Jørgen Larsen
9. november 2010

Vedr. tostikprøveproblemer i normalfordelingen: et eksempel og tre opgaver

Eksempel 1 (Sovemidler)
Det kemiske stof *hyoscyamin hydrobromid* kan anvendes som sovemiddel. Stoffet findes imidlertid i to udgaver, d-hyoscyamin hydrobromid og l-hyoscyamin hydrobromid, der afbøjer polariseret lys til hver sin side (l = laevo = mod venstre, d = dextro = mod højre (på latin)). Man er interesseret i at finde ud af om de to udgaver virker på samme måde som sovemiddel. Derfor har man udført en forsøgsrække hvor man på 10 forsøgspersoner har bestemt stoffernes søvnforlængende virkning. I tabel 1 er vist det gennemsnitlige antal ekstra søvntimer pr. nat for hver person, dels ved behandling med d-udgaven, dels ved behandling med l-udgaven af stoffet.

Da der er tale om at man på nogle forsøgspersoner har målt effekten af først en, så en anden behandling, vil det være nærliggende at søge at analysere talmaterialet ved hjælp af en model af typen »tostikprøveproblem med parrede observationer«. Derfor bestemmes differenserne mellem virkningerne af laevo- og dextroudgaven af stoffet, se tabel 2.

Vi vil opfatte tallene i tabel 2 som et »enstikprøveproblem i normalfordelingen«, og spørgsmålet om de to stoffer virker lige godt, kan da præciseres til spørgsmålet om tallenes middelværdi er signifikant forskellig fra 0. Dette kan testes som en statistisk hypotese.

Gennemsnittet af differenserne i tabellen er $\overline{d} = 1.58$ timer, og estimatet over variansen på differenserne er $s^2 = 1.51$ timer^2 (med 9 frihedsgrader), svarende til at den estimerede standardafvigelse er $s = 1.23$ timer. Den estimerede standardafvigelse på gennemsnittet er dermed $\sqrt{s^2/n} = \sqrt{1.51/10}$ timer $= 0.39$ timer. Endvidere bliver t-teststørrelsen

$$t = \frac{\overline{d} - 0}{\sqrt{s^2/n}} = \frac{1.58 \text{ timer}}{0.39 \text{ timer}} = 4.06\ .$$

I t-fordelingen med 9 frihedsgrader er 99.5%-fraktilen 3.25 og 99.9%-fraktilen 4.29, så testsandsynligheden ligger et sted mellem 0.2% og 1%. Der er således ganske klart signifikans, dvs. de to stoffer virker signifikant forskelligt (og som man ser, er l-stoffet det mest virksomme).

Dette var altså et eksempel på et tostikprøveproblem med parrede observationer, men hvad var der sket hvis man af vanvare var kommet til at analysere det som om der var tale om uparrede observationer?

Den t-størrelse man så ville udregne, var en anden. Tælleren ville være den samme fordi differensen mellem gennemsnittene er lig gennemsnittet af differenserne. Det variansestimat der

Tabel 1 Antal ekstra søvntimer ved behandling med hyoscyamin hydrobromid.

person	dextro-	laevo-
1	0.7	1.9
2	−1.6	0.8
3	−0.2	1.1
4	−1.2	0.1
5	−0.1	−0.1
6	3.4	4.4
7	3.7	5.5
8	0.8	1.6
9	0.0	4.6
10	2.0	3.4

Tabel 2 Differenser mellem l- og d-hyoscyamin hydrobromids søvnforlængende virkning.

person	differens (timer)
1	1.2
2	2.4
3	1.3
4	1.3
5	0.0
6	1.0
7	1.8
8	0.8
9	4.6
10	1.4

skulle benyttes i nævneren, er estimatet over den fælles varians i de to grupper, og det udregnes til $s_0^2 = 3.605$ timer^2 med 18 frihedsgrader, og teststørrelsen ville derfor blive

$$t = \frac{1.58 \text{ timer}}{\sqrt{3.605 \text{ timer}^2\left(\frac{1}{10} + \frac{1}{10}\right)}} = \frac{1.58}{0.85} = 1.86\ .$$

Denne gang ville vi få 18 frihedsgrader i t-fordelingen, og det vil sige at 95%-fraktilen er 1.73 og 97.5%-fraktilen 2.10. Der ville altså være et sted mellem 5% og 10% chance for at få en mere ekstrem t-størrelse end 1.86, og man vil derfor almindeligvis sige at $t_{\text{obs}} = 1.86$ *ikke* er signifikant stor. Dette test ville således ikke vise nogen signifikant forskel på de to stoffer.

Grunden til at de to analyser giver forskellige resultater, er at der er en temmelig stor forskel på forsøgspersonerne:

- I den første model (parrede observationer) elimineres en stor del af personforskellene ved at man går over til at analysere differenserne. Til gengæld får variansestimatet kun 9 frihedsgrader.
- I den anden model (uparrede observationer) skal al forskellen mellem personer beskrives af variansparameteren (fordi forskellen mellem personer i denne omgang udelukkende anses for tilfældig), og til gengæld får variansestimatet hele 18 frihedsgrader. – På den anden side indebærer det at hvis der *er* stor forskel mellem personer, så bliver variansestimatet også stort.

