Föreläsning 11: Mer om jämförelser och inferens

Matematisk statistik

David Bolin
Chalmers University of Technology
Maj 12, 2014

Oberoende stickprov

Vi antar att vi har två oberoende stickprov

- n_1 observationer $X_{11}, X_{12}, \ldots, X_{1n_1}$ från $\mathrm{N}(\mu_1, \sigma_1^2)$.
- n_2 observationer $X_{21}, X_{22}, \ldots, X_{2n_2}$ från $\mathrm{N}(\mu_2, \sigma_2^2)$.

Den intressanta parametern är nu $\theta = \mu_1 - \mu_2$ som vi skattar med $\theta^* = \bar{X}_1 - \bar{X}_2$ och vi vill nu bilda ett konfidensintervall för θ samt testa hypotesen

$$H_0 : \theta = 0,$$

vilket är samma sak som att testa $\mu_1 = \mu_2$, mot

$$H_1 : \theta \neq 0$$

eller någon ensidig hypotes, $\theta > 0$ (som motsvarar $\mu_1 > \mu_2$) eller $\theta < 0$ (som motsvarar $\mu_1 < \mu_2$). Vi skiljer på tre fall:

1. σ_1 och σ_2 är kända.
2. $\sigma_1 = \sigma_2 = \sigma$ där σ är okänd.
3. σ_2 och σ_2 är okända och ej säkert lika.

Kända σ_1 och σ_2

Om σ_1 och σ_2 är kända gäller att

$$\mathsf{V}(\theta^*) = \mathsf{V}(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Vi kan nu direkt använda detta för att bilda ett konfidensintervall för θ

$$I_\theta = \left(\theta^* \pm z_{\alpha/2}\sqrt{\mathsf{V}(\theta^*)}\right) = \left(\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Vill vi göra ett hypotestest kan vi använda detta konfidensintervall eller använda att

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\mathsf{V}(\theta^*)}}$$

är $\mathsf{N}(0,1)$-fördelad. Det vill säga, med denna teststorhet beräknar vi p-värdet som $p = 2(1 - \Phi(|T_{obs}|))$, precis som för ett vanligt normalbaserat test. Ensidiga test görs analogt.

Poolad skattning av en gemensam varians

Om vi kan anta att varianserna är lika kan vi vinna en del precision på att använda data från båda stickproven för att skatta den gemensamma variansen. En skattning när vi använder data från flera stickprov för att skatta variansen brukar kallas för "pooled estimate" på engelska.

Sats

En väntevärdesriktig skattning av ett gemensamt σ^2 från k olika normalfördelningar $\mathrm{N}(\mu_j, \sigma^2)$ fås genom den sammanvägda skattningen

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_k-1)s_k^2}{(n_1-1) + (n_2-1) + \cdots + (n_k-1)}.$$

Dessutom gäller att $(N-k)S_p^2/\sigma^2$ är $\chi^2(N-k)$-fördelad, där $N = \sum_{i=1}^{k} n_i$.

$\sigma_1 = \sigma_2 = \sigma$ där σ är okänd

Enligt satsen får vi att den sammanvägda skattningen av σ^2 ges av

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

och vi skattar variansen av θ^* med $s_p^2(1/n_1 + 1/n_2)$.

Detta ger att

$$T = \frac{\bar{X}_1 - \bar{X}_2 - \theta}{s_p\sqrt{1/n_1 + 1/n_2}}$$

är $t(n_1 + n_2 - 2)$-fördelad.

Vi kan nu bilda konfidensintervall och göra hypotestest för θ som tidigare.

Okända σ_1 och σ_2 där $\sigma_1 \neq \sigma_2$

Att testa om två väntevärden är lika i fallet med olika och okända varianser visar sig vara komplicerat.
Följande sats presenterar en approximativ metod.

Sats

Om σ_1 och σ_2 är okända och $\sigma_1 \neq \sigma_2$ så gäller att

$$T = \frac{\bar{X}_1 - \bar{X}_2 - \theta}{\sqrt{s_1/n_1 + s_2/n_2}}$$

är approximativt $t(f)$-fördelad där

$$f = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Vi kan nu bilda konfidensintervall och utföra hypotestest på samma sätt som tidigare.

Jämförelse av variansen för oberoende stickprov

För att jämföra de två varianserna inför vi teststorheten

$$T = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

Vi vet sedan tidigare att $(n_i - 1)s_i^2/\sigma_i^2$ är $\chi^2(n_i - 1)$-fördelad om observationerna är normalfördelade, och vi inför nu F-fördelningen som beskriver hur kvoten av χ^2-variabler beter sig.

Definition

Om $V_1 \sim \chi^2(f_1)$ och $V_2 \sim \chi^2(f_2)$ är oberoende så gäller att

$$\frac{V_1/f_1}{V_2/f_2}$$

är $F(f_1, f_2)$-fördelad ("F-fördelad med f_1 och f_2 frihetsgrader")

Jämfördelse av variansen (forts.)

Enligt definitionen har vi att T är $F(n_1 - 1, n_2 - 1)$-fördelad. Vi låter $F_\alpha(f_1, f_2)$ beteckna α-kvantilen i F-fördelningen och får att ett konfidensintervall för σ_1^2/σ_2^2 ges av

$$I_{\sigma_1^2/\sigma_2^2} = \left[\frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1 - 1, n_2 - 1)}, \frac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1 - 1, n_2 - 1)} \right]$$

För hypotestest av $\sigma_1^2/\sigma_2^2 = 1$ bildar vi teststorheten

$$T = s_1^2/s_2^2$$

som under H_0 är $F(n_1 - 1, n_2 - 1)$-fördelad.

Stickprov i par

En annan vanlig situation är att mätningarna uppkommer i par, till exempel om

- Man vill studera hur mycket rökare går upp i vikt när de slutar röka. Man mäter då vikten före och efter för varje person före och efter den slutar röka och jämförelsen sker för varje person.
- Man vill studera systematiska skillnader mellan två mätmetoder och använder varje metod på var och ett av ett antal prover och jämför de två metoderna för varje prov.

Modellen vi nu ansätter är att vi har n observationer i två stickprov

$$X_1, X_2, \ldots, X_n \qquad\qquad Y_1, Y_2, \ldots, Y_n$$

För varje mätning bildar vi differensen som antas vara normalfördelad:

$$D_i = X_i - Y_i \sim \mathsf{N}(\Delta, \sigma^2)$$

Vi vill nu testa om $\Delta = 0$, vilket görs som vanligt för normalfördelade mätningar.

För och nackdelar med stickprov i par

Ofta är det mer effektivt att använda stickprov i par än oberoende stickprov, särskilt om variationen mellan mätningarna är stor. Vi kan dela upp variationen i D_i som $\sigma^2 = \sigma_0^2 + \sigma_\Delta^2$ där σ_0^2 beskriver variationen mellan objekten. Vid stickprov i par har vi

$$\mathsf{V}(\bar{D}) = 2\sigma_\Delta^2/n$$

medan om vi antar modellen oberoende stickprov så har vi

$$\mathsf{V}(\bar{X} - \bar{Y}) = 2(\sigma_0^2 + \sigma_\Delta^2)/n.$$

Alltså vinner vi något på analysen om $\sigma_0 > 0$.
Det vi förlorar på att använda modellen är att vi minskar antal frihetsgrader i skattningen av variansen. För oberoende stickprov använder vi $2n - 2$ frihetsgrader vid test, medan vi använder $n - 1$ frihetsgrader vid stickprov i par. Alltså ska man vara försiktig med att använda modellen då vi har få observationer eftersom variansskattningen då blir osäkrare.