

F22, Icke-parametriska metoder.

Christian Tallberg

Statistiska institutionen

Stockholms universitet

Icke-parametriska metoder

Tidigare när vi utfört inferens, dvs utifrån stickprov gjort konfidensintervall eller hypotestest om okända populationsparametrar, har vi antagit att undersökningsvariablerna är normalfördelade. Vid stora stickprov har vi utnyttjat CGS, som säger att skattningarna av populationsparametrarna, dvs stickprovsmedelvärdena, är approximativt normalfördelade oavsett fördelningen för undersökningsvariablerna.

Hur gör vi om vi har små stickprov och inte kan anta att variablerna är normalfördelade?

Icke parametriska konfidensintervall och hypotestest.

- Inga modellantaganden om undersökningsvariablerna behövs.
- Ordinalskala tillräcklig, dvs det räcker om man kan rangordna mätvärdena. Numeriska värden ej nödvändiga.
- Konfidensintervallen blir längre och testfunktionerna vid hypotestest har sämre styrka jämfört med normalfördelningsintervall och normalfördelningstes om det är så att undersökningsvariablerna är normalfördelade. Man utnyttjar inte all information om mätvärdena (enbart ranger).
- Metoderna (konfidensintervallen och testfunktionerna) är robusta. Det vill säga längden på konfidensintervall eller utfallet av beslut vid hypotestest är inte känsligt för modellantaganden (t ex normalfördelade variabler, lika varianser).

Teckentest

Vi har parvisa observationer, dvs vi drar bara *ett stickprov* och får *två mätvärden* för varje individ. Vi har inte två *oberoende* stickprov, utan bildar därför differenser och anger om de är positiva eller negativa (jfr motsvarande parametriska test, Körner sid 212, exempel 16).

Uppgift 1201 (Körner sid 324):

Tio slumpmässigt valda personer poängsatte två diskmedelss A och B, där 10 var högsta betyget och 1 lägsta. Ger data stöd för att det ena diskmedlet är bättre än det andra?

Vi har följande data:

Person	1	2	3	4	5	6	7
A	6	8	7	8	7	8	5
B	5	9	7	6	5	7	4
Diff	+	-	0	+	+	+	+

Vi har 5 plus och 1 minus-tecken. Nollorna stryks vid beräkningarna. Vi låter n vara lika med antal tecken då nollorna stryks.

Definiera den stokastiska variabeln $X =$ "Antal minus-tecken" (välj tecken man har minst av). Fördelningen för X är alltså

$$X \sim \text{Bin}(n; p),$$

där p är sannolikheten för minus-tecken.

Om vi har enbart slumpmässighet vid val av diskmedel är $p = 0.5$. Vi vill alltså testa följande hypoteser:

$$H_0 : p = 0.5 \text{ (Ingen skillnad i preferens)}$$

$$H_1 : p \neq 0.5 \text{ (Skillnad i preferens)}$$

Beroende på hur frågeställningen ser ut kan mothypoteserna också formuleras

$$H_1 : p < 0.5 \text{ (Diskmedel A är populärare)}$$

$$H_1 : p > 0.5 \text{ (Diskmedel B är populärare).}$$

Testvariabeln och dess fördelning är

$$X \sim \text{Bin}(6; 0.5) \text{ då } H_0 \text{ är sann.}$$

Under nollhypotesen är

$$E(X) = np = 6 \cdot 0.5 = 3.$$

Det observerade testvärdet är $x = 1$. Vi har att

P – värdet

$$= 2 \Pr(X \leq 1 | \text{då } H_0 \text{ är sann, dvs } p = 0.5)$$

$$= 2 \cdot 0.1094 = 0.219$$

För $\alpha = 0.05$ kan nollhypotesen inte förkastas. Det kan inte anses statistiskt påvisat att det ena diskmedlet är bättre än det andra. Notera att om mothypotesen hade varit "Diskmedel A är populärare" hade p – värdet blivit 0.1094.

Test av median.

Uppgift 1204 (Körner sid 324):

Man vill undersöka om medianinkomsten i en stor population är större än 180000 kr. Slumpmässigt valdes $n = 100$ personer från populationen där 62 hade en inkomst över 180000.

Definiera den stokastiska variabeln $X =$ "Antal med inkomst över 180000". Fördelningen för X är alltså

$$X \sim \text{Bin}(n; p),$$

där p är sannolikheten för inkomst över 180000. Om medianinkomsten är 180000 kr är det lika sannolikt att dra en individ med inkomst över som under 180000, dvs $p = 0.5$. Vi vill alltså testa följande hypoteser

$$H_0 : p = 0.5 \text{ (Medianink = 180000)}$$

$$H_1 : p > 0.5 \text{ (Medianink > 180000)}$$

Då $n\hat{p}\hat{q} > 5$, är

$$X \sim \text{approx } N(np; np(1-p)) \text{ enl CGS.}$$

Då det observerade testvärdet i stickprovet är $x = 62$ ges p-värdet av

$$\begin{aligned} \Pr(X \geq 62 | H_0 \text{ är sann}) &= \{\text{norm.approx}\} \\ &= \Pr\left(Z \geq \frac{61.5 - 50}{5}\right) \\ &= 1 - \Pr(Z < 2.3) = 1 - 0.989 \\ &= 0.011. \end{aligned}$$

H_0 förkastas för alla $\alpha > 0.011$. Det kan anses statistiskt påvisat att medianinkomsten är större än 180000 kr.

Wilcoxons rangsummatest (Mann-Whitneys test)

Vi har två *oberoende* stickprov från två populationer, och vill undersöka om det föreligger skillnader i populationsmedianerna m_1 och m_2 (jfr motsvarande parametriska test, Körner kap 8.6 sid 208).

