
C5	25	21,22	4,87	1,00	(19,57; 22,86)
C6	25	20,47	5,66	1,00	(18,82; 22,11)
C7	25	19,11	4,13	1,00	(17,47; 20,76)
C8	25	20,18	4,06	1,00	(18,53; 21,82)
C9	25	21,70	5,14	1,00	(20,05; 23,34)
C10	25	20,33	6,37	1,00	(18,69; 21,98)
C11	25	17,50	5,08	1,00	(15,86; 19,15)
C12	25	19,79	4,68	1,00	(18,15; 21,44)
C13	25	20,34	5,21	1,00	(18,69; 21,98)
C14	25	18,44	5,51	1,00	(16,79; 20,08)
C15	25	20,63	5,41	1,00	(18,98; 22,27)
C16	25	20,91	4,74	1,00	(19,27; 22,56)
C17	25	20,36	6,09	1,00	(18,72; 22,01)
C18	25	19,11	6,01	1,00	(17,46; 20,75)

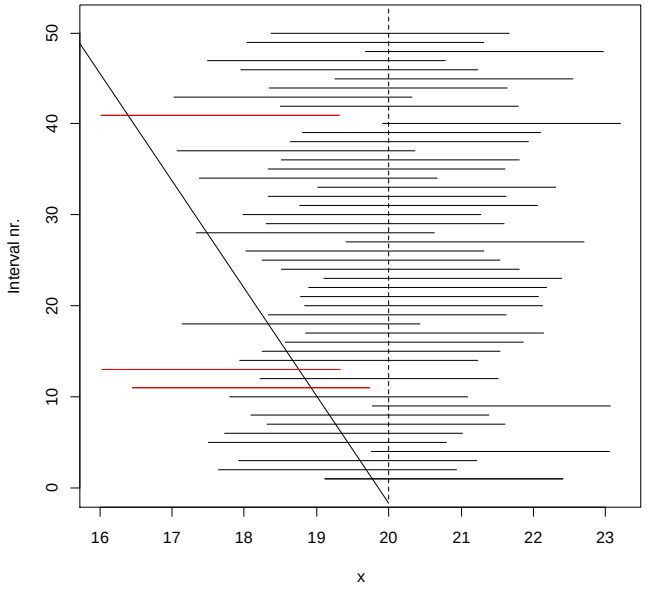
C19	25	20,66	5,08	1,00	(19,02; 22,31)
C20	25	21,32	5,55	1,00	(19,67; 22,96)
C21	25	20,59	4,58	1,00	(18,95; 22,24)
C22	25	19,71	3,69	1,00	(18,07; 21,36)
C23	25	19,66	5,12	1,00	(18,02; 21,31)
C24	25	19,28	4,83	1,00	(17,63; 20,92)
C25	25	21,06	4,55	1,00	(19,41; 22,70)
C26	25	19,32	6,20	1,00	(17,68; 20,97)
C27	25	20,15	5,35	1,00	(18,51; 21,80)
C28	25	19,70	4,66	1,00	(18,06; 21,35)
C29	25	18,95	4,87	1,00	(17,31; 20,60)
C30	25	19,99	4,93	1,00	(18,35; 21,64)
C31	25	21,09	5,73	1,00	(19,44; 22,73)
C32	25	20,80	6,09	1,00	(19,16; 22,45)
C33	25	18,85	5,22	1,00	(17,21; 20,50)
C34	25	20,55	4,65	1,00	(18,90; 22,19)
C35	25	17,93	5,25	1,00	(16,28; 19,57)
C36	25	20,84	3,66	1,00	(19,20; 22,49)
C37	25	19,92	5,15	1,00	(18,27; 21,56)
C38	25	19,86	5,08	1,00	(18,21; 21,50)
C39	25	20,00	5,41	1,00	(18,36; 21,65)
C40	25	20,67	4,60	1,00	(19,03; 22,32)
C41	25	20,51	5,11	1,00	(18,87; 22,16)
C42	25	18,91	4,15	1,00	(17,27; 20,56)
C43	25	21,02	5,82	1,00	(19,38; 22,67)
C44	25	21,87	5,44	1,00	(20,22; 23,51)
C45	25	20,15	4,53	1,00	(18,51; 21,79)
C46	25	19,80	5,45	1,00	(18,16; 21,45)
C47	25	21,93	4,84	1,00	(20,29; 23,58)
C48	25	22,00	4,74	1,00	(20,36; 23,65)
C49	25	20,84	6,24	1,00	(19,20; 22,49)
C50	25	20,67	4,02	1,00	(19,03; 22,32)


