Rapport: Oblig 3a

Gormery K. Wanjiru

Mars 2024

1 Oppgave 1 (5%)

Kapittel 11: oppgave 2 What are the two main purposes of statistical inference mentioned by Bard and Frederick? What is the difference between these two purposes, and how are the purposes related?

Svar

De to hovedformålene med statistisk inferens som nevnes av Bard og Frederick er:

1. **Estimering:** Dette innebærer å bruke data for å utlede verdier av ukjente parametere i en modell. Estimering gir punktestimater eller intervallestimater av parametere, som lar oss forstå egenskapene til populasjonen hvorfra dataene er trukket.

2. **Prediksjon:** Når modellparametrene er estimert, involverer prediksjon å bruke modellen til å forutsi fremtidige observasjoner. Dette lar oss ta informerte beslutninger basert på hva vi forventer vil skje i fremtiden.

Forskjellen mellom disse to formålene ligger i deres anvendelse: estimering er opptatt av den nåværende forståelsen av populasjonen, mens prediksjon er fokusert på fremtidige utfall basert på den forståelsen.

Disse to formålene er imidlertid sammenkoblet. En god estimering av modellparametere er essensiell for å gjøre nøyaktige prediksjoner. Estimering informerer prediksjon, ettersom presisjonen av våre prediksjoner er direkte relatert til nøyaktigheten av våre parameterestimater.

2 Oppgave 2 (10%)

(a) Hva er en parameter?

En parameter er en ukjent tallverdi som beskriver en befolkning. Parametere er sentrale i statistiske modeller fordi de hjelper til med å definere fordelingen av en datamengde.

(b) Hva er en observator?

En observator er en målbar funksjon av dataene, som brukes til å estimere verdien av en parameter.

(c) Hva er en hyperparameter?

En hyperparameter er en parameter i en bayesiansk statistisk modell som ikke er direkte relatert til datamengden, men som påvirker fordelingen av modellparametere.

(d) Hva er en Poisson-prosess?

En Poisson-prosess er en type statistisk prosess som beskriver antall hendelser som skjer i et fast tidsintervall, under forutsetning av at disse hendelsene skjer med en konstant rate og uavhengig av hverandre. Enkel sagt den beskriver hendelser som skjer tilfeldig over tid, hvor antall hendelser i et gitt tidsintervall følger en Poisson-fordeling

(e) Hva er en Bernoulli-prosess?

En Bernoulli-prosess er en sekvens av uavhengige og identisk fordelt binære tilfeldige variabler, hvor hver variabel kan anta verdien 1 med sannsynlighet p og 0 med sannsynlighet $1 - p$.

(f) Hva er en posterior fordeling?

En posterior fordeling er sannsynlighetsfordelingen for en parameter gitt observasjonsdata, basert på en kombinasjon av priorinformasjon og den likelihood som dataene gir.

(g) Hva er en prediktiv fordeling?

En prediktiv fordeling er en fordeling av fremtidige observasjoner gitt eksisterende data, som tar hensyn til usikkerheten rundt parameterestimatene.

Gaussisk

3 Oppgave 3

3.1 1k

For å finne den posterior fordelingen av middelverdien μ med ukjent μ, kjent varians σ^2, og en ikke-informativ prior, bruker jeg Bayes' teorem. Vår prior er spesifisert av hyperparameterne $\kappa_0 = 0$, $\mu_0 = -1$, og $SS_0 = 0$, som indikerer at vi starter uten å foretrekke noen spesiell verdi av μ. Med en kjent $\sigma = 100$, og gitt observasjonsstatistikkene $n = 8$, $\sum x_i = 31832$, beregner vi gjennomsnittet $\bar{x}$:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{31832}{8} = 3979.$$

Siden $\kappa_0 = 0$, blir den prior fordelingen for μ ikke-informativ, og vår posterior fordeling for μ blir derfor basert utelukkende på dataene:

$$\mu|x \sim \mathcal{N}\left(\bar{x}, \frac{\sigma^2}{n}\right).$$

Dette resulterer i en posterior fordeling for μ med et gjennomsnitt på 3979 og en varians på:

$$\frac{\sigma^2}{n} = \frac{100^2}{8} = 1250.$$

Dermed er den posterior fordelingen for μ:

$$\mu|x \sim \mathcal{N}(3979, 1250).$$

Det fordelingen gir oss den oppdaterte troen på μ etter observation av dataene.

