**Korrelasjonen mellom gjennomsnitt og median**

**Nils Lid Hjort, april 2008**

Jostein Lillestøl spør i TG 1/2008 om vi kan finne korrelasjonen mellom empirisk gjennomsnitt og median, (i) når data stammer fra normalfordelingen, og (ii) utenfor normalitetsantagelsen. Dette uskyldig utseende problem er ganske riktig noe kinkig – og det er en grunn til at Jostein deler problemet i (i) og (ii), siden det man kan kalle «en relativt enkel løsning» er mulig kun for normalsituasjonen. For det generelle tilfellet finnes det riktig nok eksplisitte formler, men pene er de ikke, og numeriske beregninger må til, med integraler og summer. Bildet er penere når man lar sin sample-størrelse $n$ bevege mot uendelig, som vi skal se. Vi trenger et minimum av notasjon: våre observasjoner $X_1, \ldots, X_n$ er i.i.d. fra en tetthet $f(x)$, og har gjennomsnitt $\bar{X}$ og median $M_n$. Vi skal utlede en generell formel for grensekorrelasjonen mellom disse to størrelser, altså uttrykt ved $f$, men vi tar normalsituasjonen først.

*For normalfordelte data:* Her gjelder det at vektoren av $X_i - \bar{X}$ er stokastisk uavhengig av $\bar{X}$. Dette kommer av at kovarians null medfører uavhengighet for multinormalt fordelte variable. Men da er også $M_n - \bar{X}$ uavhengig av $\bar{X}$, hvorav følger at

$$\operatorname{cov}(M_n, \bar{X}) = \operatorname{Var} \bar{X} = \sigma^2/n. \tag{1}$$

Skriver vi variansen til $M_n$ som $\tau_n^2/n$ får vi

$$\rho_n = \operatorname{corr}(M_n, \bar{X}) = \frac{\sigma^2/n}{(\sigma/\sqrt{n})(\tau_n/\sqrt{n})} = \frac{\sigma}{\tau_n}. \tag{2}$$

(Om vi ikke visste det fra før, vet vi derfor nå at $\bar{X}$ er mer presis enn $M_n$.) Men det er kjent at $M_n$ er asymptotisk normalfordelt med standardavvik $(\frac{1}{2}/f(\mu))/\sqrt{n}$, der $\mu$ er populasjonsmedianen. For normalfordelingen betyr dette varians $(\pi/2)\sigma^2/n$, altså at $\tau_n$ over vil konvergere mot $\sqrt{\pi/2}\,\sigma$. Korrelasjonen er derfor $\sqrt{2/\pi} = 0.7979$ for stor $n$.

Mer nøyaktig analyse må bruke den eksplisitte tettheten til $M_n$, som for tilfellet $n$ oddetall, si $n = 2m+1$, generelt er lik

$$\frac{(2m+1)!}{m!\,m!} F(x)^m \{1 - F(x)\}^m f(x),$$

der $F$ er den kumulative. Dens eksakte varians blir

$$\frac{\tau_n^2}{n} = \frac{(2m+1)!}{m!\,m!} \int_0^1 \{F^{-1}(u) - F^{-1}(\tfrac{1}{2})\}^2 u^m (1-u)^m \,\mathrm{d}u.$$

Tar man seg bryderiet med å beregne dette integralet numerisk, oppdager man at approksimasjonen $\tau_n \doteq \sigma\sqrt{\pi/2}$ er meget god, selv for temmelig lave $n$.

*Når kan knepet over benyttes?* Vi kom rimelig kjapt frem til svaret $\sqrt{2/\pi}$ over, ved å bruke det hendige faktum at vektoren med $X_i - \bar{X}$ er uavhengig av $\bar{X}$. Det er dette

som medfører (1) og (2). Men grunnlaget for (1) og (2) holder *meget sjelden*, viser det seg. La for enkelhets skyld forventningen være null og la $M(t)$ være den momentgenererende funksjon for $X_i$. Uavhengigheten medfører da

$$\mathrm{E}\exp\{t(X_1-\bar X)+t\bar X\}=\mathrm{E}\exp\{t(X_1-\bar X)\}\,\mathrm{E}\exp(t\bar X)$$

for alle $t$, som gir

$$M(t)=M((1-1/n)t)M(-t/n)^{n-1}M(t/n)^n.$$

Sier vi for enkelhets skyld også at $f$ er symmetrisk (en antagelse man kan kvitte seg med etterpå, med ytterligere analyse), får vi spesielt at $M(t)=M(t/2)^4$. Poenget er at dette etter noe analyse viser seg å holde utelukkende for tilfellet $M(t)=\exp(\frac12\sigma^2t^2)$, som karakteriserer normalfordelingen. – For enhver ikke-normal fordeling må vi altså ty til andre knep og verktøy enn (1) og (2).