Interval nr.
0
10
20
30
40
50
16
18
20
22
24
x


Interval nr.
0
10
20
30
40
50
18
20
22
24
x


Interval nr.
0
10
20
30
40
50
17
18
19
20
21
22
23
x


Interval nr.
0
10
20
30
40
50
16
18
20
22
24
x


Interval nr.
0
10
20
30
40
50
16
17
18
19
20
21
22
23
x

Kapittel 6: Intro til statistisk inferens

Hypotesetesting, setting X

- Nullhypotese H_0 vs alternativ hypotese H_a
- Test-statistikk Z = z, dvs. har observert verdi liten z
- P-verdi p=P(Mer ekstremt utfall enn Z=z - gitt H0 er sann)
- Hvis p er liten, dvs. p < α = valgt signifikansnivå
  så forkaster vi H_0 og aksepterer H_a.

Eksempel: $x_1,\cdots,x_n$ er uavhengige og $N(\mu,\sigma)$

- Hypoteser $H_0: \mu=\mu_0$ vs $H_a: \mu\neq\mu_0$
- Teststatistikk $z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$
- P-verdi $p = P(Z > |z|)$ når Z er N(0,1)
- Forkast hvis p < $\alpha = 0.05$ (f.eks.)

Oppgave 2

Vi er interesserte i å finne ut om det er forskjell mellom kjønnene når det gjelder forventet hvilepuls. Til å undersøke dette skal du bruke ditt modifiserte puls-datasett fra Oppgave 2 i det første obligatoriske oppgavesettet (se den oppgaven for detaljer). Siden vi er interesserte i hvilepulsen, skal du i hele oppgaven konsentrere deg om variabelen Pulse1. Husk å legge inn 1 eller 2 for kjønn i variabelen Sex i linje 93, som representerer deg selv.

a) Bruk kommandoen `Stat - Basic Statistics - Display Descriptive Statistics` til å finne gjennomsnittlig hvilepuls og empirisk standardavvik separat for kvinner og for menn.