4 Oppgave 4

4.1 Oppgave 3: Capacitors

Vi har målt kapasitansen av FR Electronics' minste kondensatorer. Fra et utvalg av 25 målinger har vi fått et gjennomsnitt på $\bar{c} = 49.19\,\mu F$ og en standardavvikelse av $s_c = 2.15\,\mu F$. Vi antar at kapasitansen følger en normalfordeling $\phi(\mu, \sigma)$ med ukjente verdier for μ og σ. Ved bruk av en nøytral prior, ønsker vi å finne sannsynlighetsfordelingene for μ og $\tau = 1/\sigma^2$ (presisjon), og sannsynligheten for at en tilfeldig valgt kondensator av denne typen har en kapasitans på mer enn $50\,\mu F$.

En nøytral prior betyr at vi ikke har noen spesifikk informasjon om forventet verdi av μ eller σ før vi ser dataene. Derfor kan vi bruke en

ikke-informativ prior som Jeffreys' prior, som for en normalfordeling er proporsjonal med $1/\sigma$.

Gitt vårt utvalg, kan vi oppdatere vår tro om μ og τ ved hjelp av de observerte dataene. For en normalfordeling med en ikke-informativ prior, er den posterior fordelingen for μ gitt ved:

$$\mu|\bar{c}, s_c \sim \mathcal{N}\left(\bar{c}, \frac{s_c^2}{n}\right).$$

For våre data har vi at:

$$\mu|\bar{c}, s_c \sim \mathcal{N}\left(49.19, \frac{2.15^2}{25}\right).$$

Den posterior fordelingen for presisjonen τ er mer kompleks, ettersom det krever en gammafordeling. Vi vil ikke gå inn i detaljene her, men vi kan finne sannsynligheten for at en kondensator har mer enn $50\,\mu F$ ved å bruke den kumulative fordelingsfunksjonen (CDF) for en normalfordeling:

$$P(c > 50) = 1 - P(c \leq 50) = 1 - \Phi\left(\frac{50 - \mu}{\sigma}\right),$$

hvor Φ er CDF for en standard normalfordeling. For våre data, blir dette:

$$P(c > 50) = 1 - \Phi\left(\frac{50 - 49.19}{2.15/\sqrt{25}}\right).$$

sannsynligheten er 0.0298, eller 2.98% (beregnet med kode):

```
# Gitt data
mean_c <- 49.19  # Gjennomsnittlig måling
std_c <- 2.15    # Standardavvik
n <- 25          # Antall målinger
threshold <- 50  # Kapasitansverdi for å beregne sannsynligheten

# Beregne standardfeilen for gjennomsnittet
sem_c <- std_c / sqrt(n)

# Beregne z-score for terskelen
z_score <- (threshold - mean_c) / sem_c

# Beregne sannsynligheten for at en tilfeldig valgt kondensator har mer enn 50 µF
prob_gt_50 <- 1 - pnorm(z_score)

prob_gt_50
```

Bernoulli

5 Oppgave 5

5.1 Oppgave 5: Posterior

Gitt prior hyperparametere for en binomial prosess, observerte data og en verdi for p, skal vi finne den posterior fordelingen for π og dens normale tilnærming. Videre vil vi beregne sannsynligheten $P(\pi \leq p)$ både ved eksakt beregning på π og ved bruk av den normale tilnærmingen.

(a) Prior hyperparametere og observasjoner

Gitt prior hyperparametere $a_0 = 2$ og $b_0 = 2$, og observasjoner av 17 positive og 29 negative hendelser, kan vi bruke en Beta-prior for π, som er konsistent med en binomial likelihood. Den posterior fordelingen for π vil da også være en Beta-fordeling:

$$\pi|\textbf{data} \sim \textbf{Beta}(a_0 + x, b_0 + n - x),$$

hvor x er antall suksesser (positive observasjoner) og n er totalt antall forsøk. For våre data blir dette:

$$\pi|\textbf{data} \sim \textbf{Beta}(2 + 17, 2 + 29).$$

Dette forenkles til:

$$\pi|\textbf{data} \sim \textbf{Beta}(19, 31).$$

For den normale tilnærmingen av den posterior fordelingen av π, bruker vi at for store a og b, er Beta(a, b) tilnærmet normalt distribuert med middelverdi $\mu = \frac{a}{a+b}$ og varians $\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}$. For vårt tilfelle blir middelverdien og variansen:

$$\mu = \frac{19}{19 + 31} \quad \textbf{og} \quad \sigma^2 = \frac{19 \cdot 31}{(19 + 31)^2(19 + 31 + 1)}.$$

