Metadata

Tittel: Tilstrekkelig statistikk?
Sammendrag: Suffisiens, eller tilstrekkelighet, er et sentralt begrep i statistikken. Begrepet motiveres, defineres og drøftes - inklusive tilstrekkelighetsprinsippet. Spesielt begrunnes det på hvilken måte en tilstrekkelig statistikk inneholder all informasjon om modellparameteren. Fisher-Neyman faktoriseringsteoremet bevises og brukes til bevis av suffisiens for konkrete eksempel. Videre vises det at Fisher informasjonen er uendret ved overgang til en tilstrekkelig statistikk og dermed at all informasjon er beholdt også i denne tolkningen. Spesielt er da Cramer-Rao grensen for usikkerhet uendret.
Dette er del av NTNU faget ST1201 Statistiske metoder høsten 2017.
Emneord: Statistikk; Informasjon; Fisher-Neyman faktorisering; Cramer-Rao

Tilstrekkelig statistikk ?

Preludium: Tilstrekkelig statistikk

La $X_1, \ldots, X_{1000}$ være et tilfeldig utvalg fra en fordeling.
Hva betyr påstandene under?

- Empirisk middel $\overline{X}$ er tilstrekkelig for binomisk parameter p
- Empirisk middel $\overline{X}$ er tilstrekkelig for parameter β i eksponensiell fordeling
- Minste og største verdi $(X_{(1)}, X_{(1000)})$ er tilstrekkelig for uniform fordeling på $[\mu - 1, \mu + 1]$
- Geometrisk middel og empirisk middel $(\tilde{X}, \overline{X})$ er tilstrekkelig for gamma parametre (α, β)
- Empirisk middel og varians $(\overline{X}, S^2)$ er tilstrekkelig for parametrene (μ, σ^2) i normalfordelingen

Du skal lære om

- Hva betyr det at T er tilstrekkelig?
- Faktoriseringsteoremet og $f(x \mid \theta) = g(t \mid \theta)h(x)$
- Tilstrekkelighetsprinsippet
- $\mathcal{F}_X = \mathcal{F}_T$
- Eksempel på suffisiens (=tilstrekkelighet)

Suffisient = Tilstrekkelig

- Intuisjon: T er tilstrekkelig dersom den inneholder all informasjon om modellparameteren.
- Definisjon: T er tilstrekkelig dersom $(X \mid T = t)$ har en fordeling som ikke avhenger av modellparameteren.
- Begrunnelse: Dersom vi har observert $T = t = \phi(x)$, så kan vi simulere x' fra $X \mid T = t$. Da er x' og opprinnelig x fra samme fordeling, og inneholder dermed like mye informasjon om modellparameteren.

Faktoriseringsteoremet

- $T = \phi(X)$ er tilstrekkelig hvis og bare hvis $f(x \mid \theta) = g(t; \theta)h(x)$
- Hvis: $f^\theta(x \mid t) = f(x, t \mid \theta)/f(t \mid \theta) = g(t; \theta)h(x)/[\sum_{x'} g(t; \theta)h(x')[\phi(x') = t]]$
- Bare hvis: $f^\theta(x)[\phi(x) = t] = f^\theta(x, t) = f^\theta(x \mid t)f^\theta(t) = h(x)g(t; \theta)$.
- Beviset mer generelt enn for diskrete variable utelates her.

Tilstrekkelighetsprinsippet

Dersom en statistikk er tilstrekkelig, så skal statistisk inferens være basert på denne.

Dette skal gjelde for alle tilstrekkelige statistikker!

Eksempel: X_1 som estimator for μ basert på et tilfeldig utvalg $X_1, X_2, \ldots, X_n$ fra $N(\mu, \sigma^2)$ bryter med tilstrekkelighetsprinsippet. Informasjonen i X_1 er ikke tilstrekkelig. $(\overline{X}, S^2)$ er tilstrekkelig! $\overline{X}$ er derimot en tillatt estimator.

Informasjonen i en suffisient

Dersom en statistikk $T = \phi(X)$ er tilstrekkelig, så gjelder $\mathcal{F}_T = \mathcal{F}_X$

Bevis: $\mathcal{F}_X = \mathrm{E}^\theta(L'_X/L_X)^2 = \mathcal{F}_T$ fordi $L_X = L_T \cdot h(X)$

Observer: Cramer-Rao grensen er dermed uendret dersom T brukes fremfor X i estimering!

Eksempel: La $T = X_1 + \cdots + X_n$ gitt et tilfeldig utvalg fra eksponensialfordeling. Da er $\mathcal{F}_T = \mathcal{F}_{X_1,\ldots,X_n} = n\mathcal{F}_{X_1}$ som forenkler. I tillegg vet vi og at estimatoren for parameteren skal være en funksjon av T i følge tilstrekkelighetsprinsippet!

Eksponensialfordelingen

- La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra eksponensialfordelingen.
- $L = \prod_i e^{-x_i/\beta}/\beta = e^{-t/\beta}/\beta^n$, så $t = x_1 + \cdots + x_n$ er tilstrekkelig.
- Konklusjon: $\overline{X} = T/n$ er en sann estimator med minimal varians lik Cramer-Rao grensen. (Den er i tillegg tillatt i følge tilstrekkelighetsprinsippet fordi T er minimal.)

Uniform fordeling

- La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra $U(0, \theta)$.
- $L = \prod_i [0 < x_i < \theta]/\theta = [0 < x_{(n)} < \theta]/\theta^n$, så største observasjon $t = x_{(n)}$ er tilstrekkelig.
- Konklusjon: $\hat{\theta} = x_{(n)}(n+1)/n$ er en sann estimator med minimal varians. (Den er i tillegg tillatt i følge tilstrekkelighetsprinsippet fordi T er minimal.)
- Observer: Usikkerheten til $\hat{\theta}$ er her mindre enn Cramer-Rao grensen! Suffisiensbegrepet dekker flere tilfeller enn gitt av informasjonsulikheten.

Oppsummering

- En tilstrekkelig statistikk inneholder all informasjon
- Fisher informasjonen er uendret ved bruk av en suffisient og mulig nøyaktighet er uendret
- Faktoriseringsteoremet gir de suffisiente
- Tilstrekkelighetsprinsippet
- Eksempel: normalfordeling, eksponsialfordeling, uniform fordeling, ...

Takk for oppmerksomheten!