SF1901 Sannolikhetsteori och statistik I

Jimmy Olsson

Föreläsning 9
25 november 2016

Idag

Inferensproblemet

Punktskattningar (Kap. 11.3)

Skattning av väntevärde och varians (Kap. 11.4)

Metoder för att ta fram skattningar (Kap. 11.5)

Idag

Inferensproblemet

Punktskattningar (Kap. 11.3)

Skattning av väntevärde och varians (Kap. 11.4)

Metoder för att ta fram skattningar (Kap. 11.5)

Statistisk inferens: data $\Rightarrow$ kunskap

- ”Inference is the problem of turning data into knowledge, where knowledge often is expressed in terms of entities that are not present in the data per se but are present in models that one uses to interpret the data.”

  Committee on the Analysis of Massive Data: *Frontiers in Massive Data Analysis*. The National Academies Press, Washington D.C., 2013, sid. 3.

Exempel: opinionsundersökning

- Ett antal personer, säg 1000 st, väljes på måfå ur en stor population (t.ex. Sveriges befolkning).
- En fråga (rörande t.ex. monarkins vara eller icke vara eller medlemskap i Nato), som skall besvaras med ”ja” eller ”nej”, ställs med resultatet att $x = 350$ personer svarar ”ja”.
- Vi vill uppskatta andelen ”ja”-svarare p i hela populationen.

Exempel: the German tank problem

- ”During World War II, British and U.S. statisticians working for military intelligence were keenly interested in estimating German war production (especially the production of Panzerkampfwagen V, Panther), but could hardly ask the German factories to send them reports. Instead, they based their estimates on the manufacturing serial numbers of captured equipment (especially the tank gearboxes). These numbers ($\{1, 2, 3, \dots\}$) were consecutive and did not vary (as this was rational in terms of maintenance and spare parts). These serial numbers provided a sample that was very small, but reliable.”
- Utmaningen består i att skatta det totala antalet tyska stridsvagnar av denna typ.

Inferensproblemet

- Vi antar att vi har tillgång till uppmätta värden $x_1, x_2, \ldots, x_n$.
- Denna *mätdata* kan ses som utfall av s.v. $X_1, X_2, \ldots, X_n$, vilkas fördelning (diskret eller kontinuerlig) beror av en ev. flerdimensionell *okänd parameter* θ.
- Mängden av möjliga parametrar, *parameterrummet*, betecknas Ω_θ.
- Vi vill skatta θ med hjälp av mätdatan.

Exempel: opinionsundersökning (forts.)

- Statistisk modell: det uppmätta antalet "ja"-sägare $x = 350$ kan ses som ett utfall (observation) av

$$X \in \text{Hyp}(9743087^*, 1000, p),$$

där p är okänt.

- Då $n/N = 1000/9743087 \ll 0.1$ kan vi med gott samvete bortse från det faktum att vi väljer personer utan återläggning och approximera hypergeometriska fördelningen med en binomialfördelning, dvs. anta att $x = 350$ är en observation av

$$X \in \text{Bin}(1000, p).$$

*Befolkningsmängd i Sverige den 30 nov 2014 enligt SCB.

Exempel: the German tank problem (forts.)

- Statistisk modell: varje upphittat serienummer x_i, $i = 1, \ldots, n$, ses, något förenklat, som ett utfall av en s.v. X_i med likformig fördelning över mängden

$$\{1, \ldots, \theta\},$$

där θ är det okända antalet pansarvagnar. De olika X_i:na kan anses vara oberoende.

Idag

Inferensproblemet

Punktskattningar (Kap. 11.3)

Skattning av väntevärde och varians (Kap. 11.4)

Metoder för att ta fram skattningar (Kap. 11.5)

Punktskattning

- Vi vill skatta den okända parametern θ med hjälp av en *funktion av data*. Följande definition är viktig.

Definition

En *punktskattning* av en parameter θ är en funktion θ^* som för varje uppsättning mätdata $x_1, x_2, \ldots, x_n$ ordnar ett värde i Ω_θ. Detta värde betecknas

$$\theta^*_{\text{obs}} = \theta^*(x_1, x_2, \ldots, x_n).$$

Då mätdata ses som utfall av s.v. $X_1, X_2, \ldots, X_n$ är θ^*_{obs} ett utfall — observation — av *stickprovsvariabeln* $\theta^*(X_1, X_2, \ldots, X_n)$. Den senare betecknas ofta θ^* för enkelhets skull.

