



MVE051/MSG810 - Statistik och diskret matematik (LP4 2022-2023)

Chalmers och Göteborgs universitet

Föreläsning 10 – Skattning av proportioner

May 2, 2023

Dagens föreläsning

- Skattning av proportioner
- Konfidensintervall för proportioner
- Hypotestest för proportioner

Skattning av proportioner

Mål

Skatta andelen personer i en population som har en viss egenskap, ex.

- partitillhörighet
- för/emot NATO
- infekterad i Covid
- har/har inte gen

Exempel

För att veta hur stor andel θ av befolkningen som hade röstat på Miljöpartiet om det vore val idag frågar Sifo 1000 slumpvis utvalda personer om vilket parti de skulle röstat på. Av de tillfrågade svarar 47 personer Miljöpartiet. Vi skattar $\hat{\theta} = 47/1000 = 0.047$.

Formell modell för proportioner

Antag att vi har en population där varje person antingen har eller inte har en viss egenskap. Låt θ vara sannolikheten att en slumpvis vald person har egenskapen. Då motsvarar en slumpvis vald individ j i populationen en slumpvariabel

$$X_j = \begin{cases} 1 & \text{om individen har egenskapen} \\ 0 & \text{annars} \end{cases} \sim \text{Bernoulli}(\theta).$$

Om vi drar ett stickprov av storlek n från populationen får vi oberoende slumpvariabler $X_1, X_2, \dots, X_n \sim \text{Bernoulli}(\theta)$. Om vi låter X vara det totala antalet med egenskapen i stickprovet är

$$X = X_1 + \dots + X_n \sim \text{bin}(n, \theta).$$

En skattare $\hat{\theta}$ av θ

Skattare

$$\hat{\theta} = \frac{X_1 + \cdots + X_n}{n}$$

Väntevärdesriktig?

$$\mathbb{E}[\hat{\theta}] =$$

Konsistent?

$$\text{Var}(\hat{\theta}) =$$

Hypotestest för proportioner (små n)

Antag att vi vill testa

$$H_0: \theta = \theta_0 \quad \text{mot} \quad H_1: \theta \neq \theta_0$$

Om nollhypotesen stämmer är antalet individer X med egenskapen i stickprovet

$$X = n\hat{\theta} = n\bar{X} = X_1 + \cdots + X_n \sim \text{bin}(n, \theta_0)$$

så vi ex. beräkna p -värdet

$$p = P(|\bar{X} - \theta_0| > |\hat{\theta} - \theta_0|) = P(|X - n\theta_0| \geq |n\theta_0 - n\hat{\theta}|).$$

Om p -värdet är litet förkastar vi H_0 .

Hypotestest för proportioner (små n)

Exempel

Man vill testa om blåa ögon är lika vanligt som bruna/gröna ögon hos färgblinda personer.

Nollhypotesen är att 50% av alla färgblinda har blåa ögon. Vi vill alltså testa

$$H_0: \theta = 1/2 \quad \text{mot} \quad H_1: \theta \neq 1/2.$$

Man har ett stickprov med $n = 10$ färgblinda personer. Det totala antalet X bland dessa som har blåa ögon en stokastisk variabel X med

$$X \sim \text{bin}(10, \theta).$$

Om H_0 är sann är

$$X \sim \text{bin}(10, 1/2).$$

Hypotestest för proportioner (små n)

Av de 10 personerna i stickprovet hade $x = 9$ personer blåa ögon. Om nollhypotesen stämmer $E[X] = 10 \cdot 1/2 = 5$, och $|9 - 5| = 4$ Vi kan därför beräkna p -värdet

$$\begin{aligned} p &= P(|X - 5| \geq 4) = P(X \geq 9) + P(X \leq 1) \\ &= P(X = 10) + P(X = 9) + P(X = 1) + P(X = 0) \\ &= \binom{10}{10} 0.5^{10} 0.5^0 + \binom{10}{9} 0.5^9 0.5^1 + \binom{10}{1} 0.5^1 0.5^9 + \binom{10}{0} 0.5^0 0.5^{10} \\ &= 0.0215 \end{aligned}$$

Eftersom att p -värdet är mindre än 0.05 så vi kan förkasta nollhypotesen på signifikansnivå 0.05.

Hypotestest för proportioner (stora n)

Antag att man vill testa

$$H_0: \theta = \theta_0 \quad \text{mot} \quad H_1: \theta \neq \theta_0$$

och att man har ett stort stickprov.