$P(\pi \leq 0.4)$

Eksakt beregning = 0.6222587524871719
Normaltilnærming = 0.6157194011640728
(beregnet med kode)

6 Oppgave 6

6.1 Oppgave 6: Predictive

Gitt posterior for Bernoulli-parameteren π, og tallene m, s, k, og l. Vi skal finne de prediktive fordelingene for K_{+m} og L_{+s}, og beregne sannsynlighetene $P(K_{+m} \leq k)$ og $P(L_{+s} \leq l)$.

(b) Posterior: $a_1 = 5$, $b_1 = 96$. $m = 20$, $s = 4$, $k = 3$, $l = 12$.

R kode for beregning:

```
# For K+m
pbeta(4 / (20 + 4), 5 + 4, 96 + 20 - 4)

# For L+s
pbeta(12 / (12 + 20), 5 + 20 - 12, 96 + 12)
```

Resultater fra R

- **For del (b), $P(K_{+m} \leq s) \approx 0.999$ og $P(L_{+s} \leq l) \approx 1.000$.**

7 Oppgave 7

7.1 Oppgave 11: Kvalitetsvurdering av Diamanter

Du vurderer kvaliteten på diamantene fra diamantgruvene i et nytt område med spesiell interesse for "Fancy diamonds" av kvalitet IF og VVS. Du estimerer π, andelen av diamanter fra de nye gruvene som passer til en av disse beskrivelsene. Etter å ha evaluert 172 diamanter, har du funnet 19 som er enten IF eller VVS, mens resten er av lavere kvalitetsgrader.

(a) Usikker på valg av nøytral prior

Vi vurderer tre nøytrale priorer for å finne den posterior fordelingen for π for hver av dem:

- **Uniform prior: Beta(1, 1), som representerer en ikke-informativ prior.**
- **Jeffreys' prior: Beta(0.5, 0.5), som er invariant under reparametrisering.**
- **Haldane's prior: Beta(0, 0), som er en uegnet prior med uendelig varians (bruker en grensetilnærming).**

(b) Beregne sannsynligheten for at $\pi > 0.1$ for hver av de tre priorene

Ved hjelp av R, kan vi beregne sannsynligheten for at $\pi > 0.1$ for hver prior. Her er R-koden for disse beregningene:

```
# For Bayesian-analyse av diamantkvalitet, vurderer vi 3 forskjellige
# nøytrale priorer:
# 1. Uniform prior: Beta(1, 1) som er en ikke-informativ prior.
# 2. Jeffreys' prior: Beta(0.5, 0.5) som er invariant under
#    reparametrisering.
# 3. Haldane's prior: Beta(0, 0) som er en uegnet prior med
#    uendelig varians.

# Vi har 19 fancy diamanter av enten IF eller VVS kvalitet ut av 172
# evaluerte.

# Data for analysen
n_diamonds <- 172  # totalt antall evaluerte diamanter
n_fancy <- 19      # antall fancy diamanter funnet (enten IF eller VVS)

# Definer prior parameterne for de tre priorene
priors <- list(
  "Uniform" = list("a" = 1, "b" = 1),
  "Jeffreys" = list("a" = 0.5, "b" = 0.5),
  "Haldane" = list("a" = 0, "b" = 0)  # Merk: Beta(0,0) er ikke en
  # korrekt fordeling
)

# Beregn den posterior fordelingsparameteren og sannsynligheten at \pi > 0.1
# for hver prior
posterior_params <- list()
prob_pi_gt_0_1 <- list()

for (prior_name in names(priors)) {
  params <- priors[[prior_name]]
  # Beregn posterior hyperparametere for \pi
  a_post <- params$a + n_fancy
  b_post <- params$b + n_diamonds - n_fancy

  # For uegnet Haldane's prior, bruker vi en grensetilnærming
  if (a_post == 0 || b_post == 0) {
    # For Beta(0, 0), bruker vi en grensetilnærming ved å sette en
    # veldig liten epsilon
    epsilon <- 1e-6
    a_post <- a_post + epsilon
    b_post <- b_post + epsilon
  }
```

```
  # Lagre de posterior parameterne
  posterior_params[[prior_name]] <- c(a_post, b_post)

  # Beregn sannsynligheten at \pi > 0.1 ved å bruke Beta CDF
  prob_pi_gt_0_1[[prior_name]] <- 1 - pbeta(0.1, a_post, b_post)
}

posterior_params
prob_pi_gt_0_1
```

