Fejlfordeling og middelfejl i observationsgrupper.

Af **N. P. Johansen.**

Nedenstående tavle, der er beregnet på grundlag af »Student«s teorem, viser den gennemsnitlige fordeling af de sande fejl ε i forhold til den beregnede middelfejl

$$m = \sqrt{\frac{[v^2]}{n-1}}$$

i en observationsgruppe med n observationer. For $n = 30$ ser man f. eks. at af 1000 fejl vil de 674 ligge i intervallet mellem 0 og 1 m.

n \ Interval	0—m	0—2m	0—3m	0—4m	0—5m
∞	0.683	0.954	0.997	1.000	
30	0.674	0.945	0.995	1.000	
20	0.670	0.940	0.993	0.999	1.000
15	0.666	0.935	0.990	0.999	1.000
10	0.657	0.924	0.985	0.997	0.999
8	0.649	0.914	0.980	0.995	0.998
7	0.644	0.908	0.976	0.993	0.998
6	0.637	0.898	0.970	0.990	0.996
5	0.626	0.884	0.960	0.984	0.993
4	0.609	0.861	0.942	0.972	0.985
3	0.577	0.817	0.904	0.943	0.962
2	0.500	0.705	0.795	0.844	0.874

Det ses, at denne fordeling varierer ret stærkt med observationsantallet n, og ganske særligt giver tallene det indtryk, at man kan vente væsentlig større enkeltfejl ε i grupper med få, end i sådanne med mange observationer. Men selv om dette også med en vis fortolkning kan siges at være rigtigt, er billedet dog misvisende, hvad en dybere undersøgelse viser.

Materialet til en sådan undersøgelse fremskaffes ved at tænke sig en observationsgruppe med n observationer gentaget uendelig mange gange under ensartede forhold. I dette uendelig store materiale kan man nu forfølge og fastlægge i form af formler det gennemsnitlige forløb af de forhold og elementer, man ønsker oplysning om. Undersøgelsen kan gennemføres alene på grundlag af forudsætningen om, at fejlene er tilfældige og observationsnøjagtigheden overalt den samme, karakteriseret ved sand middelfejl σ. Behandlingen er ret omstændelig og kræver i udstrakt grad anvendelse af den højere matematik.

Undersøger vi således den gennemsnitlige fordeling af de sande fejl ε indenfor de enkelte grupper, så viser det sig, at denne nøje følger den for alle tilfældige fejl gældende eksponentielle fordelingslov, uanset om der er få eller mange observationer i gruppen. Dette vil altså sige, at i gennemsnit fordeler fejlene sig i de enkelte grupper ganske som en enkelt fejl ε fordeler sig, når den bliver gentaget uendelig mange gange. Heri ligger klart udtalt, at vi ikke har grund til at befrygte større enkeltfejl i grupper med få, end i sådanne med mange observationer. Når det ikke desto mindre af tallene i tavlen ser ud, som om såvel fejlfordelingen som fejlenes størrelse er afhængige af observationsantallet n, så må dette have sin årsag i andre forhold.

Vi kan dernæst undersøge, hvorledes den af de enkelte grupper beregnede middelfejl m fordeler sig efter størrelse langs den uendelig lange kæde af observationsgrupper. Her er forholdene mere komplicerede.

En middelfejl m er i sig selv en positiv størrelse; dens værdi må derfor altid ligge i området mellem 0 og $+\infty$, og enhver værdi i dette område vil være repræsenteret i den uendelige kæde af m-værdier, der foreligger til undersøgelse. Den sande værdi σ betragtes som normalværdi, og ethvert m, der er mindre end σ, altså beliggende i størrelsesområdet mellem 0 og σ, betegnes som lille, medens et m, der er større end σ, anses for stor. Begreberne stor og lille må altså ses i relation til σ.

