SF1901 Sannolikhetsteori och statistik I

Jimmy Olsson

Föreläsning 8
21 november 2016

Idag

Förra gången

Binomialfördelningen—igen (Kap. 7.2)

Hypergeometrisk fördelning (Kap. 7.3)

Possionfördelningen—igen (Kap. 7.4)

Idag

Förra gången

Binomialfördelningen—igen (Kap. 7.2)

Hypergeometrisk fördelning (Kap. 7.3)

Possionfördelningen—igen (Kap. 7.4)

Förra gången: normalfördelningen

- En *normalfördelad* s.v. X har täthetsfunktionen

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}, \quad x \in \mathbb{R},$$

  där μ och $\sigma > 0$ är väntevärde resp. standardavvikelse (kodbeteckning: $X \in \mathrm{N}(\mu, \sigma)$).
- Om $\mu = 0$ och $\sigma = 1$ sägs X vara *standardiserad normalfördelad*. Fördelningsfunktionen finns i tabell. Om $X \in \mathrm{N}(\mu, \sigma)$ så gäller att $(X - \mu)/\sigma \in \mathrm{N}(0, 1)$, vilket är användbart vid beräkningar.
- Linjärkombinationer av normalfördelade s.v. är fortfarande normalfördelade.

Förra gången: centrala gränsvärdessatsen

- Låt X_1, X_2, ... vara en följd av oberoende och likafördelade s.v. med väntevärde μ och standardavvikelse $\sigma > 0$ och sätt

$$Y_n = X_1 + \cdots + X_n.$$

- Notera att $\mathbb{E}(Y_n) = n\mu$ och $\mathbb{V}(Y_n) = n\sigma^2$ ($\Rightarrow \mathbb{D}(X) = \sigma\sqrt{n}$).
- Då säger *centrala gränsvärdessatsen* att fördelningen för Y_n närmar sig $\mathrm{N}(n\mu, \sigma\sqrt{n})$ då n ökar. Vi skriver detta $Y_n \in \mathrm{AsN}(n\mu, \sigma\sqrt{n})$.
- Speciellt gäller $\bar{X} = Y_n/n \in \mathrm{AsN}(\mu, \sigma/\sqrt{n})$.

Idag

Förra gången

Binomialfördelningen—igen (Kap. 7.2)

Hypergeometrisk fördelning (Kap. 7.3)

Possionfördelningen—igen (Kap. 7.4)

Binomialfördelning

- Vi rekapitulerar definitionen:

Definition

Om en s.v. X har sannolikhetsfunktionen

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, 1, 2, \ldots, n\},$$

där $0 < p < 1$, sägs X vara *binomialfördelad* (kodbeteckning: $X \in \mathrm{Bin}(n, p)$).

- Betrakta n st. oberoende försök där sannolikheten att lyckas i varje försök är p.
- Låt X vara antalet lyckade försök av dessa n. Då gäller att $X \in \mathrm{Bin}(n, p)$.

Mellanspel: Bernoulli-fördelning

Definition

Om den s.v. I antar värdena 1 och 0 med sannolikheterna p respektive $1-p$ sägs I vara *Bernoulli-fördelad* (kodbeteckning: Be(p)-fördelning).

- Sannolikhetsfunktionen är

$$p_I(k) = \begin{cases} 1-p & k=0, \\ p & k=1, \\ 0 & \text{annars.} \end{cases}$$

- Dessutom gäller att

$$\mathbb{E}(I) = p,$$
$$\mathbb{V}(I) = p(1-p).$$

Summarepresentation av binomialfördelning

- Betrakta n st. oberoende försök där sannolikheten att lyckas i varje försök är p.
- Låt för $i = 1, \ldots, n$,

$$I_i = \begin{cases} 1 & \text{om försök nr. } i \text{ lyckas,} \\ 0 & \text{annars.} \end{cases}$$

- Då är I_i:na oberoende och $\mathrm{Be}(p)$-fördelade s.v.
- Låt X vara antalet lyckade försök. Då gäller att $X \in \mathrm{Bin}(n, p)$. Dessutom är

$$X = I_1 + I_2 + \cdots + I_n.$$

- Denna *summarepresentation* av binomialfördelade s.v. är ofta användbar vid beräkningar.

Väntevärde och varians för binomialfördelad s.v.

- Låt $X \in \mathrm{Bin}(n, p)$. Med hjälp att summarepresentationen får vi

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^{n} I_i\right) = \sum_{i=1}^{n} \mathbb{E}(I_i) = np,$$

$$\mathbb{V}(X) = \mathbb{V}\left(\sum_{i=1}^{n} I_i\right) \overset{\text{ober.}}{=} \sum_{i=1}^{n} \mathbb{V}(I_i) = np(1-p),$$

$$\mathbb{D}(X) = \sqrt{\mathbb{V}(X)} = \sqrt{np(1-p)}.$$

- Ibland används notationen $q = 1 - p$.

