Holme, 2007-03-10

Signifikante forskelle på hitlister

Indledning

I mange sammenhænge indgår der hitlister fx liste over sorte pletter på veje eller de mest solgte cd'er. Når samme liste opstilles en uge senere, vil der typisk være afvigelser fra den foregående, og da de ofte bliver opstillet på baggrund af stikprøver, er der god mening i at spørge om de er signifikant forskellige. Nedenfor forsøges det at tage fat på løsningen af dette problem. Det har ikke været muligt at finde referencer på Google, hvor der ellers er mange referencer til fordelingsfrie tests som vel er det aktuelle område inden for statistikken da vi ikke har nogen hypotese om en bagvedliggende fordeling.

Analyse

Når man skal se på forskelligheder mellem hitlister, kan man starte med nogle meget enkle lister med kun fire emner på fire pladser:

I	IIa	IIb	IIc
A	B	C	D
B	A	D	B
C	D	A	C
D	C	B	A

Udgangslisten er benævnt I, og IIa-IIc er nogle mulige, senere udfald af listen. Umiddelbart vil man måle forskellen på udgangslisten og de øvrige ved at definere en afstandsfunktion i form af det samlede antal flyttede pladser for alle emnerne. Normalt vil man også benytte en afstandsfunktion der straffer store forskelle ved at kvadrere afstanden. I de tre tilfælde vil den samlede afstand blive hhv. 4, 16 og 18 ($4*1^2$, $4*2^2$ og $2*3^2$).

Havde man ikke benyttet en kvadrering, havde man fået hhv. 4, 8 og 6 og dermed en anden rangordning af listerne end ovenfor. Intuitivt vil man nok mene at IIc afviger mere fra I end IIb gør, så der arbejdes videre med kvadratet på afstandene.

I tilfældet med fire emner på fire pladser er der 4! = 24 muligheder for forskellige hitlister, og for hver af disse permutationer af emnerne kan man beregne afstandsfunktionen. Det viser sig at afstandsfunktionen kan antage værdierne fra 0 til 20, og antallet af permutationer for hver af disse værdier udgør hhv.

1 3 1 4 2 2 2 4 1 3 1

eller 24 i alt. Den simpleste hitliste med to pladser har de to mulige afstande 0 og 2 med en mulighed af hver. Hvis man nu foretager samme optælling for hitlister med op til 7 pladser, fås nedenstående fordeling:

Nummer	1	2	3	4	5	6
Pladser	2	3	4	5	6	7
Permutationer	2	6	24	120	720	5040
Max samlet afstand	2	8	20	40	70	112
Antal grupper	2	5	11	21	36	57
Permutationer	1	1	1	1	1	1
per gruppe	1	2	3	4	5	6
		0	1	3	6	10
		2	4	6	9	14
		1	2	7	16	29
			2	6	12	26
			2	4	14	35
			4	10	24	46
			1	6	20	55
			3	10	21	54
			1	6	23	74
				10	28	70
				6	24	84
				10	34	90
				4	20	78
				6	32	90
				7	42	129
				6	29	106
				3	29	123
				4	42	134
				1	32	147
					20	98
					34	168
					24	130
					28	175
					23	144
					21	168
					20	144
					24	184
					14	144
					12	168
					16	144
					9	175
					6	130
					5	168
					1	98
						147
						134
						123
						106
						129
						90
						78
						90
						84
						70
						74
						54
						55
						46
						35
						26
						29
						14
						10
						6
						1

Optællingen er også foretaget for hitlister med otte og ni pladser hvorefter det bliver uhåndterligt i Excel.

Man kan bruge tabellen på følgende måde for at besvare det oprindelige spørgsmål vedr. signifikante afvigelser:
Hvis hitlisten har fem pladser, og hypotesen er at de to lister er forskellige, så vil man afvise denne hypotese hvis den målte afstand tilhører de mindste 5 % dvs. de fem tilfælde hvor afstandsfunktionen er 0 eller 2. Hvis omvendt hypotesen er at de to lister er ens, vil man afvise den hvis afstandsfunktionen er 38 eller 40.

Nu er det besværligt at danne ovenstående tabel, og hitlister med mere end 9 pladser bør også kunne vurderes, så det ville være ønskeligt om man kunne danne en af søjlerne i tabellen ud fra den eller de foregående søjler analogt med Pascals trekant.

Problemer

1. Der synes ikke at være nogen måde at danne søjlerne ud fra de foregående. Og det er karakteristisk at værdierne ikke stiger og falder monotont.

