MVE051/MSG810 2016 Föreläsning 9

Petter Mostad

Chalmers

November 28, 2016

Att välja mellan modeller

- För att använda stokastiska modeller (variabler) för prediktion behöver man välja en modell och hitta parametrarna i modellen. Vi pratade förra gången om ett sätt att ta fram parametrar i en modell. Nu skall vi fokusera på sätt att välja mellan möjliga modeller.
- Ett tillvägagångssätt baseras på att jämföra sannolikheten för att observera data under de två modellerna, när dessa sannolikheter kan beräknas (i andra fall jämförs även komplexiteterna till modellerna).
- I Milton fokuserar man i stället på två klassiska tillvägagångssätt: Hypotestesting och signifikanstesting.

Hypotestesting

1. Två modeller etableras: *Nollhypotesen* H_0 och den *alternativa* hypotesen H_1. (H_1 är oftast det man "vill visa statistiskt").
2. En *test statistika* T (alltså en funktion av ett stickprov) fastställs, så att
   - Fördelningen till test statistikan T kan beräknas om H_0 är sann.
   - Test statistikan har en typ värden (oftast små) när H_0 är sann och en annan typ värden (oftast stora) när H_1 är sann.
3. Ett förkastningsområde F etableras (generellt ett eller flera intervall), och man bestämmer sig för att *förkasta* H_0 om $T \in F$ medan man inte förkastar H_0 om $T \notin F$.
4. Man beräknar T från observerade data, jämför med F och förkastar H_0 eller inte baserat på detta.

Hypotestestets egenskaper

- Typ I och typ II fel.
- Sedan fördelningen av test statistikan förutsätts känd under H_0 så kan man beräkna sannolikheten för Typ I fel innan man har observerat data. Denna sannolikhet betecknas ofta α, och kallas testets *signifikans*. Oftast väljs T_0 så att $\alpha = 0.05$.
- Motsvarande betecknar β sannolikheten för Typ II fel. Denna sannolikhet kan inte alltid lika lätt beräknas utan bättre specifikation av H_1. $1 - \beta$ kallas testets *styrka*.
- Man försöker välja en test statistika så att styrkan maximeras medan signifikansen fixeras (oftast vid $\alpha = 0.05$).
- Enligt Milton: Om H_0 specificeras som en *samling med hypoteser* så att sannolikheten för typ I fel är *maximalt* α för varje hypotes, så har man fortfarande en hypotestest med signifikans α.

EXEMPEL: Att testa medelvärdet i en normalfördelning

- Antag $X_1, \ldots, X_n$ är ett stickprov från Normal(μ, σ) med μ, σ okända. Antag vi vill jämföra $H_0 : \mu = \mu_0$ med $H_1 : \mu \neq \mu_0$ för någon känd μ_0.
- Vi väljer som test statistika

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

som har en T fördelning med $n-1$ frihetsgrader när H_0 är sann.

- Förkastningsområdet blir alla T så att $T < -T_0$ eller $T > T_0$ för någon T_0. För att signifikansen skall bli α måste vi välja

$$T_0 = t_{\alpha/2}$$

där $t_{\alpha/2}$ är så att Pr$\left[T > t_{\alpha/2}\right] = \alpha/2$ när T har en T-fördelning med $n-1$ frihetsgrader.

- Slutligen beräknar vi vårt värde för T, jämför med T_0 och beslutar om vi förkastar H_0 eller inte.

Ensidiga och tvåsidiga tester

- Testet i exemplet över är en *två-sidig* test. Ett alternativ är en *en-sidig* test, där till exempel $H_0 : \mu \leq \mu_0$ och $H_1 : \mu > \mu_0$.
- Teststatistikan är den samma, men förkastningsområdet blir alla T så att $T > T_0$, där

$$T_0 = t_\alpha$$

  där t_α är så att $\Pr[T < t_\alpha] = \alpha$ när T har en T-fördelning med $n - 1$ frihetsgrader.
- Motsvarande kan man konstruera en en-sidig test med $H_0 : \mu \geq \mu_0$ och $H_1 : \mu < \mu_0$.