Fordelingsloven for m er:

$$F\,(m) \triangle m = \frac{(n-1)^{\frac{n-2}{2}}}{2^{\frac{n-3}{2}} \cdot \Gamma\left(\frac{n-3}{2}\right)} \cdot \left(\frac{m}{\sigma}\right)^{n-2} \cdot e^{-\frac{m^2(n-1)}{2\sigma^2}} \cdot \left(d\left(\frac{m}{\sigma}\right)\sqrt{n-1}\right)^{*)} \qquad (1)$$

Populært fortolket angiver formlen det antal gange, en bestemt m-værdi gentager sig i den foreliggende uendelige m-række. Iøvrigt kan man af formlen aflede forskellige specialoplysninger angående m. Man kan således bestemme gennemsnitsværdierne $\{m^2\}$ og $\{m\}$ af henholdsvis middelfejlkvadratet og middelfejlen i første potens.

For gennemsnitsværdien af middelfejlkvadratet fås

$$\{m^2\} = \sigma^2 \qquad (2)$$

uanset om grupperne er store eller små, altså uafhængigt af observationsantallet n. Dette resultat styrker tilliden til middelfejlkvadratet som nøjagtighedskriterium. Mindre gunstigt stiller det sig med gennemsnitsværdien af m i første potens. Denne er

$$\{m\} = \sigma \cdot \frac{\sqrt{2}}{\sqrt{n-1}} \cdot \frac{\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n-3}{2}\right)} \qquad (3)$$

og er altså varierende med n.

For n =	20	bliver	$\{m\}$ =	0.987. σ
»	10	»	»	0.973. σ
»	5	»	»	0.940. σ
»	3	»	»	0.886. σ
»	2	»	»	0.798. σ

Gennemsnitsværdien af middelfejlen i første potens bliver altså mindre og mindre, efterhånden som observationsantallet aftager. Dette vil atter sige, at middelfejlen i al almindelighed har en tendens til at blive for lille, når observationsantallet er ringe. Dette forhold er af væsentlig betydning for forståelsen af tallene i tavlen pag. 343.

[1]) Den i formlen indgående størrelse $\Gamma\left(\frac{n-3}{2}\right)$ er en såkaldt gammafunktion. Det er en særskilt art af funktioner, defineret ved $\Gamma\,(p) = \int_0^{+\infty} x^p \cdot e^{-x} \cdot dx$.

Det kan måske synes underligt, at $\{m^2\}$ kan være lig σ^2, når $\{m\}$ er forskellig fra σ, men heri er intet mærkeligt, og det er i god overensstemmelse med den almengyldige læresætning, at $f(\{x\})$ almindeligvis er forskellig fra $\{f(x)\}$.

Det fremgår af fordelingsloven, at for store værdier af n vil hovedmassen af de beregnede middelfejl gruppere sig nogenlunde snævert om σ omtrent ligelig fordelt på begge sider af denne, medens værdier, der afviger stærkt fra σ, vel kan forekomme, men er sjældne. Men når n bliver mindre, spredes m-værdierne mere, og samtidig kommer, efterhånden som n aftager, flere og flere af dem til at ligge i størrelsesområdet mellem 0 og σ, bliver altså mindre end σ, medens tilsvarende færre falder i området fra σ og opefter. Det kan således anføres, at

for $n=5$ er 60%
» $n=3$ er 63%
» $n=2$ er 67%

af middelfejlene mindre end σ. Med aftagende n sker der altså en forskydning i værdifordelingen nedad i retning mod nul, hvorved de små m-værdier kommer i overtal i forhold til de store.

Med et lille observationsantal kommer altså færre m-værdier til at ligge i størrelsesområdet fra σ og opefter, men adskillige af disse kan helt vel blive ret store. Man træffer flere helt store m-værdier i de små grupper end i de store. Dette afbøder i nogen grad virkningen af den oven omtalte værdiforskydning nedad for aftagende n, men kan dog ikke helt afbalancere denne, eftersom $\{m\}$ jo bliver mindre og mindre for aftagende n, men det er altså dog tilstrækkeligt til at sikre $\{m^2\}$ en konstant værdi, uafhængig af n.