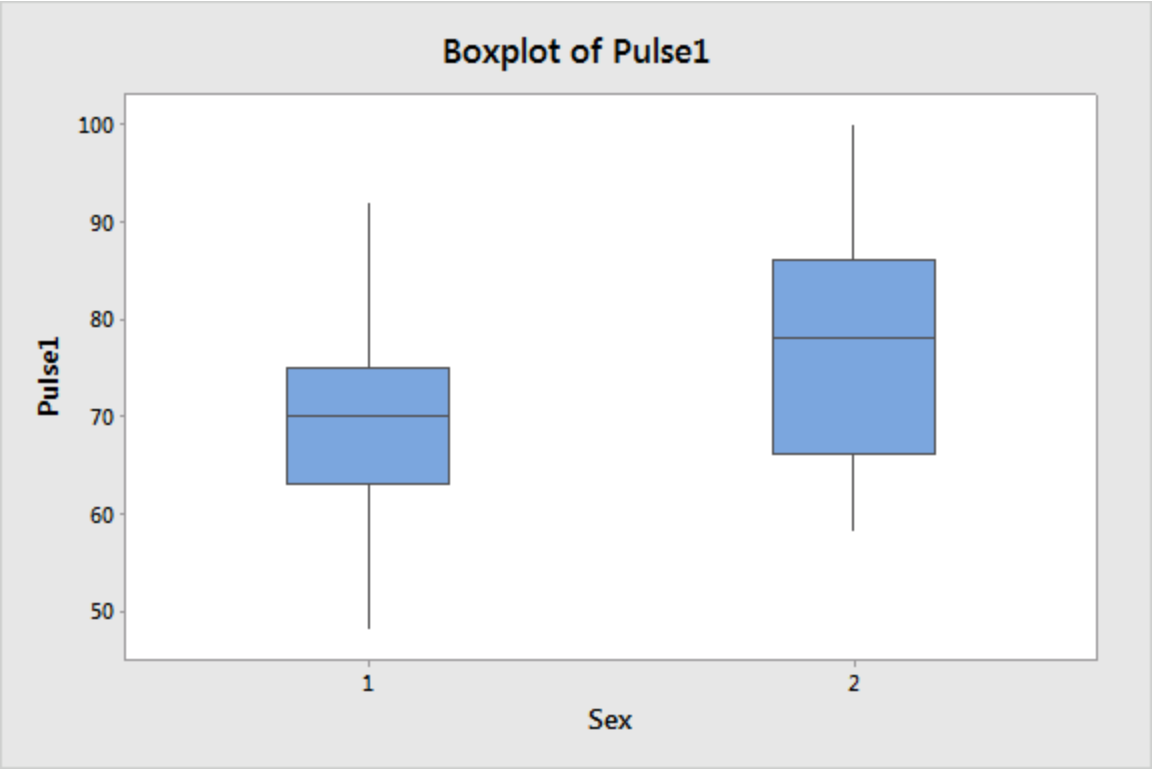
Descriptive Statistics: Pulse1

Variable	Sex	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Pulse1	1	57	0	70,42	1,32	9,95	48,00	63,00	70,00	75,00	92,00
	2	35	0	76,86	1,96	11,62	58,00	66,00	78,00	86,00	100,00

Så for menn er gjennomsnittet lik 70.42 og standardavviket lik 9.95

Og for kvinner blir gjennomsnittet 76.86 og standardavviket lik 11.62

b) Lag boksplott av pulsmålingene for kvinner og for menn i samme figur, og forklar hvilke av størrelsene fra utskriften i punkt a) du kan finne igjen i boksplottene.



Streken midt i boksene angir median.

Øvre og nedre grense for boksene er første og tredje kvartil

Linjene strekker seg til hhv. minimum og maksimum siden det ikke er outliere

c) For å svare på problemstillingen gitt først i oppgaven om at pulsnivået for kvinner er forskjellig fra pulsnivået for menn, vil vi først anta at gjennomsnittet av Pulse1 for menn (Sex=1) faktisk gir den sanne forventning for menn. Anta videre at standardavviket for begge grupper er 10.0 (slag per minutt).

Formuler problemstillingen som et hypotesetestingsproblem (for gruppen av kvinner) med nullhypotese og alternativ hypotese.
Bruk resultatene fra punkt a) til å beregne teststatistikken slik den er gitt på side 383 i læreboka.

$$H_0: \mu = \mu_0 = 70.42 \text{ vs } H_a: \mu \neq \mu_0 = 70.42$$

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{76.86 - 70.42}{10/\sqrt{35}} = 3.81$$

d) Finn P-verdien for testen. Forklar hvordan denne tolkes og hva resultatet av testen betyr.

Fra Tabell finner vi P(Z>3.81)<P(Z>3.49)=0.0002
Dermed blir p-verdien < 2x0.0002 = 0.0004 < 0.05 = vanlig valg av signifikansnivå.

Vi konkluderer at det vi forkaster nullhypotesen og at det er evidens for at hvilepuls hos kvinner er høyere enn blant menn.

e) Gjør testen direkte ved hjelp av kommandoen `Stat - Basic Statistics - 1-Sample z`. Kontroller at resultatene blir de samme som de du fikk i punktene c) og d).

One-Sample Z: Pulse1

```
Test of μ = 70,42 vs ≠ 70,42
The assumed standard deviation = 10

Variable   N   Mean  StDev  SE Mean       95% CI          Z      P
Pulse1    35  76,86  11,62     1,69  (73,54; 80,17)   3,81  0,000
```

Vi ser at Minitab også regner ut teststatistikk z=3.81 hvilket svarer til en liten p-verdi < 0.0005 i samsvar med beregningen i forrige punkt.

f) Lag et konfidensintervall for forventningen til Pulse1 for kvinner. Ved å bruke sammenhengen mellom tester og konfidensintervall, hva blir resultatet av testen da?

