

Föreläsning 11: Mer om jämförelser och inferens

Matematisk statistik

David Bolin
Chalmers University of Technology
Maj 18, 2015



Oberoende stickprov

Vi antar att vi har två oberoende stickprov

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

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

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

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

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

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

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

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

är $N(0, 1)$ -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 $N(\mu_j, \sigma^2)$ 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

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

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

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

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

Om det inte är känt från försöksupställningen kan vi behöva testa om vi kan anta $\sigma_1 = \sigma_2$

Ibland kan det också vara intressant i sig att jämföra varianserna, till exempel om

- Man vill jämföra två produktionsmetoders förmåga att skapa produkter av jämn kvalitet. Man tillverkar ett antal satser med de två metoderna och jämför varianserna.
- Man har tagit fram en ny borste i ett limaggregat och vill veta om den ger jämnare limfogar. Man mäter därför några limfogar gjorda med den gamla borsten och några med den nya borsten och jämför varianserna.

Jämförelse av variansen för oberoende stickprov

För att jämföra de två varianserna inför vi teststorheten

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

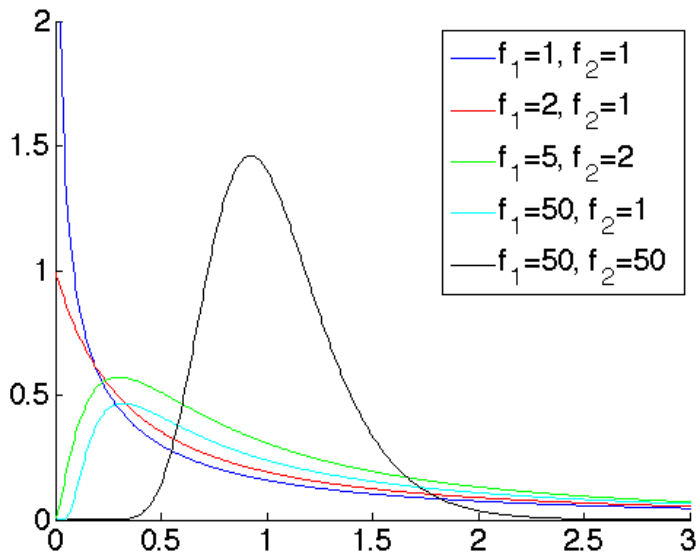
Vi vet sedan tidigare att $(n_i - 1)s_i^2/\sigma_i^2$ är $\chi^2(n_i - 1)$ -fördelad om observationerna är normalfördelade, och vi inför nu F-fördelningen som beskriver hur kvoten av χ^2 -variabler beter sig.

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")

Täthetsfunktioner för $F(f_1, f_2)$ -fördelningen

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. För att underlätta beräkning av p-värdet, välj s_1 som den större av de 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

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

För varje mätning bildar vi differensen som antas vara normalfördelad:

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

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

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

medan om vi antar modellen oberoende stickprov så har vi

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

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.

Binomialfördelningen

Låt $X \sim \text{Bin}(n, p)$. Vi observerar x och vill ha information om p .

$p^* = x/n$ är en väntevärdesriktig skattning av p med varians $V(p^*) = p(1 - p)/n$.

Centrala gränsvärdessatsen som säger att X är approximativt $N(np, np(1 - p))$ -fördelad om n är stor. En tumregel brukar vara att om $np(1 - p) > 10$ så fungerar normalapproximation bra. Vi har då att p^* är approximativt $N(p, p(1 - p)/n)$ -fördelad och att

$$\frac{p^* - p}{\sqrt{p^*(1 - p^*)/n}}$$

är approximativt $N(0, 1)$ -fördelad.

Konfidensintervall

Ett konfidensintervall med approximativ konfidensgrad $1 - \alpha$ fås som

$$I_p = [p^* \pm z_{\alpha/2} \sqrt{p^*(1-p^*)/n}]$$

Eftersom p är en sannolikhet bildas ensidiga intervall som

$$[0, p^* + z_{\alpha} \sqrt{p^*(1-p^*)/n}]$$

och

$$[p^* - z_{\alpha} \sqrt{p^*(1-p^*)/n}, 1]$$

Hypotestest

Vill vi testa

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

eller en ensidig hypotes $H_1 : p > p_0$ eller $H_1 : p < p_0$.

Under H_0 är $X \sim \text{Bin}(n, p_0)$, och vi kan direkt beräkna p-värdet för testet:

- $p = P_{H_0}(X \geq x)$ om $H_1 : p > p_0$.
- $p = P_{H_0}(X \leq x)$ om $H_1 : p < p_0$.
- $p = 2P_{H_0}(X \geq x)$ om $x \geq np_0$ och $p = 2P_{H_0}(X \leq x)$ om $x \leq np_0$ i fallet $H_1 : p \neq p_0$.

Jämförelser i binomialfördelningen

Antag att $X_1 \sim \text{Bin}(n_1, p_1)$ och $X_2 \sim \text{Bin}(n_2, p_2)$ och vi vill testa om det är rimligt att anta att $p_1 = p_2$.

$p_1^* - p_2^*$ är en väntevärdesriktig skattning av $p_1 - p_2$.

