Besvar alle oppgavene. Hver deloppgave har lik vekt.

Oppgave I

En kommune skal bygge ny idrettshall og vurderer to entreprenører, A og B. Begge gir samme pristilbud, men kommunen er bekymret for forsinkelser. For å undersøke hvem av de to som er best samler de inn informasjon om hvilke forsinkelser de to firmaene har hatt når de har bygget for andre kommuner i fylket. De finner følgende hyppigheter:

	A	**B**
Ferdig før tiden	4	7
Ferdig til avtalt tid	12	7
Forsinket 1 måned	2	6
Forsinket 2 måneder	4	4
Forsinket 3 måneder	6	2
Forsinket 4 måneder	0	2

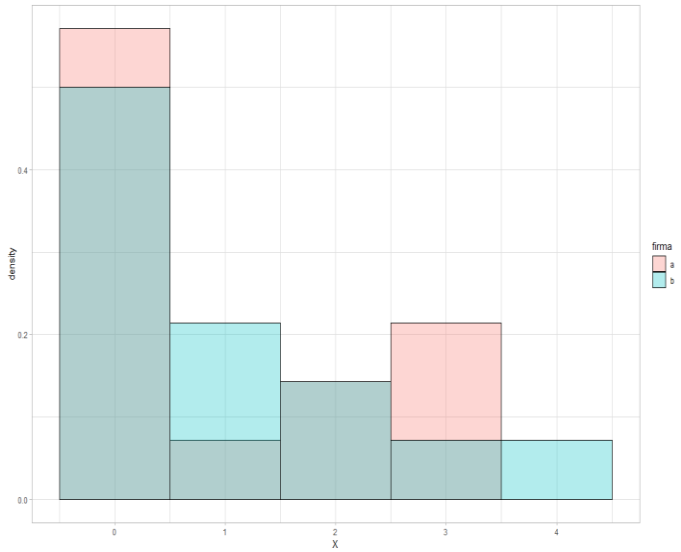
I denne oppgaven skal vi anta at kolonne A gir den sanne fordelingen for entreprenør A, og likeså for kolonne B . Dermed ser vi bort fra eventuelle inferens-problemer.

a) Hva er sannsynligheten for at entreprenør A er ferdig til avtalt tid eller før? Og hvor stor er denne sannsynligheten for entreprenør B?
Pr(A ferdig innen avtalt)=(4+12)/28=4/7=0,57
Pr(B ferdig innen avtalt)=(7+7)/28=1/2

b) Gitt at et firma ble forsinket tre måneder eller mer, hva er sannsynligheten for at det var firma A når kommuner i utgangspunktet velger firma A og B med sannsynlighet $^1/_2$ hver?
Pr(A | Forsinket 3 eller 4)=6/10=3/5=0,6

c) Definer to stokastiske variable X_A og X_B som antall måneder forsinkelse med bedrift A og B. Både ferdig før tiden og til avtalt tid regnes som 0.
Finn fordelingen til de to variablene og tegn et histogram for hver av dem.

	A	B
0	16/28=0,57	14/28=0,5
1	2/28=0,07	6/28=0,21
2	4/28=0,14	4/28=0,14
3	6/28=0,21	2/28=0,07
4	0	2/28=0,07




```
library(tidyverse)
a<-rep(0:4,c(16,2,4,6,0))
b<-rep(0:4,c(14,6,4,2,2))

tibble(a,b) %>%
  gather(key=firma,value = X) %>%
  ggplot(aes(x=X,fill=firma,stat(density))) +
    geom_histogram(bins=5,color='black',alpha=.3,
position="identity") +
    theme_light()
```

d) Finn forventning, varians og standardavvik til X_A og X_B.

$$EX_A = 1, \quad Var(X_A) = 1{,}63, \quad Sd(X_A) = 1{,}28$$
$$EX_B = 1, \quad Var(X_B) = 1{,}63, \quad Sd(X_B) = 1{,}28$$

Kandidater som har feil svar på opg c) skal i minst mulig grad straffes for det på oppgavene under.

e) Beregn $E(X_A|X_A > 0)$ og forklar hva den betyr. Sammenlikn med tilsvarende for drift B.

$$E(X_A|X_A > 0) = 1 \times \frac{2}{12} + 2 \times \frac{4}{12} + 3 \times \frac{6}{12} + 4 \times \frac{0}{12} = 2{,}33$$
$$E(X_B|X_B > 0) = 1 \times \frac{6}{14} + 2 \times \frac{4}{14} + 3 \times \frac{2}{14} + 4 \times \frac{2}{14} = 2$$

En tolkning av den betingede forventningen er forventet forsinkelse gitt at det er forsinkelser. Den blir lengre med bedrift A enn B, som kan sammenliknes med at de har like ubetingede forventninger i opg d).

f) Ved forsinkelse kan kommunen ilegge dagbøter på kr 100 000 per måned. Definer stokastiske variable Y_A og Y_B som beløp betalt i dagbøter. Finn forventning og varians til Y_A og Y_B.