*Det generelle tilfellet.* En helt presis formel, om vi for et konkret tilfelle virkelig trenger et eksakt tallsvar, må anvende noe slikt som

$$\mathrm{cov}(\bar X,M_n)=n^{-1}\sum_{i=1}^n\mathrm{cov}(X_{(i)},X_{(m+1)}),\tag{3}$$

sammen med den eksakte simultantetthet for $(X_{(i)},X_{(m+1)})$, der altså $X_{(i)}$ er den $i$-te største observasjon. Dette kan gjennomføres, men blir ikke særlig opplysende.

Alt blir så meget bedre om vi går til grensen. Sentralgrenseteoremet gir $\sqrt n(\bar X_n-\xi)\to_d A\sim\mathrm{N}(0,\sigma^2)$ direkte, der $\xi$ er forventningen, mens man finner det ovennevnte resultatet $\sqrt n(M_n-\mu)\to_d B\sim\mathrm{N}(0,\tau^2)$ i diverse lærebøker, med $\tau=\frac12/f(\mu)$. Det er derimot ikke så lett å finne det *simultane* resultatet vi nå trenger, av typen

$$\begin{pmatrix}\sqrt n(\bar X_n-\xi)\\ \sqrt n(M_n-\mu)\end{pmatrix}\to_d\begin{pmatrix}A\\B\end{pmatrix}\sim\mathrm{N}_2(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\sigma^2&\rho\sigma\tau\\ \rho\sigma\tau&\tau^2\end{pmatrix}).$$

Vi kjenner $A$ alene og $B$ alene, men hvordan samvarierer de? Jostein har altså bedt oss om å finne $\rho$ – der vi kjenner svaret for normalfordelingen, men ikke for andre $f$.

Det beste grepet her blir å se også gjennomsnittet som en operasjon utført på ordningsobservatoren, idet $\bar X_n=(1/n)\sum_{i=1}^n X_{(i)}$, og så anvende teori for ordningsobservatorprosessen. Man har

$$\sqrt n\{F_n^{-1}(t)-F^{-1}(t)\}\to_d q(t)W^0(t)\quad\text{over }(0,1),$$

der $F_n^{-1}$ er den empiriske kvantilprosessen, med $q(t)=(F^{-1})'(t)=1/f(F^{-1}(t))$, og $W^0(\cdot)$ er en såkalt Brownsk bro over $[0,1]$: en Gaußisk prosess med forventning null og kovarians $s(1-t)$ for $s\le t$. Det følger at $B$ kan ses som $q(\frac12)W^0(\frac12)$, mens

$$A=\int_0^1 q(t)W^0(t)\,\mathrm{d}t=\int W^0(F(x))\,\mathrm{d}x.$$

Dette henger sammen med at $\bar X_n$ er nær nok $\int_0^1F_n^{-1}(t)\,\mathrm{d}t$ mens $\int_0^1F^{-1}(t)\,\mathrm{d}t$ er identisk med $\xi=\int xf(x)\,\mathrm{d}x$. Svaret for $B$ krever den milde antagelse at tettheten er positiv i medianen $\mu$.

Nå har vi simultan kontroll over $A$ og $B$, og kan beregne kovariansen

$$\mathrm{E}\,AB = q(\tfrac12)\Big[\int_{-\infty}^{\mu} F(x)(1-\tfrac12)\,\mathrm{d}x + \int_{\mu}^{\infty} \tfrac12\{1-F(x)\}\,\mathrm{d}x\Big]. \tag{4}$$

Svaret på Josteins spørsmål, for en helt generell fordeling, blir dermed

$$\rho = \frac{1}{\sigma}\Big\{-\int_{-\infty}^{\mu} xf(x)\,\mathrm{d}x + \int_{\mu}^{\infty} xf(x)\,\mathrm{d}x\Big\},$$

der jeg har anvendt delvis integrasjon for å forenkle svaret. Når fordelingen er symmetrisk, la oss si rundt null, får man

$$\rho = 2\int_0^{\infty} xf(x)\,\mathrm{d}x \Big/ \Big\{2\int_0^{\infty} x^2 f(x)\,\mathrm{d}x\Big\}^{1/2}. \tag{5}$$

For normaltettheten blir svaret $\sqrt{2/\pi}$, som funnet tidligere.

*Hvilke verdier kan korrelasjonen anta?* Hvordan ser da denne $\rho$ ut, for ulike alternative tettheter? Vi finner $1/\sqrt2 = 0.7071$ for den doble eksponentialfordeling, $\sqrt3/2 = 0.8660$ for den uniforme fordeling, osv. Hvor lav kan den bli? Man kan tvinge (5) ned mot null (tilnærmet uavhengighet mellom gjennomsnitt og median, for store $n$), men essensielt bare på den noe uinteressante måten at man velger tetthet med stor $\sigma$, f.eks. en $t$-fordeling med antall frihetsgrader 2.01. Og hvor høy kan den bli? Bruker man Cauchy–Schwarz på $xf(x)^{1/2}\cdot f(x)^{1/2}$ finner man for det første at $\rho \le 1$, som man vet fra før, men også at $\rho$ når høyest når $xf(x)^{1/2}$ er omtrent proporsjonal med $f(x)^{1/2}$ overalt der det teller, som kan oversettes til at den symmetriske $f(x)$ bør ha mesteparten av sin masse sentrert nær $\pm1$ – men den må på den annen side også ha en positiv verdi i null (ellers har ikke medianen en grensefordeling). Man kan derfor eksperimentere litt med skarpe tri-miksturer av typen

