

MVE051/MSG810 2016 Föreläsning 10

Petter Mostad

Chalmers

December 6, 2016

Konfidensintervall för proportion när n är stor

- Om $X_1, \dots, X_n$ är ett stickprov där varje X_i är 1 med sannolikhet p och 0 annars, om $z_{\alpha/2}$ är så att $[-z_{\alpha/2}, z_{\alpha/2}]$ innehåller $100(1 - \alpha)\%$ av sannolikheten i standard normalfördelning, och om vi definerar $\hat{p} = \bar{X}$ och

$$L_1 = \hat{p} - z_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p})/n}$$

$$L_2 = \hat{p} + z_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p})/n}$$

så är $[L_1, L_2]$ ett *approximativt* $100(1 - \alpha)\%$ konfidensintervall för p när n är stor.

- Bevis: När n är stor har vi approximativt att $\bar{X} \sim \text{Normal}(p, \sqrt{p(1 - p)/n})$. När n är stor kan vi även approximera $p(1 - p)$ med $\hat{p}(1 - \hat{p})$.

Skattning av stickprovsstorlek

- ▶ När vi har en formel för ett konfidensintervall där L_1 och L_2 beror på n , så kan vi fråga: Hur stor n behöver jag välja för att konfidensintervallets längd skall bli mindre än något givet tal?
- ▶ Exempel 1: Vi vill skatta en proportion p och har en gissning $\hat{p}$. Hur stort stickprov behövs för att konfidensintervallets längd skall bli max $2d$?

$$n = z_{\alpha/2}^2 \frac{\hat{p}(1 - \hat{p})}{d^2}$$

- ▶ Exempel 2: Vi vill skatta en proportion p och har ingen gissning för $\hat{p}$. Hur stort stickprov behövs för att konfidensintervallets längd skall bli max $2d$?

$$n = z_{\alpha/2}^2 \frac{1}{4d^2}$$

Detta baseras på att vi alltid har $\hat{p}(1 - \hat{p}) \leq 1/4$ (därför att $0 \leq \hat{p} \leq 1$).

Hypotestest för proportion, när n är stor

- ▶ Antag $X_1, \dots, X_n$ är ett stickprov där X_i är 1 med sannolikhet p och annars 0. Antag vi vill jämföra $H_0 : p = p_0$ med $H_1 : p \neq p_0$ för någon känd p_0 .
- ▶ Vi skriver $\hat{p} = \bar{X}$ och väljer som test statistika

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

som har en approximativ standard normalfördelning när H_0 är sann och n är stor.

- ▶ Förkastningsområdet blir alla värden utanför intervallet $[-z_{\alpha/2}, z_{\alpha/2}]$, där $z_{\alpha/2}$ är så att $\Pr [Z > z_{\alpha/2}] = \alpha/2$ när Z har en standard normalfördelning.
- ▶ Motsvarande kan man också göra en-sidiga test där vi använder z_{α} i stället för $z_{\alpha/2}$, och signifikanstest, där vi beräknar ett p-värde.

Konfidensintervall för differens mellan proportioner när n är stor

- Om $X_1, \dots, X_{n_1}$ och $Y_1, \dots, Y_{n_2}$ är stickprov där varje X_i är 1 med sannolikhet p_1 och 0 annars, och motsvarande för Y_i och p_2 , om $z_{\alpha/2}$ är så att $[-z_{\alpha/2}, z_{\alpha/2}]$ innehåller $100(1 - \alpha)\%$ av sannolikheten i standard normalfördelning, och om $\hat{p}_1 = \bar{X}$, $\hat{p}_2 = \bar{Y}$ och

$$L_1 = \hat{p}_1 - \hat{p}_2 - z_{\alpha/2} \sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}$$

$$L_2 = \hat{p}_1 - \hat{p}_2 + z_{\alpha/2} \sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}$$

så är $[L_1, L_2]$ ett *approximativt* $100(1 - \alpha)\%$ konfidensintervall för $p_1 - p_2$ när n är stor.

- Bevis: När n är stor har vi approximativt att $\bar{X} - \bar{Y} \sim \text{Normal}(p_1 - p_2, \sqrt{p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2})$. När n är stor kan vi även approximera $p_1(1 - p_1)$ med $\hat{p}_1(1 - \hat{p}_1)$ och $p_2(1 - p_2)$ med $\hat{p}_2(1 - \hat{p}_2)$.