2. Talfølgerne som fx 1 3 1 4 2 2 2 4 1 3 er ikke kendt på webstedet 'The On-Line Encyclopedia of Integer Sequences', hvorimod de samme talfølger med en ikke-kvadreret afstandsfunktion er kendt. For en sådan hitliste med fem pladser kan den samlede afstand komme op på 12, og den tilsvarende talfølge er: 1 4 12 24 35 24 20 (i alt 120). Denne følge kan findes som nummer A062869 på webstedet.

3. Antallet af pladser, P, kan skrives som følgende udtryk hvor N er ovenstående søjlenummer:

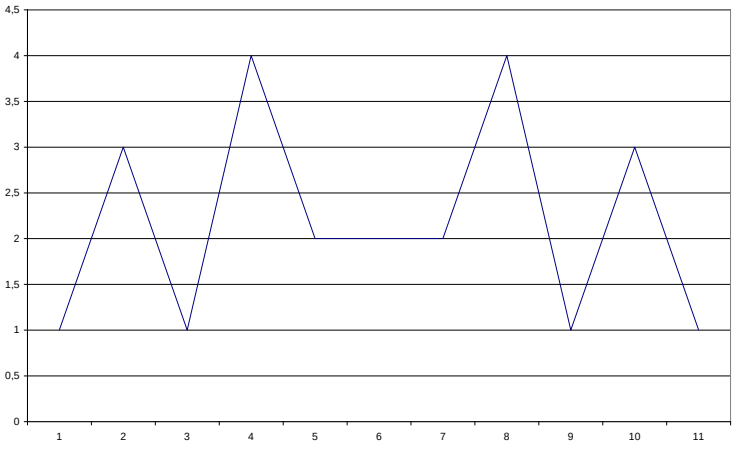
$P = N^3/6 + N^2/2 + N/3 + 1$

Det er interessant at man her har et tredjegrads polynomium med koefficienter der er rene brøker og som alligevel giver heltallige værdier for heltalligt N. Dette gør de for alle heltallige værdier af N hvilket kan kontrolleres ved restklasseberegninger. For tredjegradspolynomier med koefficienter der er brøker af formen 1/S hvor $1 < S < 10$ findes der (ved inspektion i Excel) i øvrigt kun fem der har denne egenskab.

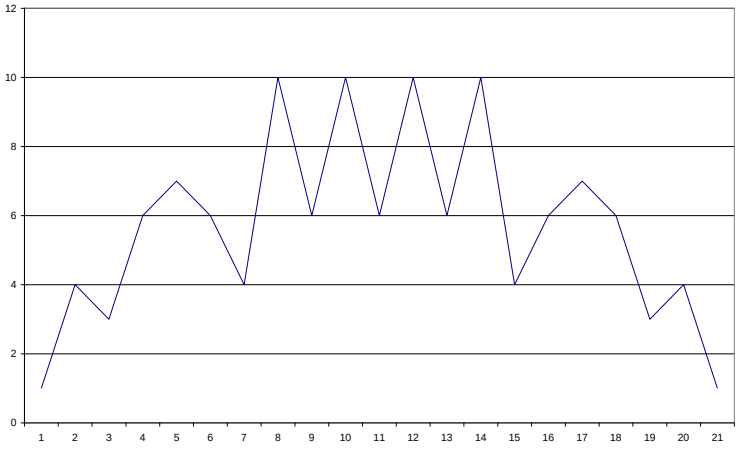
4. Som det fremgår af nedenstående grafer der viser talfølgerne fra N=3 til 8, nærmer de sig en normal fordeling ligesom rækkerne i Pascals trekant. I den sidste figur er en normalfordeling med samme middelværdi og varians som talfølgen indtegnet. Dette synes at være en ny måde at nå frem til normalfordelingen.

5. Selvom det lykkes at finde en systematik for at finde talfølgerne ud fra de(n) foregående, vil der opstå et problem med meget store – og evt. ubegrænsede – hitlister. Skal det tælle lige så meget når et emne bevæger sig 50 pladser op fra 500 til 450 som hvis det havde været fra 100 til 50? Og hvad gør man når emner helt udgår fra hitlisten fx hvis en sort plet er blevet udbedret?