Datamaterialet til dette eksempel er meget berømt fordi det blev benyttet til et illustrativt eksempel i den artikel hvor t-testet (i enstikprøveproblemet) blev introduceret ('Student', 1908). Artiklen er skrevet af W.S. Gosset der arbejdede som biometriker ved Guinnessbryggerierne, og som benyttede 'Student' som sit *nom de plume*.

Opgave 1 (Isens smeltevarme)
Man ønsker at sammenligne to forskellige metoder (A og B) til bestemmelse af isens smeltevarme.

Tabel 3 Opgave 1: Varmemængde (i kalorier) for at smelte 1 g is med en begyndelsestemperatur på -0.72°C, bestemt ved to forskellige metoder.

Metode A	Metode B
79.98	80.02
80.04	79.94
80.02	79.98
80.04	79.97
80.03	79.97
80.03	80.03
80.04	79.95
79.97	79.97
80.05	
80.03	
80.02	
80.00	
80.02	

Tabel 4 Opgave 2: Den maksimale procentdel af blodpladerne der klumper sig sammen før henholdsvis efter en given påvirkning.

før	efter
25	27
25	29
27	37
44	56
30	46
67	82
53	57
53	80
52	61
60	59
28	43

Eksperimenter har givet resultaterne i tabel 3. Undersøg om der er signifikant forskel på de to metoder.
TIP: Udregningerne bliver lettere hvis man indfører et passende beregningsnulpunkt.

Opgave 2 (Rygning og blodpropper)
På 11 forsøgspersoner har man taget blodprøver før og efter de røg en cigaret, og man har så undersøgt blodpladernes tendens til at klumpe sig sammen (sådanne klumper kan udvikle sig til regulære blodpropper). Resultaterne ses i tabel 4.

Undersøg om resultaterne tyder på at rygning påvirker blodpladernes tendens til at klumpe sig sammen. (Der er øjensynligt tale om et tostikprøveproblem af en slags; der kan så være tale om *parrede* eller *uparrede* observationer. Det kan være illustrativt at forsøge sig med begge slags modeller. Hvad er forskellen? Argumentér for at den ene af dem er mere rigtig end den anden.)

Opgave 3
Man har foretaget nogle forsøg med mus for at finde ud af om de to forskellige former for jernioner Fe^{2+} og Fe^{3+} optages med forskellig hastighed i organismen. Dette er af betydning når man skal sammensætte kosttilskud (eksempelvis vitaminpiller) til mennesker.

Som led i et større forsøg har man givet 18 mus Fe^{2+} og 18 andre mus Fe^{3+}, i begge tilfælde i 1.2 millimolar opløsninger indgivet oralt. Jernatomerne var radioaktivt mærkede således at det var muligt at måle hvor meget jern der blev optaget i musen i løbet af et fastsat stykke tid. Tabel 5 viser hvor stor en procentdel af den tilførte mængde jern der blev optaget af musen.

1. Ved data af denne type kan man erfaringsmæssigt ofte beskrive logaritmen til observationerne med en normalfordeling. Undersøg om det er rimeligt at gøre det i dette tilfælde.

Tabel 5 Opgave 3: Procentdel optaget jern samt titalslogaritmen til procentdel optaget jern, for 18 mus der har fået Fe^{2+} og 18 mus der har fået Fe^{3+}. – De enkelte søjler indeholder de *ordnede* observationer.

	y		$\log_{10} y$	
	Fe^{2+}	Fe^{3+}	Fe^{2+}	Fe^{3+}
	2.20	4.04	0.342	0.606
	2.93	4.16	0.467	0.619
	3.08	4.42	0.489	0.645
	3.49	4.93	0.543	0.693
	4.11	5.49	0.614	0.740
	4.95	5.77	0.695	0.761
	5.16	5.86	0.713	0.768
	5.54	6.28	0.744	0.798
	5.68	6.97	0.754	0.843
	6.25	7.06	0.796	0.849
	7.25	7.78	0.860	0.891
	7.90	9.23	0.898	0.965
	8.85	9.34	0.947	0.970
	11.96	9.91	1.078	0.996
	15.54	13.46	1.191	1.129
	15.89	18.40	1.201	1.265
	18.30	23.89	1.262	1.378
	18.59	26.39	1.269	1.421
sum	147.67	173.38	14.862	16.339
sum af kvadrater	1715.9265	2431.1648	13.662209	15.885527

2. Undersøg om data tyder på at Fe^{2+} og Fe^{3+} optages på samme måde (sammenlign for eksempel de to stikprøver af logaritmerede målinger).
3. Man vil planlægge et nyt forsøg af samme slags, blot med et andet antal mus. Det nye forsøg skal kunne afgøre om der er en reel forskel på 0.1 (på den logaritmiske skala) mellem Fe^{2+}- og Fe^{3+}-optagelsen. I den forbindelse kan man vælge at sige at »en reel forskel på 0.1« skal betyde at hvis tælleren i t-teststørrelsen, altså differensen mellem middeltallene, er større end eller lig 0.1 (eller mindre end eller lig -0.1), så vil testsandsynligheden blive mindre end eller lig 5% (»der er signifikans på niveau 5%«).
   Spørgsmålet er hvor mange mus der skal benyttes: Omsæt ovenstående præcisering af »en reel forskel på 0.1« til matematik, og få derved en ulighed der kan løses med hensyn til den ubekendte »antal mus«.

Litteratur

‘Student’. The probable error of a mean. *Biometrika*, 6:1–25, 1908.