Att jämföra två stickprov

Vi antar vi har ett stickprov $X_1, \dots, X_{n_1}$ från en fördelning med väntevärde μ_1 och varians σ_1^2 och ett stickprov $Y_1, \dots, Y_{n_2}$ från en fördelning med väntevärde μ_2 och varians σ_2^2 . Vad kan vi säga om $\mu_1 - \mu_2$?

- ▶ Man kan skilja mellan tre tillfällen
 1. σ_1 och σ_2 är kända från *andra källor* än våra data: Se nästa två overheads.
 2. σ_1 och σ_2 är okända, men vi antar att $\sigma_1 = \sigma_2 = \sigma$. Se senare overheads.
 3. σ_1 och σ_2 är okända och kan vara olika. Använd formler från Milton sektion 10.4.
- ▶ MÄRK: För att välja mellan fall 2 och 3 över, så kan man utföra ett hypotestest med $H_0 : \sigma_1 = \sigma_2$, $H_1 : \sigma_1 \neq \sigma_2$, se senare overheads. Testet har dock vissa problem, t.ex. kan det fungera dåligt om fördelningarna för X_i och Y_i inte är normalfördelningar.
- ▶ MÄRK: För fall 1 och 2 över har vi metoder som ger exakta resultat när fördelningarna för X_i och Y_i är normalfördelningar. Generellt gäller att approximationerna kan bli sämre ju mindre normalfördelade data är. En data-analys bör därför alltid starta med att utvärdera om fördelningarna rimligtvis kan vara normalfördelningar.

Konfidensintervall för differens mellan väntevärden, kända σ_1 och σ_2 .

Antag vi har ett stickprov $X_1, \dots, X_{n_1} \sim \text{Normal}(\mu_1, \sigma_1)$ och ett oberoende stickprov $Y_1, \dots, Y_{n_2} \sim \text{Normal}(\mu_2, \sigma_2)$.

- ▶ Om vi antar att σ_1 och σ_2 är kända (alltså fördelningsvarianserna, inte stickprovsvarianserna), om $z_{\alpha/2}$ är så att $[-z_{\alpha/2}, z_{\alpha/2}]$ innehåller $100(1 - \alpha)\%$ av sannolikheten i standard normalfördelning, och om

$$L_1 = \bar{X} - \bar{Y} - z_{\alpha/2} \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

$$L_2 = \bar{X} - \bar{Y} + z_{\alpha/2} \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

så är $[L_1, L_2]$ ett $100(1 - \alpha)\%$ konfidensintervall för $\mu_1 - \mu_2$.

- ▶ Bevis: Baserar sig på att vi kan visa att

$$\bar{X} - \bar{Y} \sim \text{Normal}(\mu_1 - \mu_2, \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}).$$

- ▶ OK approximation om fördelningarna inte är normalfördelningar men n_1, n_2 är stora.

Hypotestest för differens mellan väntevärden, kända σ_1 och σ_2 .

Antag vi har ett stickprov $X_1, \dots, X_{n_1} \sim \text{Normal}(\mu_1, \sigma_1)$ och ett oberoende stickprov $Y_1, \dots, Y_{n_2} \sim \text{Normal}(\mu_2, \sigma_2)$.

- ▶ Antag vi vill jämföra $H_0 : \mu_1 - \mu_2 = d_0$ med $H_1 : \mu_1 - \mu_2 \neq d_0$ för känd d_0 .
- ▶ Vi väljer som test statistika

$$\frac{\bar{X} - \bar{Y} - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

som har en standard normalfördelning när H_0 är sann.

- ▶ Förkastningsområdet blir alla värden utanför intervallet $[-z_{\alpha/2}, z_{\alpha/2}]$, där $z_{\alpha/2}$ är så att $\Pr[Z > z_{\alpha/2}] = \alpha/2$ när Z har en standard normalfördelning.
- ▶ Motsvarande kan man också göra en-sidiga test där vi använder z_α i stället för $z_{\alpha/2}$, och signifikanstest, där vi beräknar ett p-värde.

- ▶ Om X_1 och X_2 är oberoende och X_1 har en chi-kvadrat fördelning med γ_1 frihetsgrader och X_2 har en chi-kvadrat fördelning med γ_2 frihetsgrader så har

$$W = \frac{X_1/\gamma_1}{X_2/\gamma_2}$$

en F fördelning med γ_1 och γ_2 frihetsgrader, vi skriver $W \sim F(\gamma_1, \gamma_2)$.