Tolkning av resultat av ett hypotestest

- Korrekt tolkning: Om man använder hypotestestet som en beslutsregel, så kommer man i en lång serie med hypotestest att förkasta en ungefärlig andel α av korrekta H_0-hypoteser.
- Märk: Man kan efter ett hypotestest inte säga någonting precist om sannolikheten för att H_0 eller H_1 är sann.
- Märk: Om man förkastar H_0 så betyder det inte nödvändigtvis att H_1 är sann.
- Märk: Om man inte förkastar H_0 så betyder det inte att H_0 är bevisat sann!
- Märk: Konklutionen beror inte bara på data men också på val av test statistika.

Signifikanstestning

- Signifikanstesting representerar en vidareutveckling av hypotestesting: I stället för att först bestämma ett signifikansnivå α och sen förkasta H_0 eller inte, så beräknas först värdet av teststatistikan, och sen beräknas *det minsta signifikansnivå* α *som skulle göra det möjligt att förkasta* H_0 med detta värdet av teststatistikan.
- Detta minsta signifikansnivå kallas *p-värdet* för testet.
- Om man beräknar p-värdet för ett test och sen förkastar H_0 om p-värdet är mindre eller lika med 0.05 och annars inte, så får man tillbaka en hypotestest med signifikansnivå 0.05. Motsvarande om vi ersätter 0.05 med andra tal.

EXEMPEL: Signifikanstesting för medelvärdet i en normalfördelning

- Antag $X_1, \ldots, X_n$ är ett stickprov från Normal(μ, σ) med μ, σ okända. Antag vi vill jämföra $H_0 : \mu = \mu_0$ med $H_1 : \mu \neq \mu_0$ för någon känd μ_0.
- Vi väljer som test statistika

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

  som har en T fördelning med $n - 1$ frihetsgrader när H_0 är sann.
- Vi beräknar värdet av T för våra data, väljer $t_{\alpha/2}$ så att $T = -t_{\alpha/2}$ (om $T < 0$) eller $T = t_{\alpha/2}$ (om $T > 0$), och använder tabellen för T-fördelningen med $n - 1$ frihetsgrader för att beräkna α, som blir p-värdet.
- På motsvarande sätt kan vi beräkna p-värdet för en-sidiga tester.

Tolkning av p-värdet

- Ett p-värde säger något om hur "extrema" data är när det gäller att antyda att H_1 är sann, i förhållande till slumpmessiga variationer i data förväntad under H_0.
- Utöver att ange definitionen av p-värden är det svårt att ge korrekta sannolikhetsteoretiska tolkningar av p-värden. Det är lättare att beskriva felaktiga tolkningar:
  - p-värdet anger INTE sannolikheten för att H_0 är sann.
  - p-värdet kan heller inte relateras till sannolikheten för att H_1 är sann.
- Kom i håg att p-värdet är definerad i relation till val av test-statistika, och beror inte bara på data och hypoteser.

Exempel

- Antag vi undersöker en typ försök som varje gång resulterar i success (1) eller misslyckande (0), och sannolikheten för success är en okänd parameter p. Antag vi gör ett antal försök, och får resultaten

$$0, 1, 0, 0, 1, 0, 0, 1$$

  Vår nollhypotes för p är $H_0 : p \geq 0.6$, medan den alternativa hypotesen är $H_1 : p < 0.6$. Vad blir p-värdet?
- För att svara på frågan måste vi veta villken *test statistika* som skall användas. Olika test statistika ger olika resultat.
- ALTERNATIV 1: Test statistika: Gör 8 försök, låt X vara antalet successer.
- ALTERNATIV 2: Test statistika: Gör försök tills man fått fram 3 lyckade försök, och låt X vara antalet försök man behövde göra.
- Om vi använder alternativ 1 får vi ett p-värde på 0.174, om vi använder alternativ 2 blir p-värdet 0.095. Så med en gräns på 0.1 förkastar vi H_0 i det andra fallet men inte i det första.