95% konfidensintervall for forventet hvilepuls for kvinner gis ved

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}} = 76.86 + 1.96 \times \frac{10}{\sqrt{35}} = (73.54\ ,\ 80.17)$$

der 97.5 persentil i N(0,1) fordelingen = 1.96.

Siden verdien 70.42 ikke ligger inni dette intervallet forkaster vi nullhypotesen

g) Å anta at standardavviket er kjent, er urealistisk. Bruk teorien i avsnitt 7.1 i læreboka til å utføre en test på forventningen for kvinner når du nå ikke forutsetter kjennskap til standardavviket. (Du kan her bruke kommandoen `Stat - Basic Statistics - 1-Sample t` i MINITAB)

Hvor mange frihetsgrader får denne testen?
Hva blir dine konklusjoner nå?

T-statistikken blir $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{76.86-70.42}{11.62/\sqrt{35}} = 3.28$

Under nullhypotesen har denne en t-fordeling med 35-1 frihetsgrader.

Fra Tabell D kan vi konkludere at p-verdien er tilnærmet ligger mellom 2 x 0.0025 = 0.005 og 2 x 0.001 = 0.002, den siste verdien oppgis i Minitab.

Merk også at 70.42 fortsatt ligger utenfor konfidensintervallet.

One-Sample T: Pulse1

```
Test of μ = 70,42 vs ≠ 70,42

Variable   N   Mean  StDev  SE Mean       95% CI          T      P
Pulse1    35  76,86  11,62     1,96  (72,87; 80,85)  3,28  0,002
```

h) Antagelsen vi gjorde om at forventningen til Pulse1 for menn var kjent, er ikke riktig. Det er imidlertid mulig å teste om de to gruppene har forskjellig forventning uten å gjøre antagelser om at den ene er kjent. Teorien bak dette er beskrevet i avsnitt 7.2 i læreboka.

Utfør testen ved hjelp av denne kommandoen både når du antar at standardavvikene er like og når du ikke gjør det. Diskuter resultatene og spesielt forskjeller mellom de to testene.

Two-Sample T-Test and CI: Pulse1; Sex

```
Two-sample T for Pulse1

Sex    N    Mean   StDev   SE Mean
1     57   70,42    9,95       1,3
2     35    76,9    11,6       2,0

Difference = μ (1) - μ (2)
Estimate for difference:  -6,44
95% CI for difference:  (-11,16; -1,71)
T-Test of difference = 0 (vs ≠): T-Value = -2,72  P-Value = 0,008  DF = 63
```

Så det er signifikant forskjell mellom menn og kvinner når usikkerheten i begge gjennomsnitt tas i betrakting, selv om p-verdien er litt større.

```
Difference = μ (1) - μ (2)
Estimate for difference:  -6,44
95% CI for difference:  (-10,96; -1,91)
T-Test of difference = 0 (vs ≠): T-Value = -2,82  P-Value = 0,006  DF = 90
Both use Pooled StDev = 10,6093
```

I testen over er det antatt samme σ for menn og kvinner, mens i utskriften på forrige side var standardavvikene antatt ulike.

Tallmessig er det liten forskjell mellom resultatene, dette har sammenheng med at empiriske standardavvik var nokså like for menn og kvinner.

Vi se imidlertid at testen med lik σ gir flere frihetsgrader.

Kapittel 7: Inferens – i praksis!

- T-fordelingen
- Ett-utvalgs konfidensintervall og hypotesetest basert på t-fordeling
- Parede t-tester
- Sammenligning av to utvalg under antagelse av
  - Ulik varians i utvalgene
  - Lik varians i utvalgene

Kapittel 10: Enkel lineær regresjon i statistisk ramme

- Statistisk modell for enkel lineær regresjon
- Egenskaper ved minste kvadraters estimatorene
- Testing og konfidensintervaller minste kvadraters estimatorene
- Konfidensintervall for forventet verdi og prediksjonsintervall for ny verdi
- Modellsjekk og mulighet for forbedring ved transformasjon

Kapittel 11: Multippel lineær regresjon, flere forklaringsvariable

- Modell, minste kvadraters estimatorer og deres egenskaper
- Testing og konfidensintervall
- Forklart andel av variasjon R^2 generalisert til flere forklaringsvariable