Additionssatsen för binomialfördelade s.v.

- Följande resultat följer direkt av summarepresentationen.

Sats

Om $X \in Bin(n_1, p)$ *och* $Y \in Bin(n_2, p)$ *är oberoende så gäller att*

$$X + Y \in Bin(n_1 + n_2, p).$$

- Notera att p-parametern måste vara samma för båda variablerna för att resultatet skall gälla.

Normalapproximation

- Låt igen $X \in \mathrm{Bin}(n, p)$. Genom att tillämpa centrala gränsvärdessatsen på summarepresentationen av X erhålls

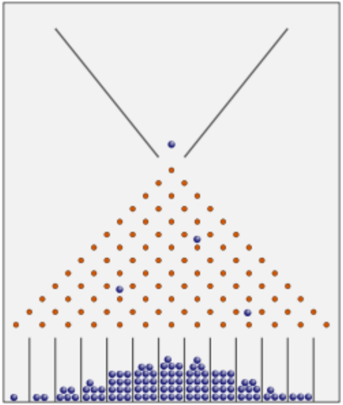
$$X \in \mathrm{AsN}(np, \sqrt{np(1-p)}).$$

- Detta betyder att

$$\begin{aligned}\mathbb{P}(a < X \leq b) &= \mathbb{P}\left(\frac{a-np}{\sqrt{np(1-p)}} < \frac{X-np}{\sqrt{np(1-p)}} \leq \frac{b-np}{\sqrt{np(1-p)}}\right)\\ &\approx \Phi\left(\frac{b-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-np}{\sqrt{np(1-p)}}\right).\end{aligned}$$

- Som tumregel fungerar denna approximation väl sålänge $np(1-p)$ är minst 10.

Galtons bräda



Figur: Galtons bräda simulerar en Bin(n, 0.5)-fördelning.

Exempel: överbokning

- Ett flygbolag tillämpar 8% överbokning på vissa turer. Detta gör man i den trygga förvissningen om att inte alla som bokat plats kommer att utnyttja denna. I själva verket tror man sig veta, att i medeltal var tionde inbokad resenär aldrig utnyttjar sin bokade plats. Anta att ett plan har 120 platser och att man till en tur bokar in 130 passagerare. Vad är sannolikheten för att någon eller några regelrätt inbokade passagerare inte kommer att få någon plats på resan? Lämpliga approximationer och oberoendeantaganden får göras.

  [svar: 0.1521 med MATLAB, 0.1922 med normalapproximation.]

Poisson-approximation

- Om $X \in \text{Bin}(n, p)$ och $Y \in \text{Po}(np)$ så kan man bevisa att
$$| \, p_X(k) - p_Y(k)| \leq np^2.$$
- Detta betyder att en $\text{Bin}(n, p)$-fördelning kan approximeras med en $\text{Po}(np)$-fördelning då p är litet.
- En tumregel är att p skall vara *högst* 0.1 för att approximationen skall vara tillräckligt noggrann.
- Eftersom Poissonfördelningen endast har en parameter är denna fördelning ofta enkel att räkna med.

Exempel: överbokning—igen

- Ett flygbolag tillämpar 8% överbokning på vissa turer. Detta gör man i den trygga förvissningen om att inte alla som bokat plats kommer att utnyttja denna. I själva verket tror man sig veta, att i medeltal var tionde inbokad resenär aldrig utnyttjar sin bokade plats. Anta att ett plan har 120 platser och att man till en tur bokar in 130 passagerare. Vad är sannolikheten för att någon eller några regelrätt inbokade passagerare inte kommer att få någon plats på resan? Lämpliga approximationer och oberoendeantaganden får göras.

[svar: 0.1658 med Poisson-approximation.]

Idag

Förra gången

Binomialfördelningen—igen (Kap. 7.2)

Hypergeometrisk fördelning (Kap. 7.3)

Possionfördelningen—igen (Kap. 7.4)

Två klassiska urnproblem—igen

- Under Föreläsning 2 diskuterades följande exempel.
- I en urna finns s svarta och v vita kulor. Man drar slumpmässigt n kulor ur urnan.
  (a) Antag att kulorna dras *med* återläggning. Vad är sannolikheten att få k vita kulor?

$$\left[\text{svar: } \binom{n}{k}\left(\frac{v}{s+v}\right)^k\left(\frac{s}{s+v}\right)^{n-k}\right]$$

  (b) Antag nu att kulorna istället dras *utan* återläggning. Vad är nu sannolikheten att få k vita kulor?

$$\left[\text{svar: } \frac{\binom{v}{k}\binom{s}{n-k}}{\binom{s+v}{n}}\right]$$

- Notera binomialfördelningen i (a) ($p = v/(s+v)$, $q = s/(s+v)$).

