Sensorveiledning

Litt om karaktersetting:
- For å få bestått ha tre delspørsmål uten massive svakheter. Hvis den ene hoveddelen er svært mangelfullt besvart bør kravet settes noe strengere.
- For å få en C bør en besvarelse svare på oppgaver fra begge hoveddelene og ha et forholdsvis dekkende svar på hoveddelen av oppgavene
- For en A bør en besvarelse ha dekkende besvarelser på alle spørsmålene, men svakheter på noen av deloppgavene kan tolereres.

I. Livet som oppfinner

1) En oppfinner jobber med å finne opp dippedutter. Vi skal her anta at antall nye dippedutter hun finner opp i løpet av et år kan beskrives ved en Poisson-fordeling med rate $\lambda = 6$. Diskuter om det er en fornuftig beskrivelse av en oppfinner.
   Her er vi ute etter forutsetninger for at Poisson-fordelingen kan brukes, dvs. at oppfinnelser kommer med en fast rate og uavhengig av hverandre. Kan gjerne se på det som mange uavhengige hendelser hver med en lav sannsynlighet.
2) Finn sannsynligheten for
   a) At hun gjør mer enn 10 oppfinnelser i løpet av et år
   b) At hun gjør mellom 7 og 9 oppfinnelser i løpet av et år
   c) Ikke finner opp noe en *måned*
   d) At hun finner opp noe i løpet av en måned, gitt at hun ikke fant opp noe i forrige måned
      Ren anvendelse av vedlagt Poisson-tabell. Merk at månedlig rate blir 6/12=1/2. Betinging i d) påvirker ikke svaret pga. uavhengighet.

For de resterende oppgaver vil ikke egenskaper ved Poisson-fordelinger utover forventningen og variansen være nødvendige å kjenne til.

3) Oppfinner-livet gir lite sikker inntekt, så vår oppfinner har slått seg sammen med to andre kolleger i et oppfinner-kollektiv. Vår oppfinner finner som kjent opp dippedutter med en årlig rate $\lambda_1 = 6$, mens de to kollegene finner opp dippedutter med årlige rater $\lambda_2 = 4$ og $\lambda_3 = 7$. Finn forventningen og variansen til årlig oppfinnelses-produksjon fra kollektivet.
   Anvendelse av regneregler for forventning og varians. En god besvarelse bør påpeke at uavhengighet er nødvendig for å beregne variansen. Kollektivet får forventning og varians 6+4+7=17.
4) I kollektivet deles alle inntektene likt på de tre medlemmene. En typisk oppfinnelse er verdt kr 100 000. Definer en stokastisk variabel Y_A som er oppfinnerens inntekt som selvstendig og en stokastisk variabel Y_K som er inntekten som kollektiv-medlem. Finn forventet inntekt med de to ordningene samt variansene til inntekten. Hva er de lønnsmessige fordelene og ulempene for oppfinneren av å være en del av kollektivet?
   Finner $EY_A = 6 \times 100\,000 = 600\,000$ *og* $Var(Y_A) = 6 * 100\,000^2 = 60\,000\,000\,000$
   samt $EY_K = 17 \times \frac{100\,000}{3} = 566\,666$ *og* $Var(Y_K) = 17 \times \left(\frac{100\,000}{3}\right)^2 = 18\,888\,888\,888$
   Ulemper med kollektivet r lavere forventa lønn, men det er også lavere varians dvs mindre variasjon. Forventet nytte og risikoaversjon er ikke dekket i kurset så det siste er ikke trivielt
5) Forklar hvordan du vil lage et R-skript som simulerer andelen av oppfinnelsene fra kollektivet som er produsert av vår oppfinner i løpet av et år. Hvordan kan du bruke dette til å finne sannsynlighet for

at hun i et år har stått for mindre enn 1/3 av oppfinnelsene i kollektivet?
Du kan enten skrive dette ut som R-kode eller pseudo-kode, men vær presis på hvilke trinn du vil gå gjennom
De fleste varianter av beskrivelse av kode bør godtas. Trekke tre vektorer I_1, I_2, I_3 av Poisson-fordelte variable med rater 6,4 og 7, lik lengde. Lage en binær vektor med verdi 1 hviss $I_1 < (I_1 + I_2 + I_3)/3$. Brukke gjennomsnitt av denne vektoren som estimat.