Vi har $Y_i = 100000X_i$. Derfor er $EY_i = 100000EX_i$ og $Var(Y_i) = 100000^2 Var(X_i)$

g) Kommunen vurderer å bestille tre bygg (idrettshall, kulturhus og eldresenter) fra samme entreprenør. De ønsker å finne sannsynligheten for at total forsinkelse for de tre prosjektene blir over 6 måneder. Forklar hvordan du vil lage en simulering som beregner dette for hver av de to bedriftene, og forklar hvert enkelt skritt i algoritmen din.

Du kan svare med å sette opp et R-skript eller en algoritme i pseudo-kode. Du kan se på hvert av de tre byggeprosjektene som uavhengige av hverandre.

Dette kan for A gjøres med

```
mean(replicate(1e4,sum(sample(0:4,3,replace=T,prob=c(.57,.07,.14,.21,0)))>6))
```

Trekker tre verdier fra fordelingen med tilbakelegging; summerer og ser om det er over 6; gjentar 10 000 ganger, finner andel over 6. Tilsvarende for B med ny sannsynlighetsvektor.

Oppgave II

I denne oppgaven skal vi se på lønnsinntekt. Det er en vanlig antakelse at logaritmen til inntekten er normalfordelt. Vi definerer en stokastisk variabel X som den naturlige logaritmen av månedsinntekten i dollar til en familie og antar $X \sim N(\mu, \sigma^2)$.

a) Vi ønsker å estimere den ukjente μ ved hjelp av gjennomsnittet $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$. Dette er en forventningsrett estimator på μ. Forklar hva det betyr at estimatoren er forventningsrett og hvorfor det er en heldig egenskap.
Ved gjentatt trekning/innsamling av utvalg er forventningen til estimatoren lik den sanne μ. Dette innebærer at vi ikke systematisk bommer på parameteren vi forsøker å estimere.

b) Anta at $\mu = 6.6$, $\sigma = 0.7$ og at vi har et utvalg på $n = 100$ familier. Hva er sannsynligheten for å observere $\bar{X} < 6.5$?
Vet at $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ så $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Da er

$$\Pr(\bar{X} < 6.5) = \Pr\left(Z < \frac{6.5 - 6.6}{0.7/\sqrt{100}}\right) = \Pr(Z < -1.43) = 0.076$$

c) En alternativ estimator for parameteren μ er medianen. Hvilke kriterier vil du legge til grunn for å velge hvilken av de to estimatorene gjennomsnitt og median som er best. Forklar hvordan du vil gjennomføre en simulering i R for å undersøke hvilken estimator som er best. Forklar hvert enkelt skritt i algoritmen din.
Du kan enten skrive et R-skript eller forklare algoritmen du vil bruke i pseudo-kode.
Kriterier: Forventningretthet og lav varians (evt. tilsvarende asymptotisk, men ikke viktig). De skal vite at medianen treffer μ. Simulere for å sammenlikne varians. Et R-skript er

```
var(replicate(1e4,mean(rnorm(100,mean=6.6,sd=0.7))))
var(replicate(1e4,median(rnorm(100,mean=6.6,sd=0.7))))
```

Forklaring: Trekker 100 datapunkter fra $N(6.6, 0.7^2)$, beregner gjennomsnitt og median, gjentar 10 000 ganger, og sammenlikner variansen til estimatene.

Videre i denne oppgaven skal vi bruke et utdrag fra den amerikanske NLSY-undersøkelsen[1]. Vi har en data frame `dta` med data på log inntekt (`loginntekt`) og om noen i familien har lånekort på biblioteket (`bibkort`). Noen beregninger på dataene er gjengitt nedenfor:

[1] Se https://www.bls.gov/nls/nlsy79.htm for detaljer om undersøkelsen.

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

> dstats(dta$loginnt)
      mean          sd           n
 6.5864336   0.7216815 920.0000000