Det særegne ved m er altså:

a) Gennemsnitsværdien $\{m\}$ bliver mindre med aftagende n

b) Med aftagende n kommer m-værdier mindre end σ i overtal i forhold til de værdier, der er større end σ.

Det er let at se, hvorledes det under a) nævnte forhold må indvirke på fejlfordelingen, når m benyttes som fordelingsmålestok i stedet for σ, og det under b) nævnte forhold virker i samme retning omend på en langt mere indviklet måde. Det er den samlede virkning af disse to forhold, der frembringer den i tavlen anførte fordeling.

Som eksempel er i hosstående tavle angivet den gennemsnitlige værdifordeling af m svarende til de i hovedet anførte observationstal n. Tallene i tavlen angiver, hvormange m-værdier af 1000, der har den i kolonnen længst tilvenstre opførte størrelse angivet i brøkdele af den sande middelfejl σ. Ved f. eks. m = 0.7 σ forstås samtlige m-værdier, der i størrelse ligger mellem 0.65 σ og 0.75 σ.

m \ n	21	11	7	5	3
0.1 σ				1	20
0.2 »			1	6	38
0.3 »			5	18	55
0.4 »		3	17	37	67
0.5 »	1	15	40	61	78
0.6 »	9	44	71	84	84
0.7 »	47	91	104	103	86
0.8 »	132	142	130	114	84
0.9 »	226	176	140	115	80
1.0 »	250	176	135	108	74
1.1 »	187	145	115	95	66
1.2 »	98	100	89	78	57
1.3 »	37	59	63	60	48
1.4 »	10	30	41	44	39
1.5 »	2	13	24	31	32
1.6 »		5	13	20	25
1.7 »		2	8	11	19
1.8 »			3	7	14
1.9 »			1	4	10
2.0 »				2	7
2.2 »				1	3
2.5 σ					1

Til yderligere orientering skal anføres, at »Student«s teorem med en let omskrivning består i fordelingsloven for $\frac{\varepsilon}{m}$. Er n lille, f. eks. 2 eller 3, kan det jo helt vel indtræffe, at de til gruppen hørende ε-værdier alle har samme fortegn, muligvis endog ligger hinanden nær i størrelse. I så fald bliver m lille, kan endog nærme sig til nul, og $\frac{\varepsilon}{m}$ bliver følgelig i tilsvarende grad stor. Men bliver der flere observationer i gruppen, vil konstellationer af den art være sjældnere, og der kommer altså færre helt små værdier af m. Dette forhold, der kommer ind under det under b) omtalte, øver en ret betydelig indflydelse på fejlfordelingen i forhold til m.

Som anført i det foregående har middelfejlen tendens til at blive for lille i grupper med få observationer. Dette kunne lede tanken hen på, at divisoren n—1 i den almindelige middelfejlsformel er for stor og burde ombyttes med n—a, hvor a er et tal større end 1. Dette spørgsmål har flere gange været taget op til overvejelse, men uden at føre til afklaring.

Man kan direkte bevise, at

$$\frac{\{[v^2]\}}{n-1} = \sigma^2,$$

hvad jo kun er en gentagelse af ligning (2) blot afledet ad en anden vej. Dette er et tungtvejende argument til fordel for n—1. Ved at bruge n—a ville man jo vel kunne opnå, at $\{m\}$ formindskedes mindre stærkt med n, ligesom også den »Student«ske fejlfordeling i tavlen pag. 442 ville blive mere afdæmpet, men $\frac{\{[v^2]\}}{n-a}$ eller, hvad der er det samme, $\{m^2\}$ vil så ikke mere blive σ^2, og herved tabes formentlig mere, end der vindes. Divisoren n—1 er afledet direkte ud fra fejlenes væsen og natur, og den almindeligt benyttede middelfejlsformel har derved fået karakter af en naturbestemt funktion, og man skal være varsom med vilkårlige indgreb i en sådan funktion, der derved almindeligvis vil miste sine mest værdifulde egenskaber.