Resultatene fra R:

- **For den uniforme prioren er den posterior fordelingen Beta(20, 154), og sannsynligheten $P(\pi > 0.1)$ er ca. 0.719.**
- **For Jeffreys' prior er den posterior fordelingen Beta(19.5, 153.5), og sannsynligheten $P(\pi > 0.1)$ er ca. 0.686.**
- **For Haldane's prior er den posterior fordelingen Beta(19, 153), og sannsynligheten $P(\pi > 0.1)$ er ca. 0.650.**

Poisson

8 Oppgave 8

utelatt helt

9 Oppgave 9

utelatt helt

10 Oppgave 10: Tre inferenser

a) Bernoulli (25%)

i. Grunnleggende

For å generere 30 Bernoulli-forsøk med parameter $p = 0.349$, kan vi bruke følgende R-kode:

```
# Genererer 30 uavhengige Bernoulli-forsøk
rbinom(30, 1, 0.349)
# Genererer antall suksesser i 30 Bernoulli-forsøk
rbinom(1, 30, 0.349)
```

Den første linjen utfører 30 uavhengige forsøk, som returnerer et vektor med 30 elementer, hvor hvert element representerer utfallet av hvert forsøk (suksess eller fiasko). Den andre linjen utfører et sett av 30 forsøk og returnerer totalt antall suksesser. Den første metoden gir mer detaljert informasjon om hvert enkelt forsøk, mens den andre gir en summarisk oversikt over antall suksesser.

ii. Observasjons-versjon

Denne koden simulerer 50 grupper av 10 Bernoulli-forsøk hver, beregner suksessraten for hver gruppe, og visualiserer distribusjonen av disse ratene gjennom et histogram. Koden markerer også gjennomsnittet, standardavviket, og den sanne verdien av p.

```
p = 0.349
m = 50
n = 10
z = rbinom(m, n, p) / n
hist(z, breaks = seq(-0.5, n + .5, 1) / n)
m = mean(z)
s = sd(z)
```

```
abline(v = m, col = "green", lwd = 1)
abline(v = m + c(-s, s), col = "pink", lwd = 1)
abline(v = p, col = "blue", lwd = 1)
```

Dette illustrerer hvordan empiriske suksessrater distribuerer seg rundt den sanne parameteren p, med gjennomsnitt og spredning visualisert.

iii. Effekt av større m og n

Økningen i m (antall grupper av forsøk) og n (antall forsøk per gruppe) påvirker resultatene på følgende måter:

- Økning i m gir en jevnere og mer stabil estimasjon av fordelingen av suksessraten siden vi har flere datapunkter å basere estimatene på.
- Økning i n vil typisk føre til en smalere distribusjon rundt den sanne verdien av p, ettersom loven om store tall er at gjennomsnittet av en stor prøve vil nærme seg den forventet verdien.

iv. Inferens-versjon

For inferens bruker vi Jeffreys' nøytrale prior for en Bernoulli-prosess, som er en Beta-distribusjon med $a_0 = 0.5$ og $b_0 = 0.5$. Dette reflekterer en usikkerhet før vi ser data, og er spesielt valgt fordi den er ikke-informativ.

utfører deretter forsøkene for $n = 10, 100, 1000, 10000$ og oppdaterer vår tro basert på observasjonene:

```
# Jeffreys' nøytrale prior
a0 = 0.5
b0 = 0.5

# Simulerer forsøk og oppdaterer troen for forskjellige n
n_values = c(10, 100, 1000, 10000)
for (n in n_values) {
  k = sum(rbinom(n, 1, p))  # Antall suksesser
  l = n - k  # Antall fiaskoer
  a1 = a0 + k
  b1 = b0 + l
  # Tegn posterior sannsynlighetsfordeling for p
  curve(dbeta(x, a1, b1), 0, 1, main = paste("Posterior for n=", n))
}
```

Den demonstrerer hvordan vår kunnskap om p forbedres med flere data, og hvordan posterior distribusjonen blir mer konsentrert rundt den sanne verdien av p som n øker.

b) Poisson (25%)

utelatt for nå

c) Gaussisk (25%)

utelatt for nå