Hypergeometrisk fördelning

- Antag att i en population omfattande N element har Np element en viss egenskap, säg, A. Här är p *relativa frekvensen* A-element.
- I urnexemplet ovan är $A =$ "vit", $N = s + v$ och $Np = v (\Rightarrow p = v/(s + v))$.
- Om n element uttas slumpmässigt *med* återläggning är antalet A-element X bland dessa $\text{Bin}(n, p)$-fördelat.
- Om däremot n element uttas slumpmässigt *utan* återläggning är antalet A-element X bland dessa *hypergeometriskt fördelat*, dvs. har sannolikhetsfunktion

$$p_X(k) = \frac{\binom{Np}{k}\binom{N-Np}{n-k}}{\binom{N}{n}}$$

(kodbeteckning: $X \in \text{Hyp}(N, n, p)$).

Hypergeometrisk fördelning (forts.)

- Man drar slumpmässigt fem kort ur en kortlek utan återläggning. Vad är sannolikheten att två av dessa är hjärter?
- MATLAB-lösning (notera parametriseringen):

```
>> hygepdf(2,52,13,5)

ans =

    0.2743
```

Hypergeometrisk fördelning (forts.)

- Man kan visa att

$$\mathbb{E}(X) = np,$$
$$\mathbb{V}(X) = np(1-p)\frac{N-n}{N-1}.$$

- Väntevärdet ovan är detsamma som för binomialfördelningen. Varianserna skiljer med faktorn $d_n^2 = (N-n)/(N-1)$.
- När kvoten n/N är liten (säg, mindre än 0.1) så kommer det faktum att vi drar utan återläggning att bli försumbart och fördelningen *närmar sig asymptotiskt binomialfördelningen*.
- Detta syns även på att $d_n^2 \to 1$ då $N \to \infty$. Faktorn d_n^2 kallas *korrektionsfaktorn för ändlig population*.

Exempel: bus spotting in Stockholm

- En bussentusiast promenerar en dag omkring i Stockholms innerstad och noterar registreringsnummer på alla bussar han ser, vilket resulterar i 30 st. olika nummer. En vecka senare ger han sig ut igen och påträffar då 25 bussar. Av de senare fanns 5 st. bland dem han noterade en vecka tidigare. Uppskatta hur många bussar det finns i innersta'n!

[svar: ca 150 st.]

- Denna inferensteknik kallas *capture-recapture* och används t.ex. inom ekologi och populationsdynamik.

Idag

Förra gången

Binomialfördelningen—igen (Kap. 7.2)

Hypergeometrisk fördelning (Kap. 7.3)

Possionfördelningen—igen (Kap. 7.4)

Poissonfördelning

- Vi rekapitulerar följande definition.

Definition

Om en s.v. X har sannolikhetsfunktionen

$$p_X(k) = \frac{\mu^k}{k!}e^{-\mu}, \quad k \in \{0, 1, 2, \ldots\},$$

där $\mu > 0$, sägs X vara *Poissonfördelad* (kodbeteckning: $X \in \mathrm{Po}(\mu)$).

- Intressant nog är väntevärdet lika med variansen för denna fördelning:

$$\mathbb{E}(X) = \mathbb{V}(X) = \mu.$$

Summor av oberoende Poissonfördelade s.v.

- Med hjälp av den s.k. *faltningsformeln* för fördelning av summa (se Kap. 4.7) kan man visa följande.

Sats

Om $X \in Po(\mu_1)$ *och* $Y \in Po(\mu_2)$ *är oberoende så gäller att*

$$X + Y \in Po(\mu_1 + \mu_2).$$

- Ovan additionsegenskap leder till att Poissonfördelningen närmar sig normalfördelningen då μ är stort.

Normalapproximation av Poissonfördelningen

- Antag att μ är ett heltal och låt $X \in \text{Po}(\mu)$.
- Enligt ovan sats kan då X skrivas som

$$X = V_1 + V_2 + \ldots + V_\mu,$$

  där $V_1 \in \text{Po}(1), \ldots, V_\mu \in \text{Po}(1)$ är oberoende.
- Centrala gränsvärdessatsen ger nu att X närmar sig normalfördelningen då μ är stort, dvs.

$$X \in \text{AsN}(\mu, \sqrt{\mu}) \quad \text{då } \mu \to \infty.$$

- Resultatet gäller fortfarande då μ inte är ett heltal.
- En tumregel är att μ skall vara *minst* 15 för att approximationen skall vara tillräckligt noggrann.

Nästa föreläsning

- Inferensteori!