Punktskattning (forts.)

- Det gäller alltså att θ^*_{obs} är en observation av den s.v. $\theta^*(X_1, X_2, \ldots, X_n) = \theta^*$.
- Om ny mätdata samlas in ses denna som ett nytt utfall av X_1, X_2, $\ldots$, X_n, vilket i sin tur leder till ett nytt utfall θ^*_{obs} av θ^*.
- Skattningens *osäkerhet* beror på hur θ^* varierar kring θ, och att bestämma fördelningen för θ^* är sålunda en viktig (och ibland svår) uppgift.

Väntevärdesriktighet

- Det är givetvis önskvärt att fördelningen för θ^* inte har något systematiskt fel. Detta kan formaliseras med hjälp av följande definition.

Definition

En punktskattning sägs vara *väntevärdesriktig* om

$$\mathbb{E}(\theta^*) = \theta$$

oavsett värdet på den okända parametern $\theta \in \Omega_\theta$.

Konsistens

- Ju fler mätningar (information) vi har tillgång till, desto bättre kan vi förvänta oss att den okända parametern kan skattas.
- För n mätdata betecknar vi nu stickprovsvariabeln med θ_n^*.
- Fördelningen för θ_n^* bör sålunda koncentreras mer och mer kring θ när n ökar. Detta kan formaliseras med hjälp av följande definition.

Definition

Om för varje $\theta \in \Omega_\theta$ och för varje givet (litet) $\varepsilon > 0$,

$$\mathbb{P}(|\theta_n^* - \theta| \geq \varepsilon) \to 0$$

då stickprovsstorleken $n \to \infty$, sägs punktskattningen vara *konsistent*.

Intermezzo: Markovs olikhet

- I samband med konsistensundersökningar är följande enkla olikhet användbar.

Sats (Markovs olikhet)

Låt X vara en icke-negativ s.v. Då gäller för varje fixt $\varepsilon > 0$,

$$\mathbb{P}(X \geq \varepsilon) \leq \frac{1}{\varepsilon}\mathbb{E}(X).$$

- Markovs olikhet implicerar direkt *Tjebysjovs olikhet* (som finns i formelsamlingen).
- Genom att tillämpa Markovs olikhet på X^p, för $p \geq 0$, får man även

$$\mathbb{P}(X \geq \varepsilon) = \mathbb{P}(X^p \geq \varepsilon^p) \leq \frac{1}{\varepsilon^p}\mathbb{E}(X^p).$$

Opinionsundersökning (forts.)

- Då vi vill skatta en andel p är punktskattningen

$$p^*_{\text{obs}} = p^*(x) = \frac{x}{n} = \frac{350}{1000} = 0.35.$$

rimlig.

- Denna skattning är väntevärdesriktig, ty då $X \in \text{Bin}(1000, p)$ gäller att

$$\mathbb{E}(p^*) = \mathbb{E}(p^*(X)) = \mathbb{E}\left(\frac{X}{1000}\right) = \frac{\mathbb{E}(X)}{1000} = \frac{1000p}{1000} = p.$$

Opinionsundersökning (forts.)

- För ett godtyckligt n ges skattningens varians av

$$\mathbb{V}(p_n^*) = \mathbb{V}\left(\frac{X}{n}\right) = \frac{1}{n^2}\mathbb{V}(X) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

- Detta medför att skattningen även är konsistent, ty Markovs olikhet ger för varje fixt $\varepsilon > 0$,

$$\mathbb{P}(|p_n^* - p| \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\mathbb{E}(|p_n^* - p|^2) = \frac{1}{\varepsilon^2}\mathbb{V}(p_n^*) = \frac{p(1-p)}{n\varepsilon^2},$$

  där högerledet går mot noll då $n \to \infty$.
- Alltså, ju fler personer vi frågar, desto mer koncentrerad blir skattningen kring den okända andelen p!

Medelkvadratfel

- Ett annat sätt att beskriva hur koncentrerad θ^* är kring θ är följande.

Definition

Medelkvadratfelet (MSE, Mean Square Error) för en punktskattning är

$$\text{MSE} = \mathbb{E}(\{\theta^* - \theta\}^2).$$

- Man visar enkelt (övning!) att

$$\text{MSE} = \mathbb{V}(\theta^*) + \{\mathbb{E}(\theta^*) - \theta\}^2.$$

- Termen $\mathbb{E}(\theta^*) - \theta$ kallas *systematiskt fel* eller *bias* och är noll om skattningen är väntevärdesriktig. I det senare fallet är alltså $\text{MSE} = \mathbb{V}(\theta^*)$.