$$f = \tfrac12(1-p)\,\mathrm{N}(-1,\sigma^2) + p\,\mathrm{N}(0,\sigma_0^2) + \tfrac12(1-p)\,\mathrm{N}(1,\sigma^2),$$

med lave standardavvik $\sigma$ og $\sigma_0$ og lav $p$. Da vil $\rho$ komme opp mot sin teoretiske maksimalverdi 1. Her er imidlertid konvergensen $\rho_n \to \rho$ langsom.

*Andre spesialtilfeller.* En formel for den eksakte korrelasjon trenger altså både variansen til $M_n$ og kovariansene i (3). Kun for noen ganske få tettheter kan man finne rimelig eksplisitte formler, og med mitt presumptivt øvede øye ser jeg bare to gode kandidater.

For den *eksponensielle modellen* (der $\bar X$ og $M_n$ har to forskjellige siktemål) leder formel (4) til grensekorrelasjon $\log 2 = 0.6931$, men eksakte formler lar seg utlede med bakgrunn i representasjonen

$$X_{(i)} = E_1/n + E_2/(n-1) + \cdots + E_i/(n-i+1),$$

der $E_i$-ene er i.i.d. og standard eksponensielle; poenget er at mellomrommene $X_{(i)} - X_{(i-1)}$ er uavhengige akkurat for denne fordelingen. Dette leder med litt plunder til $\text{corr}(M_n, \bar{X}) = c_n/\tau_n$, der

$$\tau_n^2 = n \operatorname{Var} M_n = (2m+1)\Big\{ \frac{1}{(2m+1)^2} + \frac{1}{(2m)^2} + \cdots + \frac{1}{(m+1)^2} \Big\},$$
$$c_n = n \operatorname{cov}(M_n, \bar{X}) = \sum_{i=1}^{m} \Big\{ \frac{1}{(2m+1)^2} + \cdots + \frac{1}{(2m+2-i)^2} \Big\}$$
$$+ (m+1)\Big\{ \frac{1}{(2m+1)^2} + \cdots + \frac{1}{(m+1)^2} \Big\}.$$

Det er nå en utmerket gymnastikkøvelse å vise at $\tau_n \to 1$ (ganske raskt) mens $c_n \to \log 2$ (en anelse langsommere).

For den *uniforme modellen* vet vi igjen grensefasitsvaret $\sqrt{3}/2$ (fra (5)), men kan supplere dette med eksakte formler for endelig $n$, via resultatene

$$\mathrm{E}\, X_{(i)} = \frac{i}{n+1} \quad \text{og} \quad \operatorname{cov}(X_{(i)}, X_{(j)}) = \frac{1}{n+2} \frac{i}{n+1} \Big( 1 - \frac{j}{n+1} \Big) \text{ for } i \le j$$

(disse kan utledes direkte, eller via det hjelperesultat at $(X_{(i)}, X_{(j)} - X_{(i)}, 1 - X_{(j)})$ følger en Dirichlet-fordeling med parametre $(i, j-i, n-j+1)$). Dette leder til $\rho_n = c_n/(\sigma\tau_n)$, der $\sigma = \sqrt{1/12}$ er standardavviket til $X_i$, og der

$$\tau_n^2 = n \operatorname{Var} M_n = \frac{n}{n+2} \frac{1}{4},$$
$$c_n = \sum_{i=1}^{n} \operatorname{cov}(X_{(i)}, X_{(m+1)}) = \frac{1}{n+2} \Big\{ \sum_{i=1}^{m} p_i \tfrac{1}{2} + \tfrac{1}{2}\tfrac{1}{2} + \sum_{i=m+2}^{n} (1-p_i)\tfrac{1}{2} \Big\} = \frac{1}{4} \frac{m+1}{n+2}.$$

Sluttresultatet blir

$$\rho_n = \frac{\sqrt{3}}{2} \frac{n+1}{\sqrt{n(n+2)}}.$$

Her er $p_i = i/(n+1)$, og $n = 2m+1$ er altså odde.

PS: Jeg skrev først til Jostein i en mail at jeg hadde kommet frem til svaret $\sqrt{2/\pi}$ på den proverbiale «baksiden av en konvolutt» – men da brukte jeg det generelle resonnementet, og formelen (4), ikke dette knepet forbundet med (1) og (2) som altså gir regning som får plass på en enda mindre konvolutt, for normalsituasjonen. Det kom jeg først på etterpå ...