

Oppgave 1: Matbutikk

Småbarnsforeldre trenger i blant å ta seg en øl. Anta at blant personer som handler på en bestemt matbutikk er $1/10$ småbarnsforeldre, og at $1/3$ av småbarnsforeldrene kjøper øl. Blant de som ikke er småbarnsforeldre, men som handler på denne butikken, kjøper $1/6$ øl. Definer begivenhetene « \emptyset : personen kjøper øl» og « S : personen er småbarnsforelder».

[Hint: Det kan være nyttig å bruke sammenhengen $P(A \cap B) = P(A \text{ og } B) = P(B)P(A|B)$.]

- a) Finn sannsynlighetene $P(\emptyset|S)$ og $P(\emptyset \cap S)$.
- b) Hva er sannsynligheten for at en tilfeldig person på denne matbutikken kjøper øl?
- c) Hva er sannsynligheten for at en tilfeldig person på butikken er småbarnsforelder, gitt at personen kjøper øl?
- d) Hva er sannsynligheten for at en tilfeldig person på butikken er småbarnsforelder, gitt at personen *ikke* kjøper øl?

Oppgave 2: Smittsomt virus

Det har brutt ut et virus i samfunnet. Myndighetene ønsker å finne andelen smittede i befolkningen på et tidspunkt. Personer med symptomer, samt helsearbeidere og barnehage- og skoleansatte blir testet, og man finner at 15% tester positivt.

- a) Er 0,15 et godt estimat på andelen smittede i hele befolkningen? Begrunn svaret ditt.
- b) Betegn sannsynligheten for at en tilfeldig utvalgt person fra befolkningen er smittet som p . Helsemyndighetene trekker et tilfeldig utvalg på 1000 personer fra befolkningen og tester. Anta at testen ikke er helt pålitelig, det vil si, erfaring fra tidligere virusutbrudd tilsier at sannsynligheten for «falskt negativt» resultat er 0.02. Det betyr at sannsynligheten for at en syk person tester positivt er 0.98. Du kan anta at sannsynligheten for «falske positive», det vil si at en frisk person tester positivt, er null. Av de 1000 som blir trukket, tester X personer positivt. Diskuter forutsetningene som må være oppfylt for at X skal være binomisk fordelt: $X \sim \text{bin}(1000, 0.98p)$.
- c) I resten av oppgaven kan du anta at $X \sim \text{bin}(1000, 0.98p)$. La den sanne smitteandelen i befolkningen være 0.05. Sett opp et uttrykk for sannsynligheten for at 50 personer i utvalget tester positivt. Det holder å sette opp uttrykket, du trenger ikke regne ut endelig verdi.
- d) Forklar hva følgende kode i R regner ut:

```
> sum(dbinom(0:50, 1000, 0.05*0.98))  
[1] 0.5950827
```

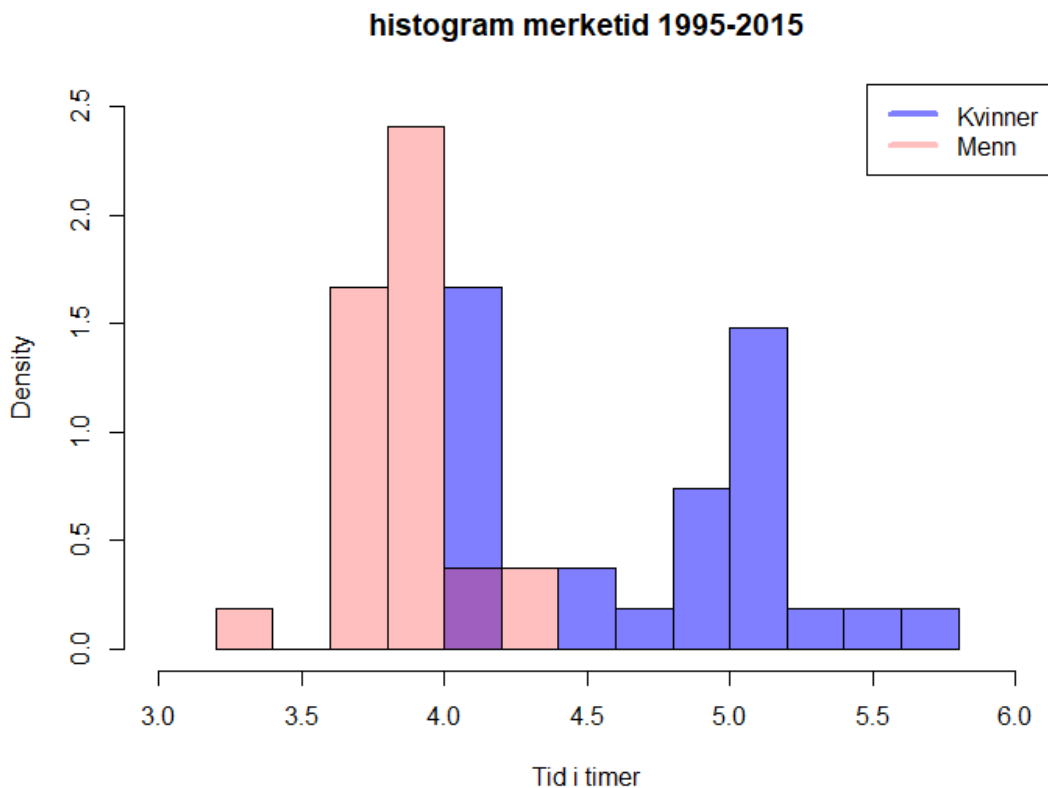
- e) Vis at $\hat{p} = \frac{1}{0.98} \frac{X}{n}$ er en forventningsrett estimator for sannsynligheten, p , for at en tilfeldig trukket person i befolkningen er smittet. Finn standardfeilen til estimatoren.
- f) Det blir trukket et tilfeldig utvalg på 1000 personer fra befolkningen. Av de 1000 personene i utvalget, er det 760 personer som sier ja til å ta testen. Blant disse tester 64 positivt. Finn et estimat for p , og konstruer et 95% konfidensintervall, basert på de 760 personene som tar testen. [Hint: Du kan ta for gitt at X , og dermed \hat{p} , vil være tilnærmet normalfordelt når utvalget er stort. For å finne konfidensintervallet kan du bruke at den estimerte standardfeilen til \hat{p} er 0.0103.]
- g) Gi en tolkning av konfidensintervallet. Ut ifra hvordan utvelgelse og testing ble foretatt, kan du tenke deg noen grunner til at den sanne smitteandelen p ligger utenfor konfidensintervallet?