Normalapproximation

Vi vet att

$$X = X_1 + \cdots + X_n \sim \text{bin}(n, \theta)$$

Om n är nog stort kan vi använda CLT för att säga att

$$\hat{\theta} = \bar{X} \approx N(\theta, \sqrt{\theta(1-\theta)/n})$$

och

$$\frac{\theta - \hat{\theta}}{\sqrt{\theta(1-\theta)/n}} \approx N(0, 1)$$

Hypotestest för proportioner (stora n)

Antag att man vill testa

$$H_0: \theta = \theta_0 \quad \text{mot} \quad H_1: \theta \neq \theta_0$$

Hypotestest

Om nollhypotesen stämmer och n är stort nog gäller

$$\frac{\theta_0 - \hat{\theta}}{\sqrt{\theta_0(1 - \theta_0)/n}} \approx N(0, 1)$$

Vi förkastar H_0 på signifikantnivå α om

$$\left| \frac{\theta_0 - \hat{\theta}}{\sqrt{\theta_0(1 - \theta_0)/n}} \right| > z_{(1+\alpha)/2}$$

p -värde

$$p = P\left(|Z| > \left| \frac{\theta_0 - \hat{\theta}}{\sqrt{\theta_0(1 - \theta_0)/n}} \right| \right)$$

Konfidensintervall för proportioner (stora n)

Vi vet att

$$X = X_1 + \cdots + X_n \sim \text{bin}(n, \theta)$$

Om n är nog stort kan vi använda CLT för att säga att

$$\frac{\theta - \hat{\theta}}{\sqrt{\theta(1 - \theta)/n}} \approx N(0, 1).$$

Konfidensintervall

För att kunna beräkna konfidensintervall ersätter vi $\theta(1 - \theta)$ med $\hat{\theta}(1 - \hat{\theta})$ och får

$$\frac{\theta - \hat{\theta}}{\sqrt{\hat{\theta}(1 - \hat{\theta})/n}} \approx N(0, 1)$$

Obs: $\sigma(X_j) = \theta(1 - \theta)$ och $s = (1 + \frac{1}{n-1})\hat{\theta}(1 - \hat{\theta})$

Konfidensintervall för proportioner (stora n)

$$\frac{\theta - \hat{\theta}}{\sqrt{\hat{\theta}(1 - \hat{\theta})/n}} \approx N(0, 1)$$

Konfidensintervall

$$CI_{\alpha} = \hat{\theta} \pm z_{(1+\alpha)/2} \sqrt{\hat{\theta}(1 - \hat{\theta})/n}$$

Hypotestest vs konfidensintervall

Eftersom att vi använde $\frac{\theta - \hat{\theta}}{\sqrt{\theta(1 - \theta)/n}}$ och inte $\frac{\theta - \hat{\theta}}{\sqrt{\hat{\theta}(1 - \hat{\theta})/n}}$ när vi gjorde hypotestest får vi ingen direkt motsvarighet mellan konfidensintervall och hypotestest. För att få det måste vi använda $\frac{\theta - \hat{\theta}}{\sqrt{\hat{\theta}(1 - \hat{\theta})/n}}$ även för hypotestestet.

Konfidensintervall för proportioner (stora n)

Exempel

För att veta hur stor andel θ av befolkningen som hade röstat på Miljöpartiet om det vore val idag frågar Sifo 1000 slumpvis utvalda personer om vilket parti de skulle röstat på. Av de tillfrågade svarar 47 personer Miljöpartiet. Vi skattar $\hat{\theta} = 47/1000 = 0.047$.

Ett konfidensintervall på konfidensnivå 0.95 ges av

$$\begin{aligned} CI_{0.95} &= \hat{\theta} \pm z_{(1+\alpha)/2} \sqrt{\hat{\theta}(1 - \hat{\theta})/n} \\ &= 0.047 \pm z_{0.975} \sqrt{0.047(1 - 0.047)/1000} = [0.034, 0.060]. \end{aligned}$$

och ett konfidensintervall på konfidensnivå 0.90 ges av

$$\begin{aligned} CI_{0.90} &= \hat{\theta} \pm z_{(1+\alpha)/2} \sqrt{\hat{\theta}(1 - \hat{\theta})/n} \\ &= 0.047 \pm z_{0.975} \sqrt{0.047(1 - 0.047)/1000} = [0.036, 0.058] \end{aligned}$$

Eftersom att konfidensintervallet innehåller talet 0.04 kan vi inte säga att Miljöpartiet ligger över riksdagsspärren med signifikansnivå 0.025.

För testet $H_0: \theta = 0.04$ mot $H_1: \theta > 0.04$ får vi p -värdet

$$p = P\left(Z > \frac{0.047 - 0.04}{\sqrt{0.04(1 - 0.04)/n}}\right) = 0.129$$

Konfidensintervall för proportioner (små n)

Vi kan inte ta fram konfidensintervall för små n med hjälp av binomialfördelningen eftersom att vi inte kan lösa ut θ ut någon statistika som innehåller både θ och $\hat{\theta}$.

