Hypotesetesting

Notat til STK1110

Ørnulf Borgan
Matematisk institutt
Universitetet i Oslo

September 2010

De grunnleggende idéene og begrepene ved hypotesetesting er beskrevet i avsnitt 9.1 læreboka til Devore og Berk. Formålet med dette notatet er å illustrerer disse idéene og begrepene ved hjelp av et eksempel. Det gir et supplement til framstillingen i avsnitt 9.1 i læreboka (men det erstatter ikke dette avsnittet).

Problemstilling

Et farmasøytisk firma ønsker å finne ut om en ny salve mot eksem er bedre enn den gamle. For å gjøre det, utfører de et forsøk – en såkalt klinisk prøving. Hundre pasienter, som har eksem på begge hendene (omtrent like alvorlig på hver hånd), tar del i forsøket. Hver pasient får ved loddtrekning den nye salven på en hånd og den gamle salven på den andre.

For å unngå at subjektive vurderinger skal påvirke resultatet, gjøres forsøket "dobbelt blindt". Det betyr at hverken pasienten eller behandlende lege vet hvilken salve som brukes på hver hånd. Det er det bare det farmasøytiske firmaet som kjenner til. Det får en til ved at firmaet pakker salvene i nøytrale tuber som bare er merket med pasientnummer og hvilken hånd tuben skal brukes på.

Etter tre uker avgjør legen hvilken av de to hendene som nå er best, og det farmasøytiske firmaet finner ut av sine lister om denne hånden har blitt behandlet med den gamle eller den nye salven. De teller så opp antall pasienter der den nye salven ga best resultat. (Vi forutsetter at en alltid klarer å avgjøre hvilken hånd som er best. I praksis vil det imidlertid alltid være noen pasienter hvor en ikke klarer å gjøre det.)

La oss si at den nye salven ga best resultat for 60 av pasientene. Kan det farmasøytiske firmaet *med rimelig grad av sikkerhet* konkludere med at den nye salven er bedre enn den gamle?

Nullhypotese og alternativ hypotese

For å avgjøre om det farmasøytiske firmaet med rimelig grad av sikkerhet kan konkludere med at den nye salven er bedre enn den gamle, formulerer vi problemstillingen som et hypotesetestingsproblem.

La X være antall pasienter hvor den nye salven er best. I det konkrete forsøket fikk X verdien 60. Men hvis vi hadde gjentatt forsøket ville nok X ha fått en annen verdi. Det er

derfor naturlig å se på X som en stokastisk variabel. En rimelig sannsynlighetsmodell er at X er binomisk fordelt med $n = 100$ og sannsynlighet p.

Problemstillingen til det farmasøytiske firmaet kan nå "oversettes" til hypoteser om parameteren p:

- Hvis den nye salven *er bedre* enn den gamle, er $p > 0.50$.
- Hvis den nye salven *ikke er bedre* enn den gamle, er $p \leq 0.50$.

Når vi formulerer problemstillingen som et hypotesetestingsproblem, velger vi som *nullhypotese* (H_0) at den nye salven ikke er bedre enn den gamle, mens den *alternative hypotesen* (H_a) er at den nye salven er bedre enn den gamle. For hvis vi da forkaster H_0, kan vi "med rimelig grad av sikkerhet" konkludere med at H_a er sann, dvs. at den nye salven er best (jf. nedenfor). Vi ønsker altså å teste nullhypotesen $H_0 : p \leq 0.50$ mot den alternative hypotesen $H_\text{a} : p > 0.50$.

Test, testobservator og forkastningsområde

En *test* er en regel som forteller oss for hvilke verdier av X vi skal *forkaste* H_0. Den stokastiske variabelen vi baserer testen på, kalles en *testobservator*. Her er det altså X som er testobservatoren.

Siden alternativet er at den nye salven er bedre enn den gamle, er det rimelig å forkaste nullhypotesen (og altså konkludere med at H_a er sann) hvis den nye salven er best for tilstrekkelig mange pasienter. Vi vil altså forkaste nullhypotesen hvis X er tilstrekkelig stor, det vil si hvis $X \geq k$ for en passende valgt k. Spørsmålet er hvordan k bør velges.

La oss først prøve med $k = 55$, det vil si at testen forkaster H_0 så sant $X \geq 55$. De verdiene av testobservatoren som gir forkastning av H_0, kaller vi *forkastningsområdet*. Her er altså forkastningsområdet $X \geq 55$.

Feil av type I og feil av type II

Når vi utfører en hypotesetest kan vi gjøre to typer feil:

- Hvis vi *forkaster H_0 når den er sann*, gjør vi en *feil av type I*.
- Hvis vi *ikke forkaster H_0 når den er gal*, gjør vi en *feil av type II*.