- ▶ Sedan $\mathbb{E}[X_1] = \gamma_1$ och $\mathbb{E}[X_2] = \gamma_2$ är det naturligt att väntevärdet till W är ungefär 1; mera precist har vi $\mathbb{E}[W] = \frac{\gamma_2}{\gamma_2 - 2}$.
- ▶ F fördelningen liknar en normalfördelning när γ_1 och γ_2 är mycket stora.
- ▶ Det finns tabeller för fördelningsfunktionen för F med olika kombinationer av frihetsgrader.
- ▶ Märk att om $W \sim F(\gamma_1, \gamma_2)$ så får vi $1/W \sim F(\gamma_2, \gamma_1)$.

Hypotestest för identiska varianser

Antag vi har ett stickprov $X_1, \dots, X_{n_1} \sim \text{Normal}(\mu_1, \sigma_1)$ och ett oberoende stickprov $Y_1, \dots, Y_{n_2} \sim \text{Normal}(\mu_2, \sigma_2)$.

- ▶ Antag vi vill jämföra $H_0 : \sigma_1 = \sigma_2$ med $H_1 : \sigma_1 \neq \sigma_2$.
- ▶ Om S_1^2 och S_2^2 är stickprovsvarianserna för X 'erna respektive Y 'erna, och vi väljer test statistikan

$$S_1^2/S_2^2$$

så har denna en $F(n_1 - 1, n_2 - 1)$ fördelning.

- ▶ Förkastningsområdet blir alla värden utanför intervallet $[F_{1-\alpha/2}(n_1 - 1, n_2 - 1), F_{\alpha/2}(n_1 - 1, n_2 - 1)]$, där $F_{\alpha/2}(\gamma_1, \gamma_2)$ är så att $\Pr[W > F_{\alpha/2}(\gamma_1, \gamma_2)] = \alpha/2$ när W har en $F(\gamma_1, \gamma_2)$ -fördelning.
- ▶ Märk: Vid användning av testet väljas alltid vad som är X och vad som är Y så att $S_1^2 > S_2^2$.

Konfidensintervall för differens mellan väntevärden, okänd $\sigma_1 = \sigma_2 = \sigma$.

Antag vi har ett stickprov $X_1, \dots, X_{n_1} \sim \text{Normal}(\mu_1, \sigma)$ och ett oberoende stickprov $Y_1, \dots, Y_{n_2} \sim \text{Normal}(\mu_2, \sigma)$.

- ▶ Om vi låter

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

beteckna det sammansatta ("pooled") variansestimater, om $t_{\alpha/2}$ är så att $[-t_{\alpha/2}, t_{\alpha/2}]$ innehåller 100(1 - α)% av sannolikheten i en T-fördelning med $n_1 + n_2 - 2$ frihetsgrader, och om

$$L_1 = \bar{X} - \bar{Y} - t_{\alpha/2} \sqrt{S_p^2(1/n_1 + 1/n_2)}$$

$$L_2 = \bar{X} - \bar{Y} + t_{\alpha/2} \sqrt{S_p^2(1/n_1 + 1/n_2)}$$

så är $[L_1, L_2]$ ett 100(1 - α)% konfidensintervall för $\mu_1 - \mu_2$.

- ▶ Bevis: Baserar sig på att

$$(\bar{X} - \bar{Y} - (\mu_1 - \mu_2)) / \sqrt{S_p^2(1/n_1 + 1/n_2)} \sim T(n_1 + n_2 - 2).$$

Hypotestest för differens mellan väntevärden, okänd $\sigma_1 = \sigma_2 = \sigma$.

Antag vi har ett stickprov $X_1, \dots, X_{n_1} \sim \text{Normal}(\mu_1, \sigma)$ och ett oberoende stickprov $Y_1, \dots, Y_{n_2} \sim \text{Normal}(\mu_2, \sigma)$.

- ▶ Antag vi vill jämföra $H_0 : \mu_1 - \mu_2 = d_0$ med $H_1 : \mu_1 - \mu_2 \neq d_0$ för någon känd d_0 .
- ▶ Vi väljer som test statistika

$$\frac{\bar{X} - \bar{Y} - d_0}{\sqrt{S_p^2(1/n_1 + 1/n_2)}}$$

som har en $T(n_1 + n_2 - 2)$ fördelning när H_0 är sann.

- ▶ Förkastningsområdet blir alla värden utanför intervallet $[-t_{\alpha/2}, t_{\alpha/2}]$, där $t_{\alpha/2}$ är så att $\Pr[T > t_{\alpha/2}] = \alpha/2$ när T har en $T(n_1 + n_2 - 2)$ fördelning.
- ▶ Motsvarande kan man också göra en-sidiga test där vi använder t_α i stället för $t_{\alpha/2}$, och signifikanstest, där vi beräknar ett p-värde.