> dstats(dta$loginnt[dta$bibkort==0])
      mean          sd           n
 6.5302692   0.7182616 238.0000000
> dstats(dta$loginnt[dta$bibkort==1])
     mean          sd           n
 6.606033   0.722368 682.000000
```

d) Vi skal til å begynne med se bort fra lånekort på biblioteket. Lag et estimat på parameteren μ og sett opp et 95 % konfidensintervall for estimatet. Vær klar på hvilke forutsetninger du gjør. Forklar hva konfidensintervallet betyr.

Estimatet kan leses rett av utskriften, $\bar{X} = 6{,}59$.

Vi kan anta at data er trukket uavhengig fra en populasjon med $X_i \sim N(\mu, \sigma^2)$, evt. trekke på sentralgrenseteoremet. Da følger det at $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, så $Pr\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = \alpha$. Derfor blir $Pr\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \alpha$. Dette gir et konfidensintervall på $6.59 \pm 1.96 \times \frac{0.72}{\sqrt{920}} = [6.54; 6.64]$.

Tolkning: Intervaller konstruert på denne måten vil i gjentatte utvalg dekke den sanne μ med sannsynlighet 0.95.

e) Det er mye som tyder på at personer med bedre sosio-økonomisk bakgrunn gjør det bedre i arbeidsmarkedet enn de med dårligere bakgrunn. Det er blitt foreslått at en indikator på høy sosio-økonomisk status er at noen i familien har lånekort på biblioteket.
Test hypotesen at forventet log inntekter er lik i gruppene med og uten lånekort. Du kan selv velge signifikansnivå. Vær klar på hvilke hypoteser du setter opp og hvilke forutsetninger du gjør.

Vi antar at populasjonen med lånekort følger $X_i \sim N(\mu_{Med}, \sigma^2)$ og de uten $X_i \sim N(\mu_{Uten}, \sigma^2)$. Hvis kandidaten velger å anta ulik varians bør hun følge det opp, men vi har bare vært sporadisk innom det. Mest plausibelt med en en-sidig test, men begge deler greit så lenge kandidaten er klar på hva hun gjør. Da er H_0: $\mu_{Med} \leq \mu_{Uten}$ vs H_1: $\mu_{Med} > \mu_{Uten}$. Viktig å presisere hypoteser. Testobservatoren er $T = \frac{\bar{X}_{Med} - \bar{X}_{Uten}}{s\sqrt{\frac{1}{n_{Med}} + \frac{1}{n_{Uten}}}} \sim t$ med 920-2=918 frihetsgrader. Greit å gå rett på Z-test, men bør forklare hvorfor. En samlet estimator for s er $s^2 = \frac{(n_{Med}-1)s^2_{Med} + (n_{Uten}-1)s^2_{Uten}}{n_{Med}+n_{Uten}-2}$.

På 5% nivå kan vi forkaste H_0 hvis vi observerer $T > 1.645$ (andre nivåer er ok).

Vi finner $s^2 = \frac{681 \times 0.722^2 + 237 \times 0.718^2}{682+238-2} = 0.520 = 0.721^2$. Da blir $T = \frac{6.61-6.53}{0.721\sqrt{\frac{1}{682}+\frac{1}{238}}} = 1.47$. Siden 1.47<1.645 kan vi ikke forkaste nullhypotesen.

NB: Det bør utvises en del toleranse for hvilke formler som brukes til å beregne standardfeil og testobservatorer så lenge intuisjonen er på plass.

f) Beregn styrken på testen i e) hvis den sanne differansen er $\mu_{Med} - \mu_{Uten} = 0.1$. Du kan anta en felles kjent $\sigma = 0.72$. Hvordan tolker du svaret ditt?

Vi ønsker å beregne $\Pr(Forkast\ H_0 | \mu_{Med} - \mu_{Uten} = 0.1) = \Pr(T > 1.645)$. Bruker vi resultatene over får vi $T > 1.645$ hvis $\bar{X}_{Med} - \bar{X}_{Uten} > 1.645 \times s\sqrt{\frac{1}{n_{Med}} + \frac{1}{n_{Uten}}}$. Videre er

$$\Pr(\bar{X}_{Med} - \bar{X}_{Uten} > 0.089) = \Pr\left(\frac{\bar{X}_{Med} - \bar{X}_{Uten} - 0.1}{s\sqrt{\frac{1}{n_{Med}} + \frac{1}{n_{Uten}}}} > \frac{1.645 \times s\sqrt{\frac{1}{n_{Med}} + \frac{1}{n_{Uten}}} - .1}{s\sqrt{\frac{1}{n_{Med}} + \frac{1}{n_{Uten}}}} = 1.645 - \frac{.1}{s\sqrt{\frac{1}{n_{Med}} + \frac{1}{n_{Uten}}}}\right) = \Pr(T > 1.645 - 1.84) = 0.577.$$

Hvis den sanne differansen mellom de med og uten lånekort er 0.1 er sannsynligheten for å forkaste nullhypotesen ved gjentatt trekning av tilsvarende data 0.577.

g) Gå tilbake til tilfellet uten å se på lånekort. Vi ønsker å finne forventet lønn ϕ, det vil si uten å ta logaritmer. Hvis vi fortsatt har X som log inntekten, er en foreslått estimator $\hat{\phi} = \exp(\bar{X})$ hvor exp er eksponensialfunksjonen (dvs. den inverse av logaritmen) og $\bar{X}$ gjennomsnittet av log inntekt. Bruk estimatoren til å estimere forventningen ϕ. Er denne estimatoren forventningsrett og/eller konsistent? Beregn også et 95 % konfidensintervall for ϕ.

Estimatet er $\hat{\phi} = \exp(\bar{X}) = \exp(6.59) = 727.8$. Siden eksponensialfunksjonen ikke er lineær er estimatoren generelt ikke forventningsrett, men siden funksjonen er kontinuerlig er den konsistent. Et konfidensintervall kan vi finne ved å ta eksponensialfunksjonen av intervallet fra opg d), som gir $[\exp(6.54); \exp(6.64)] = [692.3; 765.1]$. Denne prosedyren har vi ikke sett i kurset, så bare de aller beste kandidatene kan forventes å få til dette.