Det farmasøytiske firmaet gjør altså en feil av type I hvis den nye salven ikke er bedre enn den gamle ($p \leq 0.50$), men de likevel forkaster nullhypotesen (og dermed konkluderer med at den nye salven er best). Firmaet gjør en feil av type II hvis den nye salven er bedre enn den gamle ($p > 0.50$), men de likevel ikke forkaster nullhypotesen (slik at de ikke oppdager at den nye salven er best).

Signifikansnivå

Vi ser i første omgang på sannsynligheten for å gjøre en feil av type I. For testen som forkaster H_0 så sant $X \geq 55$, har vi at

$$\begin{aligned} P(\text{feil type I}) &= P(\text{forkaste } H_0 \,|\, H_0 \text{ er sann}) \\ &= P(X \geq 55 \,|\, p \leq 0.50) \\ &\leq P(X \geq 55 \,|\, p = 0.50) \\ &= 0.184 \end{aligned}$$

Du kan bruke `R` til å finne at $P(X \geq 55 \,|\, p = 0.50) = 0.184$. Du gir da kommandoen `1-pbinom(54,100,0.50)`.

Testen som forkaster nullhypotesen når $X \geq 55$ har en sannsynlighet for feil av type I som er 18.4%. Det vil en normalt regne som en alt for stor sannsynlighet for å forkaste H_0 når den er sann. Vi må derfor velge en større verdi av k.

Ved å gi `R`-kommandoene

```
k=56:62
cbind(k,1-pbinom(k-1,100,0.50))
```

finner vi:

k	56	57	58	59	60	61	62
$P(X \geq k \,\|\, p = 0.50)$	0.136	0.097	0.067	0.044	0.028	0.018	0.010

Hvis vi ønsker at sannsynligheten for feil av type I skal være høyst 5%, ser vi at vi må velge en verdi av k som er minst lik 59. På den andre siden vil vi ikke velge k større enn nødvendig, for da øker sannsynligheten for feil av type II. Vi får derfor en rimelig test hvis vi forkaster H_0 så sant $X \geq 59$. For denne testen er sannsynligheten for feil av type I lik 4,4%. Vi sier at testen har *signifikansnivå* $\alpha = 0.044$.

Tradisjonelt er det vanlig å kreve at signifikansnivået til en test, dvs. sannsynligheten for feil av type I, er høyst lik 1% eller høyst lik 5%. Grunnen til at vi vil ha en så liten sannsynlighet for feil av type I, er at hvis vi da forkaster H_0, kan vi være rimelig sikre på at H_a er sann.

Sannsynlighet for feil av type II

Vi ser så på sannsynligheten for feil av type II, som vi betegner β. Vi finner:

$$\beta = P(\text{feil type II}) = P(\text{ikke forkaste } H_0 \,|\, H_0 \text{ er gal}) = P(X \leq 58 \,|\, p > 0.50)$$

Vi ser at sannsynligheten for feil av type II avhenger av verdien til p. Vi kan bruke `R` til å regne ut sannsynligheten for feil av type II for ulike verdier av $p > 0.50$. Ved å gi kommandoene

```
p=seq(0.55,0.75,0.05)
cbind(p,pbinom(58,100,p))
```

finner vi:

p	0.55	0.60	0.65	0.70	0.75
P(feil type II)	0.759	0.377	0.088	0.007	0.0001

Vi ser for eksempel at hvis det er slik at $p = 0.60$, så er sannsynligheten 37.7% for at vi gjør en feil av type II. Så selv om den nye salven faktisk er bedre enn den gamle for 60% av pasientene i det lange løp, så er sannsynligheten så stor som 37.7% for at det farmasøytiske firmaet ikke vil oppdage det. Det at sannsynligheten for feil av type II kan være så stor, betyr at hvis vi ikke forkaster H_0 kan vi *ikke* konkludere med at H_0 er sann. Det *kan* være at H_0 er sann. Men det *kan også* være at H_0 er gal, men at vi ikke har nok "utsagnskraft" i dataene våre til å vise det.

Hvis det derimot er slik at $p = 0.70$, så er sannsynligheten bare 0.7 % for at vi ikke vil forkaste nullhypotesen. Det betyr at hvis den nye salven er klart bedre enn den gamle, er det farmasøtiske firmaet rimelig sikre på å oppdage det.