Om $n_1 p_1 (1 - p_1)$ och $n_2 p_2 (1 - p_2)$ båda är stora (säg större än 10) så är $p_1^* - p_2^*$ approximativt normalfördelad. Vi har

$$V(p_1^* - p_2^*) = \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}$$

ett konfidensintervall för $p_1 - p_2$ fås som

$$I_{p_1 - p_2} = \left(p_1^* - p_2^* \pm z_{\alpha/2} \sqrt{\frac{p_1^*(1 - p_1^*)}{n_1} + \frac{p_2^*(1 - p_2^*)}{n_2}} \right)$$

Hypotestest

Om vi vill testa $H_0 : p_1 = p_2$ kan vi använda en poolad variansskattning eftersom vi har under H_0 att

$$V_{H_0}(p_1^* - p_2^*) = p(1 - p)\left(\frac{1}{n_1} - \frac{1}{n_2}\right)$$

Här är p det gemensamma värdet under H_0 , vilket skattas som

$$p^* = \frac{x_1 + x_2}{n_1 + n_2}$$

Vi bilar då teststorheten

$$T = \frac{p_1^* - p_2^*}{\sqrt{p^*(1 - p^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

som under H_0 är approximativt $N(0, 1)$ fördelad.

Teckentest för medianen

- Medianen för en kontinuerlig fördelning är definierad som det tal M så att $P(X < M) = P(X > M) = 1/2$.
- Givet ett stickprov av storlek n vill vi testa $H_0 : M = M_0$ mot en ensidig eller tvåsidig mothypotes.
- Låt Q_+ vara antalet observerade värden större än M .
- Låt Q_- vara antalet observerade värden mindre än M .
- Under H_0 är $Q_+ \sim \text{Bin}(n, 1/2)$ och $Q_- \sim \text{Bin}(n, 1/2)$.
 - Om $H_1 : M < M_0$, förkasta H_0 om Q_+ är för liten.
 - Om $H_1 : M > M_0$, förkasta H_0 om Q_- är för liten.
 - Om $H_1 : M \neq M_0$, förkasta H_0 om $\min(Q_+, Q_-)$ är för liten.
- Teststorheten är binomialfördelad, vi kan direkt beräkna p-värdet.
- Om $X_i - M = 0$ räknar vi den observationen till den minsta av Q_+ eller Q_- eftersom det inte motsäger H_0 .

Wilcoxons ranktest

- Beräkna alla absolutdifferenser $|X_i - M_0|$ och ordna dessa i ökande storleksordning.
- Låt R_i vara ranken för den i te absolutdifferensen gånger tecknet på motsvarande avvikelse.
- Beräkna Wilcoxons teststorheter

$$W_+ = \sum_{i: R_i > 0} R_i, \quad |W_-| = \sum_{i: R_i < 0} |R_i|$$

- Definiera teststorheten $W = \min(W_+, |W_-|)$. Kritiska värden för W finns tabulerad för olika n och α .
- Om två absolutdifferenser är lika stora så tilldelar vi dem medelvärdet av motsvarande ranker. Till exempel om vi observerar differenser 2 och -2 som skulle få rankerna 4 och 5 så tilldelar vi dem båda rank 4.5 gånger motsvarande tecken.

Exempel för stickprov i par

I fallet med parade observationer $(X_1, Y_1), \dots, (X_m, Y_m)$ bildar vi först differenserna $X_i - Y_i$ och använder sedan ett ranktest på differenserna för att testa om differensernas median är noll.

Exempel 10.6.2 i boken. Vi vill testa hur mycket minne två statistiska paket använder vid analys av ett dataset. Vi gör mätningar (i Kb).

Program	Paket X	Paket Y	$D_i = X_i - Y_i$
1	512	500	12
2	650	600	50
3	890	890	0
4	410	400	10
5	1050	1025	25
6	1500	1400	100
7	600	625	-25
8	750	710	40

Exempel (forts)

Vi vill testa

$$H_0 : M_X = M_Y$$

$$H_1 : M_X \neq M_Y$$

Vi ordnar differenserna och beräknar rankerna

$ D_i :$	0	10	12	25	25	40	50	100
Rank:	1	2	3	4.5	4.5	6	7	8
$R_i :$	-1	2	3	-4.5	4.5	6	7	8

Vi har nu

$$W_+ = 2 + 3 + 4.5 + 6 + 7 + 8 = 30.5$$

och $|W_-| = |-1| + |-4.5| = 5.5$. Eftersom vi har ett tvåsidigt test använder vi $W = \min(|W_-|, W_+) = 5.5$. Från tabell ser vi att den kritiska punkten för ett tvåsidigt test med $\alpha = 0.1$ är 6.

Eftersom $W = 5.5 < 6$ så kan vi förkasta H_0 .

Wilcoxons ranksummetest (I)

Antag att vi har två oberoende stickprov $X_1, \dots, X_m$ och $Y_1, \dots, Y_n$ från några fördelningar X respektive Y .

Vi vill jämföra medianerna för de två variablerna:

$$H_0 : M_X = M_Y$$

$$H_1 : M_X \neq M_Y$$

Wilcoxon's ranksummetest (II)

- Ordna de $m + n$ värdena i ökande storleksordning och ge varje observation en rank R_i från 1 till $m + n$.
- Om vi har värden som är lika stora tilldelar vi dem medelranken precis som i rank-testet.
- Vi antar att $m \leq n$ och bildar teststorheten W_m som summan av rankerna associerade med variablerna $X_1, \dots, X_m$.
- Det kritiska värdet för W_m finns tabulerat för olika värden på m , n , och α .