Effektivitet

Definition

Om två olika skattningar θ^* och $\hat{\theta}^*$ är väntevärdesriktiga och det gäller att

$$\mathbb{V}(\theta^*) \leq \mathbb{V}(\hat{\theta}^*)$$

för alla $\theta \in \Omega_\theta$ (med sträng olikhet för något θ) sägs θ^* vara *effektivare* on $\hat{\theta}^*$.

- En väntevärdesriktig skattning med liten varians är helt enkelt bättre!

Idag

Inferensproblemet

Punktskattningar (Kap. 11.3)

Skattning av väntevärde och varians (Kap. 11.4)

Metoder för att ta fram skattningar (Kap. 11.5)

Skattning av väntevärde och varians

- Låt mätdata $x_1, x_2, \ldots, x_n$ vara observationer av oberoende s.v. $X_1, X_2, \ldots, X_n$ från någon fördelning med väntevärde μ och varians σ^2, dvs. för alla i gäller att

$$\mathbb{E}(X_i) = \mu,$$
$$\mathbb{V}(X_i) = \sigma^2.$$

- Antag att vi vill skatta μ och σ^2 med hjälp av mätdata $x_1, x_2, \ldots, x_n$.
- Lämpliga skattningar visar sig vara *stickprovsmedelvärdet* resp. *stickprovsvariansen*

$$\mu^*_{\text{obs}} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$
$$(\sigma^2)^*_{\text{obs}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \overset{\text{not.}}{=} s^2.$$

Skattning av väntevärde och varians (forts.)

- Normeringen $1/(n-1)$ istället för $1/n$ i s^2 gör stickprovsvariansen väntevärdesriktig.
- Ovan skattningar visar sig ha goda egenskaper:

Sats

Stickprovsmedelvärdet $\bar{x}$ och stickprovsvariansen s^2 är väntevärdesriktiga och konsistenta skattningar av μ resp. σ^2.

- Notera att dessa skattningar inte kräver kännedom om den gemensamma fördelningen för X_i:na!

Idag

Inferensproblemet

Punktskattningar (Kap. 11.3)

Skattning av väntevärde och varians (Kap. 11.4)

Metoder för att ta fram skattningar (Kap. 11.5)

Maximum-likelihood-metoden

- Hittills har vi konstruerat skattningar *ad hoc*. *Maximum-likelihood-metoden* (ML-metoden) ger ett systematiskt tillvägagångssätt.
- Idé: välj det parametervärde som ger högst sannolikhet för den givna mätdatan!

Definition

Funktionen

$$L(\theta) = \begin{cases} p_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta) & \text{(diskreta fallet)}, \\ f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta) & \text{(kontinuerliga fallet)}, \end{cases}$$

kallas *likelihood-funktionen* (L-funktionen).

- Funktionen $\ln L(\theta)$ kallas *log-likelihood-funktionen*.

Maximum-likelihood-metoden (forts.)

- Notera att $L(\theta)$ beror på den observerade mätdatan.
- I det diskreta fallet är sålunda $L(\theta)$ precis sannolikheten att just mätdatan $x_1, x_2, \ldots, x_n$ erhållas om θ vore modellparameter.
- Det kontinuerliga fallet kan tolkas analogt.
- Som skattning väljer vi sedan det parametervärde som maximerar $L(\theta)$.

Definition

Det värde θ^*_{obs} för vilket $L(\theta)$ antar sitt största värde inom Ω_θ kallas *maximum-likelihood-skattningen* (ML-skattningen) av θ.

- Ofta är det *enklare att maximera log-likelihood-funktionen*, vilken har samma maximum.