Teststyrke og styrkefunksjon

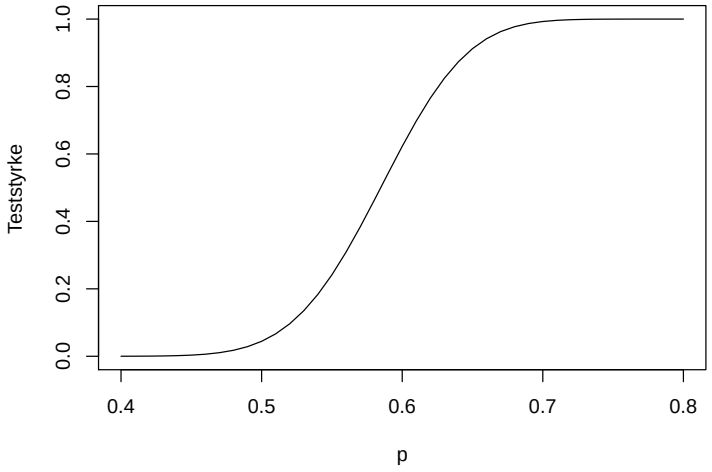
For å se hvilke egenskaper en test har, er det nyttig å se på *styrken* til testen. Teststyrken er sannsynligheten for at testen vil forkaste nullhypotesen (uansett om den er sann eller ikke). På samme måte som at sannsynligheten for feil av type II avhenger av verdien til p, vil også teststyrken gjøre det. Styrkefunksjonen $\gamma(p)$ er teststyrken som en funksjon av p:

$$\gamma(p) = P(\text{forkaste } H_0 \,|\, p) = P(X \geq 59 \,|\, p)$$

Vi kan beregne og plotte styrkefunksjonen i `R` ved å gi kommandoene

```
p=seq(0.40,0.80,0.01)
plot(p,1-pbinom(58,100,p),type="l",xlab="p",ylab="Teststyrke")
```

Det gir figuren nedenfor:



Av styrkefunksjonen kan vi lese av sannsynligheten for å forkaste nullhypotesen for ulike verdier av p. Når $p \leq 0.50$ er teststyrken høyst lik signifikansnivået α. Det er som det skal være, for vi vil ha liten sannsynlighet for å forkaste H_0 når den er sann. Når $p > 0.50$ ønsker vi at teststyrken skal være så stor som mulig. Av figuren ser vi imidlertid at teststyrke bare er omtrent 25% når $p = 0.55$. Men hvis $p = 0.65$ er den er over 90%.

Konklusjon

Vi har over kommet fram til at en test med signifikansnivå 4.4% forkaster nullhypotesen $H_0 : p \leq 0.50$ til fordel for den alternative hypotesen $H_\text{a} : p > 0.50$ så sant $X \geq 59$. I den kliniske prøvingen fant en at $X = 60$ (jf. avsnittet "Problemstilling" på side 1). Det betyr at nullhypotesen forkastes *på signifikansnivå* 4.4%, og det farmasøytiske firmaet kan med rimelig grad av sikkerhet konkluder med at den nye salven er bedre enn den gamle.

P-verdi

Slik vi har beskrevet hypotesetesting ovenfor, er det en skarp grense for forkastning av nullhypotesen. Den forkastes (på signifikansnivå 4.4%) hvis $X \geq 59$ og den forkastes ikke hvis $X \leq 58$. Denne måten å gjøre det på er hensiktsmessig når vi skal forklare idéene og begrepene i hypotesetesting og når vi skal studere egenskapene til en test (f.eks. sannsynligheten for feil av type II).

Men framgangsmåten har også sine begrensninger. For det virker ikke rimelig at konklusjonen skal bli den samme (nemlig forkast H_0 på signifikansnivå 4.4%) om vi observerer at den nye salven er best for 60 pasienter som den hadde blitt om vi observerer at den nye salven er best for 65 pasienter. I det siste tilfellet vil vi føle oss sikrere på at H_0 skal forkastes enn vi er i det første tilfellet.

Når vi skal gi resultatet av en hypotesetest, bør vi derfor ikke nøye oss med å raportere om nullhypotesen ble forkastet eller ikke på et gitt signifikansnivå. Vi bør i stedet oppgi *P-verdien* (eller signifikanssannsynligheten). Den er definert som det *minste signifikansnivået* vi kan bruke hvis vi skal få forkastet nullhypotesen.

Vi har at $P(X \geq 60 \,|\, p = 0.50) = 0.028$ og $P(X \geq 65 \,|\, p = 0.50) = 0.002$. Hvis den nye salven er best for 60 av pasientene, er derfor P-verdien 2.8%, mens P-verdien er 0.2% hvis den nye salven er best for 65 pasienter. Vi ser at P-verdien gir oss informasjon om hvor sikre vi er på at nullhypotesen skal forkastes.

En fyldigere diskusjon av P-verdien er gitt i avsnitt 9.4 i læreboka.