Föreläsning 11: Mer om jämförelser och inferens

Matematisk statistik

David Bolin
Chalmers University of Technology
Oktober 5, 2015

Oberoende stickprov

Vi antar att vi har två oberoende stickprov

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

Vi vill nu testa om vi kan anta att $\mu_1 = \mu_2$.

Inför $\theta = \mu_1 - \mu_2$ som vi skattar med $\theta^* = \bar{X}_1 - \bar{X}_2$. Testa

$$H_0 : \theta = 0,$$
$$H_1 : \theta \neq 0$$

Alternativt en ensidig mothypotes, $\theta > 0$ eller $\theta < 0$.

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.

Fall 1: Kända σ_1 och σ_2

Om σ_1 och σ_2 är kända gäller att

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

För hypotestest använder vi att under H_0 är

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\mathsf{D}(\theta^*)} \sim \mathsf{N}(0, 1)$$

Så p-värdet ges av $p = 2(1 - \Phi(|T_{obs}|))$.
Ett konfidensintervall för $\theta = \mu_1 - \mu_2$ ges av

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

Fall 2: $\sigma_1 = \sigma_2 = \sigma$ där σ är okänd

Sats: Poolad variansskattning

För k normalfördelade stickprov $\mathrm{N}(\mu_j, \sigma^2), j = 1, \ldots, k$ fås en väntevärdesriktig skattning av σ^2 som

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

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

Med $d(\theta^*) = s_p\sqrt{1/n_1 + 1/n_2}$ har vi att under H_0 är

$$T = \frac{\bar{X}_1 - \bar{X}_2}{d(\theta^*)} \sim t(n_1 + n_2 - 2)$$

Konfidensintervall ges nu av $I_\theta = \big(\theta^* \pm t_{\alpha/2}(n_1 + n_2 - 2)d(\theta^*)\big)$.

Fall 3: $\sigma_1 \neq \sigma_2$ okända

Sats

Baserat på två normalfördelade stickprov gäller att

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

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: $I_\theta = \left(\theta^* \pm t_{\alpha/2}(f)\sqrt{s_1/n_1 + s_2/n_2}\right)$.

Jämförelse av variansen för oberoende stickprov

Om det inte är känt från försöksuppställningen kan vi behöva testa om vi kan anta $\sigma_1 = \sigma_2$.
Vi undersöker detta genom att utföra hypotestestet

$$H_0 : \frac{\sigma_1}{\sigma_2} = 1$$
$$H_1 : \frac{\sigma_1}{\sigma_2} \neq 1$$

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.)

Låt $F_\alpha(f_1, f_2)$ beteckna α-kvantilen i F-fördelningen. 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. För att underlätta beräkning av p-värdet, välj s_1 som den större av det två stickprovsvarianserna.

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.

Ett exempel

Man vill veta om en ny vetesort ger större skörd än den existerande sorten. Man väljer ut sex åkrar som skiljer sig i bördighet och klimat och delar in vardera åker i två delar där varsin sort odlas.

Åker nr	1	2	3	4	5	6
Skörd sort 1, kg/ha	7529	8913	6534	6503	6896	8023
Skörd sort 2, kg/ha	7239	8726	6129	6351	6644	7711
Skillnad D_i	290	187	405	152	252	312

Vi testar $H_0 : \Delta = 0$ mot $H_1 : \Delta \neq 0$.

Vi har $\bar{D} = 266.3$ och $s_D = 91$ och får

$$I_\Delta = (\bar{D} \pm t_{0.025}(5)s_D/\sqrt{6}) = (171, 362)$$

Eftersom $0 \notin I_\Delta$ kan vi förkasta H_0.

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.