6) Forklar hvordan du vil lage et R-skript som beregner sannsynligheten for at vår oppfinner har funnet opp 3 eller færre dippedutter i løpet av et år, gitt at oppfinnerkollektivet totalt fant opp 10 dippedutter.
Du kan enten skrive dette ut som R-kode eller pseudo-kode, men vær presis på hvilke trinn du vil gå gjennom
Bruke samme kode som over for å trekke I_1, I_2, I_3. Behold elementer hvor $(I_1 + I_2 + I_3) = 10$. Definer for denne sub-vektoren en binær vektor med verdi 1 hviss $I_1 \leq 3$.

7) Hvis kollektivet har mange medlemmer som alle har $\lambda = 6$, hvordan vil fordelingen til inntekten til vår oppfinner være? Som før antar vi at inntektene deles likt på alle medlemmer. Hva er sannsynligheten for at hun tjener mindre enn kr 550 000 et år hvis kollektivet har 100 medlemmer?
Du kan her se bort fra medlemmene med $\lambda \neq 6$.
Trekke på sentralgrenseteoremet for å argumentere for normalfordeling. Med n *medlemmer blir inntekt nå $Y_n \sim N(100\,000 \times \lambda, \frac{100\,000 \times \lambda}{n})$. Trenger å finne $\Pr(Y_{100} < 550\,000)$. Vet at $\frac{Y_{100} - 600\,000}{6\,000} \sim N(0,1)$, så da blir $\Pr(Y_{100} < 550\,000) = Pr(Z < -8.33) \approx 0$.*

8) De siste tre årene har oppfinneren vår produsert henholdsvis 3, 5 og 4 nye oppfinnelser. Hun er bekymret for at hun ikke er like produktiv som før. Gjennomfør en hypotesetest hvor du tester om hun er mindre produktiv på 5% signifikansnivå. Vær klar på hvilket hypotesepar (nullhypotese og alternativ hypotese) du tester og hvilke forutsetninger du gjør.
Følgende kan være nyttig, men vær klar på hvorfor det kan brukes:

```
n<-1e5
x1<-rpois(n,lambda=6)
x2<-rpois(n,lambda=6)
x3<-rpois(n,lambda=6)
q<-x1<=3&x2<=5&x3<=4
mean(q)

[1] 0.01926
```

Setter opp H0: Like god som før eller bedre ($\lambda \geq 6$) mot HA: Dårligere enn før ($\lambda < 6$). Hvis sannsynligheten for å observere det vi gjør eller verre i de tre årene vi har data for er 5% eller lavere kan vi forkaste H0 på 5% nivå. Skriptet regner ut nettopp dette: Tar H0 for gitt og finner andelen av scenariene hvor vi obsrverer produksjon lavere enn (3,5,4). Andelen, som kan tolkes som en sannsynlighet, er 0.019<0.05 så vi kan forkaste H0.

II. Fordommer

For å studere fordommer mot svarte delte en gruppe forskere en utvalg hvite amerikanske respondenter tilfeldig inn i en behandlings- og en kontrollgruppe. De som var i kontrollgruppa fikk lest opp følgende tekst:

Jeg skal nå lese opp tre ting som noen ganger gjør folk irriterte og sinte. Etter at jeg har lest opp alle tre ønsker jeg at du forteller meg HVOR MANGE av dem som irriterte deg. Jeg vil ikke vite hvilke, bare HVOR MANGE.

- *Myndighetene øker bensinavgiftene*
- *Idrettsutøvere får million-lønninger*
- *Store bedrifter forurenser miljøet*

Respondentene i behandlingsgruppa fikk lest opp samme tekst, men fikk i tillegg et ekstra punkt på sin liste:

- *En svart familie flytter inn i nabolaget*

Her er litt deskriptiv statistikk fra en data frame `'race'`, hvor variabelen `'y'` er antall kulepunkter som ble oppgitt å være irriterende, `'treat'` er en binær variabel som er 1 for de i behandlingsgruppa og 0 i kontrollgruppa og `'south'` en tilsvarende indikatorvariabel for å bo i en sørstat.

```
> dstats = function(x) c(mean=mean(x), sd=sd(x), n=length(x))

> stats(race$y)
      mean         sd             n
2.15587349 0.86840818 1328.00000000

> aggregate(y ~ treat, race, dstats)
  treat     y.mean       y.sd          y.n
1 0     2.11901082 0.80424201 647.00000000
2 1     2.19089574 0.92448997 681.00000000

> aggregate(y ~ treat*south, race, dstats)
  treat south     y.mean       y.sd          y.n
1 0     0     2.17391304 0.80376617 483.00000000
2 1     0     2.18714556 0.90711893 529.00000000
3 0     1     1.95731707 0.78600184 164.00000000
4 1     1     2.20394737 0.98558361 152.00000000
```