Från ena populationen drar vi ett slumpmässigt stickprov av observationer från den stokastiska variabeln X_1 . Oberoende härav drar vi från den andra populationen ett slumpmässigt stickprov av observationer från den stokastiska variabeln X_2 . Vi antar att fördelningarna lika sånär som på läget.

Vi vill därför testa

$$H_0 : m_1 = m_2 \text{ (Populationsmedianerna lika)}$$

mot något av följande alternativ

$$H_1 : m_1 \neq m_2 \text{ (Populationsmedianerna ej lika)}$$

$$H_1 : m_1 < m_2$$

$$H_1 : m_1 > m_2$$

Stickprovsstorlekarna är n_1 respektive n_2 , där $n_1 \leq n_2$. Fördelningen för testvariabeln då H_0 är sann får vi fram på följande sätt. Rangordna samtliga $n_1 + n_2$ observationer. Om två eller flera observationer har samma värde får de samma rangtal, nämligen medelvärdet av de rangtal de skulle fått om man hade kunnat skilja dem åt. Beräkna rangsummorna R_1 i det mindre stickprovet och R_2 i det större stickprovet.

Om $n_1 = 2$ och $n_2 = 3$, är R_1 summan av två av dessa fem möjliga rangtal, 1, 2, 3, 4 och 5. Antal möjliga sätt att välja två av dessa fem rangtal är $\binom{5}{2} = 10$. De tio möjliga utfallen och motsvarande rangsummor R_1 blir:

Utfall	R_1
1, 2	3
1, 3	4
1, 4	5
1, 5	6
2, 3	5
2, 4	6
2, 5	7
3, 4	7
3, 5	8
4, 5	9

Rangsumman R_1 är alltså vår testvariabel och sannolikheten för varje utfall är $\frac{1}{10}$ om nollhypotesen är sann, dvs fördelningsfunktionen för R_1 under nollhypotesen blir:

$$\begin{aligned}\Pr(R_1 \leq 3) &= \frac{1}{10} \\ \Pr(R_1 \leq 4) &= \frac{2}{10} \\ \Pr(R_1 \leq 5) &= \frac{4}{10} \\ \Pr(R_1 \leq 6) &= \frac{6}{10} \\ \Pr(R_1 \leq 7) &= \frac{8}{10} \\ \Pr(R_1 \leq 8) &= \frac{9}{10} \\ \Pr(R_1 \leq 9) &= 1\end{aligned}$$

Kritiska värden anges enbart i den vänstra svansen, dvs för låga värden på R_1 . Notera först att fördelningen för R_1 är symmetrisk. Om $R_1 < R_2$ slår man direkt i tabellen. Om $R_1 > R_2$ måste man beräkna dess symmetrivärde R_s i vänstra svansen på följande sätt

$$R_s = n_1(n_1 + n_2 + 1) - R_1.$$

Nollhypotesen förkastas om R_1 eller R_s är mindre eller lika med det kritiska värdet. Vid ensidigt test ligger signifikansnivån α i vänstra svansen, vid dubbelsidigt test ligger $\alpha/2$ i vänstra svansen.

Uppgift 1205 (Körner sid 324):

Man vill jämföra två undervisningsmetoder (experimentell och traditionell). Medför experimentundervisning bättre resultat än vanlig undervisning? Låt de stokastiska variablerna X_1 och X_2 vara resultaten från experimentell och traditionell undervisning, respektive. Fördelningarna för X_1 och X_2 antas vara lika förutom läget.

Vi vill testa

H_0 : $m_1 = m_2$ (Populationsmedianerna lika)

H_1 : $m_1 > m_2$ (Populationsmedianen större för exp.und.)

för $\alpha = 0.05$. Två slumpmässiga stickprov av storlek $n_1 = 8$ och storlek $n_2 = 12$ dras från X_1 och X_2 . Det kritiska värdet enl. tabell 9 är 62. Vi har följande data:

Exper. Result	Rangtal	Trad Result	Rangtal
29	16	37	20
27	13.5	34	19
24	11	33	18
24	11	31	17
21	7	28	15
20	5.5	27	13.5
16	4	15	3
5	1	24	11
		23	9
		22	8
		20	5.5
		6	2
$R_1 =$	69	R_2	141

Då $R_1 < R_2$ blir 69 testvariabelns observerade värde. Då $69 > 62$ kan nollhypotesen inte förkastas. Det kan inte anses statistiskt påvisat att medianresultatet för experimentundervisningen är högre än medianresultatet för traditionella undervisningen.

Stora stickprov

Om $n_1 \geq 20$ och $n_2 \geq 20$ är

$$R_1 \sim \text{approx } N(E(R_1), V(R_1)),$$

där väntevärde och varians är

$$\begin{aligned}E(R_1) &= n_1 \frac{n_1 + n_2 + 1}{2} \\ V(R_1) &= \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}.\end{aligned}$$

Föregående exempel:

Vi har att

$$\begin{aligned}E(R_1) &= 8 \frac{8+12+1}{2} = 84 \\V(R_1) &= \frac{8 \cdot 12(8+12+1)}{12} = 2016.\end{aligned}$$

Om n_1 och n_2 hade varit större än 20 hade R_1 varit approximativt normalfördelad och p-värdet givits av

$$\begin{aligned}P\text{-värde} &= \Pr(R_1 \leq 62) \\&= \Pr\left(Z \leq \frac{62 - 84}{\sqrt{2016}}\right) \\&= \Pr(Z \leq -0.49) = 1 - \Pr(Z \leq 0.49) \\&= 1 - 0.6879 = 0.312.\end{aligned}$$

Då p-värdet är större än 5% kan H_0 ej förkastas.