Oppgave 3: Birken

Birkebeinerrennet er Norges største skirenn. Deltakerne får «merket» dersom de går rennet minst like raskt som merketiden. Merketiden beregnes separat for hver aldersklasse ved at det legges til 25% på gjennomsnittstiden for de fem beste som starter i den respektive aldersklassen.

A)

Under er et histogram for merketider for kvinner 30 år (blå søyler) og menn 30 år (rosa søyler) i Birkebeinerrennet fra 1995-2015. Gi en kort beskrivelse og sammenlikning av fordelingene for kvinner og menn.



B)

Histogrammet er laget basert på data er fra en data frame «merketid», der variabelene «tid_kvinner» og «tid_menn» angir merketid for henholdsvis kvinner og menn. Under er en utskrift fra R med deskriptiv statistikk, der variabelnavnene «tid_kvinner» og «tid_menn» er erstattet med «tid_###» og «tid_###».

```
> deskr_stat=function(x) c(mean=mean(x), sd=sd(x), p=quantile(x, c(.10, .50, .90)))
> deskr_stat(merketid$tid_###)
  mean      sd  p.10%  p.50%  p.90%
4.7025926 0.5073126 4.0700000 4.8800000 5.1720000
> deskr_stat(merketid$tid_###)
  mean      sd  p.10%  p.50%  p.90%
3.8697531 0.1959001 3.6666667 3.8500000 4.0666667
```

- (i) Hvilken linje i utskriften er for kvinner, og hvilken er for menn? Hva er gjennomsnittlig merketid for kvinner? Begrunn kort svaret ditt.
- (ii) Forklar kort hva $sd=0.507$ betyr, og sammenlikn «sd» for de to kjønnene.
- (iii) Kan du si noe om formen på histogrammet ved å sammenlikne «mean» og «p.50%» i utskriften?

c)

Dersom man sammenlikner verdensrekorden for kvinner og menn i forskjellige idretter, finner man ofte at rekordtiden for menn utgjør ca. 90% av rekordtiden for kvinner. Vi ønsker å teste om vi kan forkaste teorien at forventet tid en mann bruker på å gå et skirenn som Birkebeinerrennet utgjør 90% av en kvinnes forventede tid. Det vil si, vi ønsker å teste følgende hypoteser:

$$H_0: \mu_m = 0.90\mu_k \text{ mot } H_1: \mu_m \neq 0.90\mu_k,$$

der μ_m er forventet tid menn 30-34 bruker, og μ_k er forventet tid kvinner 30-34 bruker på Birkebeinerrennet. Utskriften nedenfor angir litt deskriptiv statistikk fra en data frame «tid_birken» basert på observasjoner av tiden kvinner 30-34 år og menn 30-34 år brukte på å gå Birkebeinerrennet i 2019. Du kan tenke på tidene fra 2019 som et utvalg fra alle mulige tider menn og kvinner kan gå et renn som Birkebeinerrennet på – både opp til og med 2019, og i fremtiden. Anta at tidene for menn og kvinner er uavhengige av hverandre.

- (i) Diskuter antakelsen om at tidene for menn og kvinner er uavhengige av hverandre.
- (ii) Test hypotesen over på signifikansnivå 1%. Forklar eventuelle forutsetninger du gjør.

```
> deskr_stat=function(x) c(mean=mean(x), sd=sd(x), n=length(x))
> deskr_stat(tid_birken$tid)
      mean      sd      n
4.740716  1.332650 796.000000
> aggregate(tid ~ kjonn, tid_birken, deskr_stat)
kjonn  time.mean  time.sd  time.n
1   kvinner    5.535799  1.320323 169.000000
2     menn     4.526411  1.253358 627.000000
```

- D) Anta at Birkebeinerrenn-tider for kvinner i aldersgruppen 30-34 år er normalfordelte stokastiske variabler med forventning 5,5 og standardavvik 1,3 timer. Nedenfor er en utskrift med en simulering fra R. Forklar hva tallet 0.00989236 er. Er dette tallet over eller under det teoretiske motstykket? Hva er det teoretiske motstykket?

```
> sim <- replicate(1e4, mean(rnorm(169, mean=5.5, sd=1.3)))
> var(sim)
[1] 0.00989236
```