R. A. Fisher

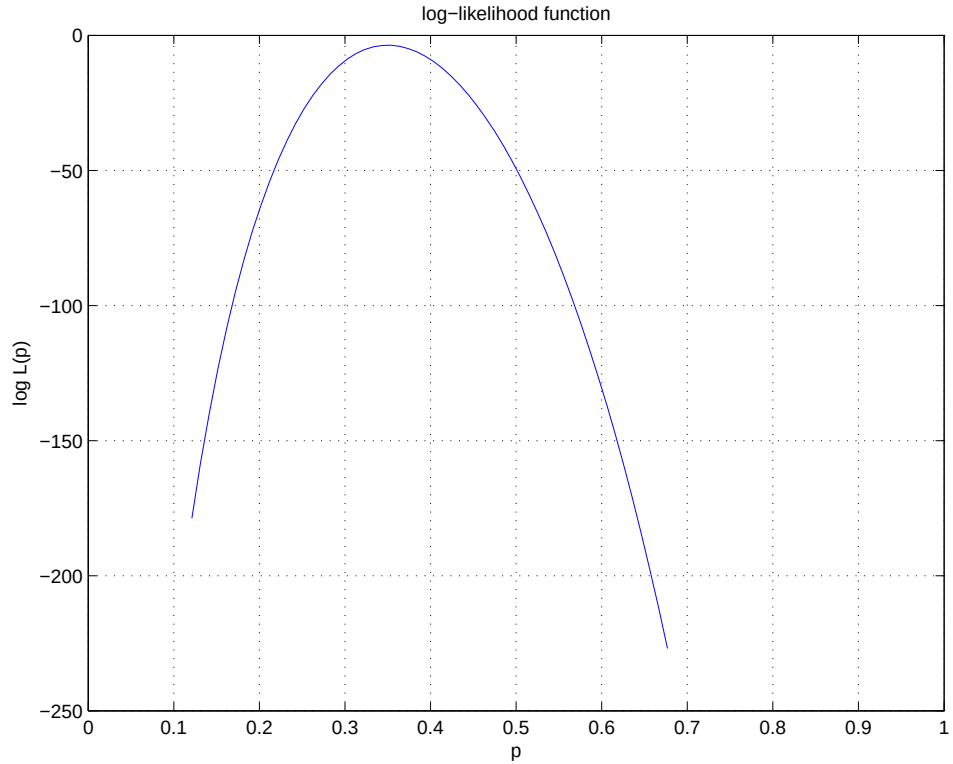
Figur: R. A. Fisher (1890–1962), ”a genius who almost single-handedly created the foundations for modern statistical science” (A. Hald), ”the greatest biologist since Darwin” (R. Dawkins).

Opinionsundersökning (forts.)

- Beräkna ML-skattningen av p!

$$\left[\text{svar: } p^*_{\text{obs}} = \frac{x}{n} = \frac{350}{1000} = 0.35 \right]$$

Opinionsundersökning (forts.)



Figur: Plot av log-likelihood-funktionen.

Exempel: the German tank problem (forts.)

- Statistisk modell: varje upphittat serienummer x_i, $i = 1, \ldots, n$, ses som ett utfall av en s.v. X_i med likformig fördelning över mängden

$$\{1, \ldots, \theta\},$$

där θ är det okända antalet pansarvagnar. De olika X_i:na kan anses vara oberoende.

- Beräkna ML-skattningen av θ!

$$[\text{ svar: } \theta^*_{\text{obs}} = \max\{x_1, \ldots, x_n\}]$$

Exempel: the German tank problem (forts.)

- I detta fall är inte ML-skattningen väntevärdesriktig, ty

$$\mathbb{E}(\theta^*) = \mathbb{E}(\max\{X_1, \ldots, X_n\}) = \ldots = \frac{n(\theta+1)}{n+1},$$

där högerledet dock är nära θ då n är stort.

- Dock kan vi enkelt korrigera skattningen genom att använda

$$\hat{\theta}^*_{\text{obs}} = \theta^*_{\text{obs}} \frac{n+1}{n} - 1;$$

denna korrigerade skattning väntevärdesriktig då

$$\mathbb{E}(\hat{\theta}^*_{\text{obs}}) = \mathbb{E}(\theta^*_{\text{obs}}) \frac{n+1}{n} - 1 = \frac{n(\theta+1)(n+1)}{(n+1)n} - 1 = \theta.$$

Exempel: the German tank problem (forts.)

- "By using the previous estimator, statisticians reportedly estimated that the Germans produced 246 tanks per month between June 1940 and September 1942. At that time, standard intelligence estimates had believed that the number was at around 1400. After the war, the allied captured German production records of the Ministry, which was in charge of Germany's war production, showing that the true number of tanks produced in those three years was 245 per month, almost exactly what the statisticians had calculated, and less than one fifth of what standard intelligence had thought likely, and were more accurate and timely than Germany's own estimates."

Nästa föreläsning

- Mer om punktskattningar,
- konfidensintervall.