Beräkning av p-värdena i exemplet

- ALTERNATIV 1: Om vi antar att $p = 0.6$ så får vi $X \sim \text{Binomial}(8, 0.6)$. De möjliga värdena för X och deras sannolikheter finns i tabellen:

0	1	2	3	4	5	6	7	8
0.001	0.008	0.041	0.124	0.232	0.279	0.209	0.090	0.017

Summan av sannolikheterna för $X = 0, 1, 2$, eller 3 är 0.174. Detta är sannolikheten för att teststatistikan ger det observerade värdet eller något *mera extremt* i förhållande till att "motbevisa" H_0, och är därmed p-värdet.

- ALTERNATIV 2: Om vi antar $p = 0.6$ så får vi $X \sim \text{Neg-Binomial}(3, 0.6)$. De möjliga värdena för X och deras sannolikheter finns i tabellen:

3	4	5	6	7	8	9	10	11
0.216	0.259	0.207	0.138	0.083	0.046	0.025	0.013	0.006

12	13	14	15	16,17,. . .
0.003	0.001	0.001	0.000	totalt 0.000

Summan av sannolikheterna för att $X = 8, 9, 10, \ldots$ är 0.095. Detta är sannolikheten för att teststatistikan ger det observerade värdet eller något *mera extremt* i förhållande till att "motbevisa" H_0, och är därmed p-värdet.

Exempel: p-värden inom utveckling av nya mediciner

- Godkänning av nya mediciner är hårt reglerat och processen baseras på p-värden.
- Efter att de medicinska försöken är gjort skulle det vara möjligt för företaget att välja bland möjliga hypoteser, data och test-statistika de som ger bäst resultat för företaget.
- För att undvika detta behöver företaget submitta en beskrivning av hur de skall beräkna p-värdet för resultatet av försöket, *innan* försöket görs: Med andra ord behöver de submitta sin test statistika på förehand.
- Därmed kan det hända att två företag, som har utfört exakt samma försök och fått exakt samma data, får olika svar om godkänning av läkemedlet: Ett företag som har satsat på en test statistika som i praktiken fungerat dåligt kan få et p-värde med $p > 0.05$ och ingen godkänning, medan ett annat föredag kan få $p < 0.05$ och därmed godkänd läkemedel.

Exempel: Användning av kontextuell information

- Antag du vill undersöka om en mynt är rättvis, dvs. att sannolikheten p för "kron" är 0.5. Du kastar myntet 8 gångar och får "kron" 2 gångar. Vad tror du om sannolikheten p, och hur säker kan du vara?
- Antag en läkare har fått tillstånd för en ny experimentell procedur. Efter 8 procedurer är 2 lyckade. Vad tror du om annolikheten p för en lyckad procedur och hur säker kan du vara?
- Antag en fabrik vill göra kvalitetskontroll på sina produkt. Av 8 slumpmässingt valda produkt har 2 fel. Vad tror du om sannolikheten p för att en produkt har fel, och hur säker kan du vara?
- I praktiken, om du skall göra de bästa möjliga prediktionerna, så skulle du använda kontextuell information och förmodligen komma till olika konklutioner om p i de tre fallen. Om man utvärderar data via konfidensintervaller eller signifikanstesting, så kan man inte fånga upp sådan information. Utvärderingen beror i stället bara på val av estimator eller test statistika.

Nyttovärdet av p-värden

- I Milton står att användning av p-värden "is coming into widespread use" på grund av metodens "logical appeal".
- Realiteten är att p-värden speciellt, och hypotestesting generellt, är mycket omdebatterad.
- Exempel: Tidsskriftet "Basic and Applied Social Psychology" beslutade februari 2015 att inte längre acceptera artikler som innehåller p-värden.
- Jag skulle säga att
  - Om p-värden tolkas korrekt så har de ett mycket begränsad användningsområde.
  - I de flesta tillämpningar är den informationen man kan få från en korrekt tolkning av ett p-värde inte intressant.
  - Det största problemet i tillämpningar är att p-värdet misstolkas.
  - Dock vill den felaktiga tolkningen i ett flertal tillämpningar vara rimlig.
  - Det finns goda alternativ till att använda p-värden.
  - Min personliga rekommendation: Använd aldrig p-värden, ej heller hypotestesting, om du känner alternativa metoder.