1) Vil du forvente at noen *uten* rasefordommer svarer annerledes på spørsmålene i behandlingsgruppa enn spørsmålene i kontrollgruppa? Hva med en som har rasefordommer? Forklar svaret ditt.
Bare personer med rasefordommer vil telle med det fjerde spørsmålet. Dvs. i kontrollgruppa vil fordelingen være lik mellom personer med og uten fordommer, i gruppa som får det siste spørsmålet forventer vi at y er 1 høyere for de med fordommer.

2) Beregn andel av hele populasjonen som ser ut til å ha rasefordommer. Beregn standardfeilen og konstruer et 95 % konfidensintervall for estimatet. Gi en kort tolkning av resultatet.
Proportion = 2.19-2.12=0.07. We find that the standard error is $\sqrt{0.80^2/647 + 0.92^2/681} =$

0.05 *and the 95% confidence interval is -0.02, 0.16. The confidence interval is telling us that if we were to run this experiment many times, the 95% interval would include the true sample proportion of those who hold racial prejudice 95% of the time. Note that this is not the same as an interval that contains the true sample proportion with 95% probability.*

3) Gjennomfør en tosidig hypotesetest hvor nullhypotesen er at andelen med rasefordommer i populasjonen er null. Beregn z-verdien og p-verdien. Gjennomfør deretter hypotesetesten med 5 % konfidensnivå. Hvilke forutsetninger må du gjøre for at testen skal være gyldig?

We obtain a z-score of 0.07/ 0.05=1.51 and an associated p -value of 0.13. At the 0.05 level of significance, we fail to reject the null hypothesis of no treatment effect. This means that the difference in the mean number of items selected in the treatment condition is not statistically significant. Substantively, this means that we cannot reject the claim that there is no racial prejudice in the survey sample. In obtaining the p-value, we are assuming that our test statistic (e.g. the difference in means between treated and control) is normally distributed. If the t -statistic was used, we would be assuming that the difference is distributed according to the Student's t distribution with n−1 degrees of freedom.

4) Forsett med nullhypotesen om ingen rasefordommer og 5 % konfidensnivå, men anta nå at den sanne andelen som har fordommer er 0.1. Hva er teststyrken med den utvalgsstørrelsen vi har? Anta at variansen i `'y'` er 1 i den behandla gruppa og 0.8 i kontrollgruppa.

Remember that power=1-prob(type2error), and $\hat{b} = \overline{y_1} - \overline{y_0}$*. We then know that* $se(\hat{b}) = \sqrt{1/681 + 0.8/647} = 0.052$ *so that under H1::* $\hat{b} \sim N(0.1, 0.052^2)$*. We then get that power* $\approx Pr((\hat{b} - .1)/0.052 > 1.96 - .1/052) = .49$

5) Beregn differansen mellom populasjonsandelen som utviser fordommer i sørstatene og i ikke-sørstatene, og konstruer et 95 % konfidensintervall for denne differansen. Tolk differansen og diskuter om den er signifikant. Skill mellom praktisk og statistisk signifikans.

We see that the confidence interval is quite large, ranging from a difference of slightly less than 1% to a difference of nearly 50%. As such, while the difference we detect is statistically significant, it is also relatively imprecise. The true difference could be of a magnitude that is either substantively very large (0.4573039) or very small (0.00949166).

6) Kan vi tolke forskjellen i andelen respondenter som viser rasefordommer mellom sørstatene og ikke-sørstatene kausalt – er den en effekt av å bo i en sørstat? Bruk kontrafaktiske scenarier i forklaringen din.

To estimate the effect of living in the south on racial prejudice for those living in the south, we need to know how their prejudice would have been if they would have lived in the north. This is the counterfactual which is not observed. The estimate here uses the racial prejudice of those in the north as an estimate of the counterfactual racial prejudice of those in the South. This is only possible if those in the north and south are comparable in everything else but their location. This means that south needs to be as-if randomly assigned.