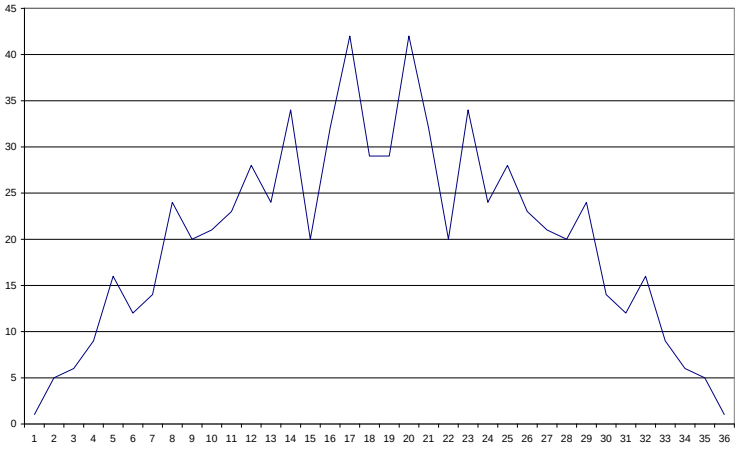
Grafer



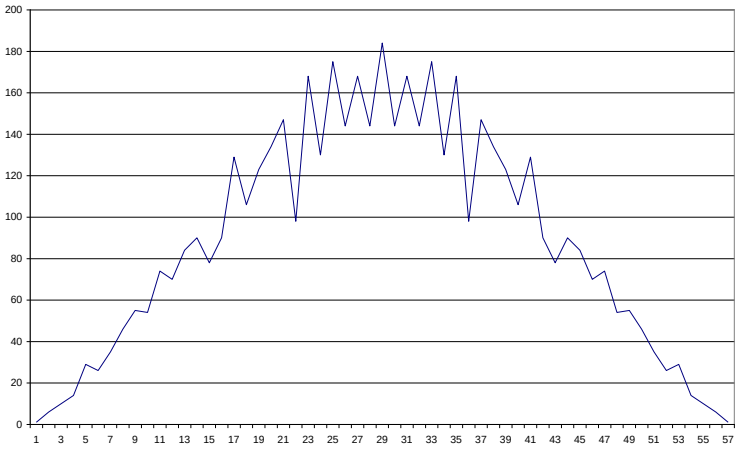
N = 3



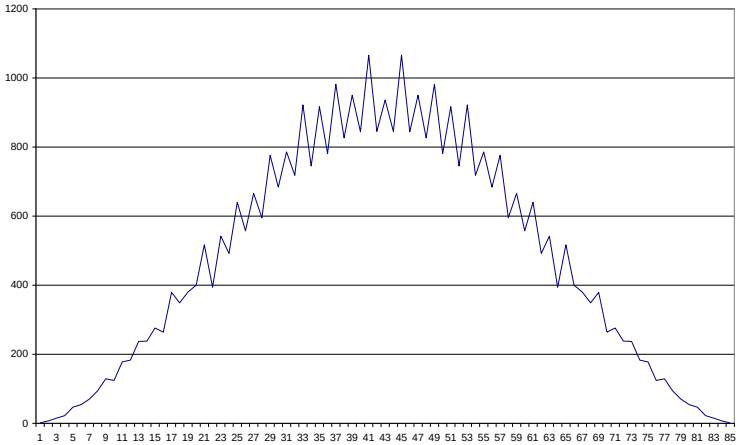
N = 4



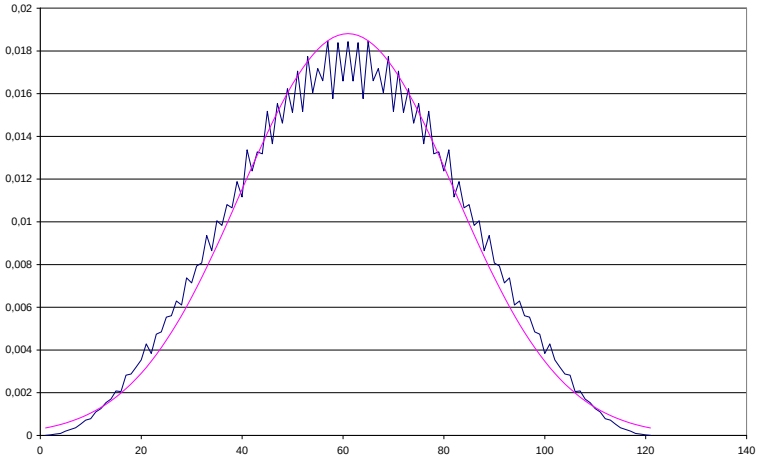
N = 5



N = 6



N = 7



N = 8

Jens Rasmussen
inl_jenr@mail.tele.dk