Hur stort stickprov?

Antag att man ska testa

$$H_0: \theta = \theta_0 \quad \text{mot} \quad H_1: \theta \neq \theta_0$$

men att man egentligen tror att $\theta = \theta_1$. Hur stort stickprov behöver man ha för att vara ganska säker på att förkasta nollhypotesen?

Om $\hat{\theta} \approx \theta_1$ så blir konfidensintervallet

$$CI_\alpha = \hat{\theta} \pm z_{(1+\alpha)/2} \sqrt{\hat{\theta}(1-\hat{\theta})/n} \approx \theta_1 \pm z_{(1+\alpha)/2} \sqrt{\theta_1(1-\theta_1)/n}$$

Om konfidensintervallet inte innehåller θ_0 så måste vi ha

$$|\theta_1 - \theta_0| > z_{(1+\alpha)/2} \sqrt{\theta_1(1-\theta_1)/n}$$

Löser vi ut n får vi

$$\sqrt{n} > \left(\frac{z_{(1+\alpha)/2} \sqrt{\theta_1(1-\theta_1)}}{|\theta_1 - \theta_0|} \right)^2 \Leftrightarrow n > \underbrace{\frac{\theta_1(1-\theta_1)}{(\theta_1 - \theta_0)^2}}_{\text{effektstorlek}} \cdot z_{(1+\alpha)/2}^2$$

Om vi använder samma resonemang för en normalfördelning får vi

$$n > \frac{s^2}{(\theta_1 - \theta_0)^2} \cdot z_{(1+\alpha)/2}^2$$

Hur stort stickprov?

Exempel

I december 2020 skrev SvD:

Sedan 1967 har Sifo för SvD varje månad undersökt den svenska väljaropinionen. Aldrig tidigare har stödet för Liberalerna varit så lågt som nu. Bara 3 procent av väljarna skulle rösta på partiet om det vore riksdagsval i dag, enligt decembers väljarbarometer. Den nya lägstanoteringen innebär att Liberalerna ligger signifikant under riksdagsspärren (4%). Sedan förra valet har partiet förlorat nästan hälften av sina väljare.

1. Vad betyder det att *"Liberalerna ligger signifikant under riksdagsspärren (4%)"*
2. Vad för statistika kan Sifo ha använt, och vad har den för fördelning?
3. Hur stort stickprov måste Sifo ha haft för att kunna dra slutsatsen att *"Liberalerna ligger signifikant under riksdagsspärren (4%)"*.

Hur stort stickprov?

Exempel

1. Vad betyder det att "*Liberalerna ligger signifikant under riksdagsspärren (4%)*"

$$\theta < 0.04$$

2. Vad för statistika kan Sifo ha använt, och vad har den för fördelning?

$$Z = \frac{\hat{\theta} - 0.04}{\sqrt{0.04(1-0.04)/n}} \approx N(0, 1).$$

3. Hur stort stickprov måste Sifo ha haft för att kunna dra slutsatsen att "*Liberalerna ligger signifikant under riksdagsspärren (4%)*".

$$z = \frac{0.03 - 0.04}{\sqrt{0.04(1-0.04)/\sqrt{n}}} = -\frac{0.01}{\sqrt{0.04 \cdot 0.94/n}} = -0.052\sqrt{n}$$

$$p = P(Z < z) = \Phi(-0.052\sqrt{n}) \Leftrightarrow n = (\Phi^{-1}(p)/0.052)^2$$

Om $p = 0.05$ får vi $n = 1000.57$. De måste alltså ha frågat minst 1001 personer.

Skilnaden mellan två proportioner (n_1 och n_2 stora)

Om vi har oberoende stickprov från två olika populationer kan vi vilja testa

$$H_0: \theta_1 = \theta_2 \quad \text{mot} \quad H_1: \theta_1 \neq \theta_2$$

Vi använder då att

$$\begin{cases} X_1 \sim \text{bin}(n_1, \theta_1) \Rightarrow \bar{X}_1 \approx N(\theta_1, \sqrt{\theta_1(1-\theta_1)/n_1}) \\ X_2 \sim \text{bin}(n_2, \theta_2) \Rightarrow \bar{X}_2 \approx N(\theta_2, \sqrt{\theta_2(1-\theta_2)/n_2}) \end{cases}$$

$$\Rightarrow \bar{X}_1 - \bar{X}_2 \approx N(\theta_1 - \theta_2, \sqrt{\theta_1(1-\theta_1)/n_1} + \sqrt{\theta_2(1-\theta_2)/n_2})$$