# Korrelationens betydelse vid GUM-analyser

Hela konceptet GUM genomsyras av antagandet att ingående mätningar är okorrelerade. Guiden betonar i och för sig att ev. korrelation spelar in, men ger inte mycket vägledning för hur man då ska gå tillväga. I denna PM redovisas några enkla metoder för korrelationsanalys samt de korrektioner som det vid påvisad korrelation blir nödvändigt att göra.

Texten är praktiskt inriktad – bakomliggande teoriresonemang redovisas endast summariskt. Scenariot är en serie mätningar i tidsföljd, en s.k. tidsserie, vars väntevärde och mätosäkerhet är okända och ska skattas. Samma tidsavstånd mellan mätningarna förutsätts.

Ett numeriskt exempel illustrerar beräkningsgången och ytterligare beräkningsdetaljer redovisas i ett appendix.

## *Okorrelerade mätningar – t-fördelningen*

Låt oss utgå från okorrelerade mätningar och det förfarande som då tillämpas. Väntevärdet $\mu$ skattas med medeltalet $\bar{x}$. Standardosäkerheten skattas med den traditionella standardavvikelsen

$$s = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1)}$$

där $n$ är antalet mätningar och $n-1$ är antalet frihetsgrader (överbestämningar). Det är den enskilda mätningens standardosäkerhet. Medeltalets standardosäkerhet blir

$$u(\bar{x}) = s / \sqrt{n}$$

och ur t-fördelningen får vi täckningsfaktorn

$$k_\alpha = t_\alpha(n-1)$$

där $\alpha$ är konfidensnivån. Ett $\alpha$ %-igt konfidensintervall för väntevärdet ges sedan av uttrycket

$$P\{\mu \in \bar{x} \pm k_\alpha * u(\bar{x})\} = P\left\{\mu \in \bar{x} \pm t_\alpha(n-1) * s / \sqrt{n}\right\} = \alpha\%$$

Det är samma sak som att säga att den utvidgade mätosäkerheten för medeltalet är

$$U_\alpha(\bar{x}) = k_\alpha * u(\bar{x}) = t_\alpha(n-1) * s / \sqrt{n}$$

Exempel: Beräkna den utvidgade mätosäkerheten för medeltalet av följande mätserie, under antagandet att mätningarna är okorrelerade. Tillämpa konfidensnivån 95%.

| | | | |
|---|---|---|---|
| 10,090 | 10,003 | 9,947 | 9,864 |
| 10,063 | 10,012 | 9,955 | 9,895 |
| 10,083 | 10,110 | 9,954 | 9,917 |
| 10,043 | 10,133 | 9,884 | 10,020 |
| 9,987 | 10,023 | 9,921 | 10,096 |

(Kolumn 1 är mätning nr. 1-5, kolumn 2 är mätning nr. 6-10 osv.)

Mätmaterialet ger $\overline{x} = 10,000$, $s = 0,081$, $n = 20$ och 19 frihetsgrader. Det innebär att

$$u(\overline{x}) = s / \sqrt{20} \approx 0,018$$

$$k_{95} = t_{95}(19) = 2,093$$

Medeltalets utvidgade mätosäkerhet på 95% konfidensnivå blir alltså

$$U_{95}(\overline{x}) = k_{95} * u(\overline{x}) \approx 0,038$$

dvs.

$$P\{\mu \in 10,000 \pm 0,038\} = 95\%$$

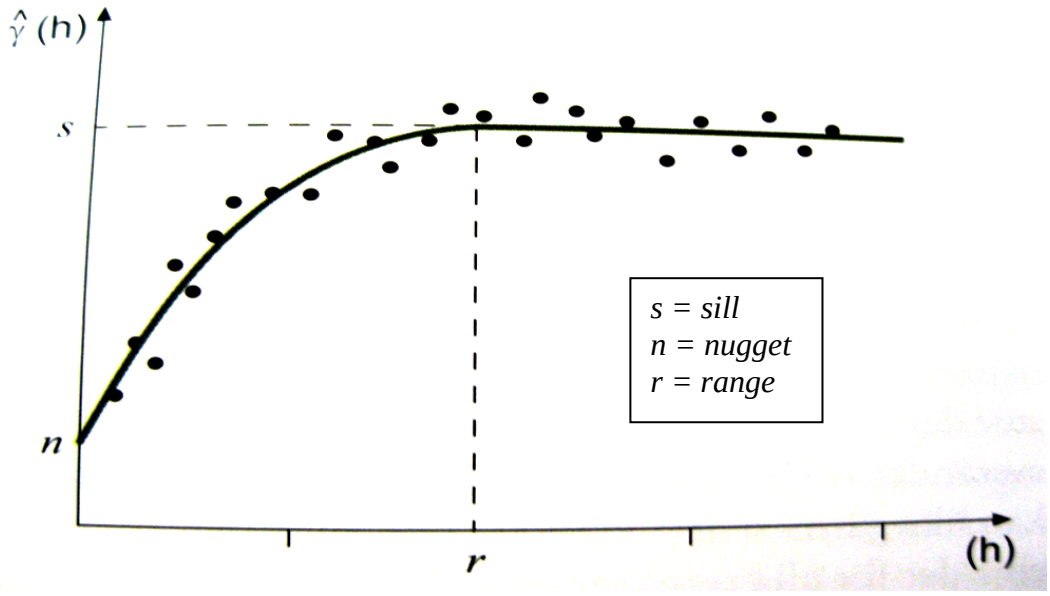
## *Korrelerade mätningar*

Vad händer då om mätningarna är korrelerade?

I vårt scenario betraktar vi den tidsmässiga korrelationen, dvs. korrelationen mellan mätningar i tidsserien som ligger nära varandra i tiden.

Korrelation innebär att resultaten av närliggande mätningar tenderar att följas åt på ett mer eller mindre regelbundet sätt. Den brukar vara som störst på korta avstånd, för att därefter successivt avta och så småningom upphöra helt. Den centrala storheten i dessa sammanhang är korrelationens *räckvidd* eller ”verkningsområde” (eng. *range*), dvs. det (tids)avstånd som krävs för att mätningarna ska kunna betraktas som okorrelerade.

Ett sätt att bestämma räckvidden är att beräkna semivariansen för olika tidsavstånd och upprätta ett *experimentellt variogram* (se GIB, avsnitt 8.4.7, sid. 177 varifrån nedanstående figur är hämtad). Ett annat sätt är att bestämma den s.k. *kovarians*- eller *korrelationsfunktionen*.




Om vi utgår från att vi (på det ena eller andra sättet) har bestämt räckvidden – här betecknad $\Omega$ – så får vi följande modifierade formler för skattning av standardosäkerhet och utvidgad mätosäkerhet.

Det ”effektiva antalet mätningar” ges av

$$n_* = n / \Omega$$

dvs. antalet okorrelerade mätningar som ingår i tidsserien. Om $n_* < 5$ så är fortsatt analys meningslös – mätmaterialet är för litet. Det ”effektiva antalet frihetsgrader” blir $n - \Omega$.

Med dessa förändrade ingångsdata kan hela analysapparaten från föregående avsnitt modifieras till att även gälla korrelerade data.

Väntevärdet $\mu$ skattas fortfarande med medeltalet $\bar{x}$. Med det effektiva antalet frihetsgrader i stället för $n-1$ övergår skattningen av standardosäkerheten till

$$s_* = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 / (n - \Omega)} = s\sqrt{\frac{n-1}{n-\Omega}}$$

Medeltalets standardosäkerhet blir

$$u_*(\bar{x}) = s_* / \sqrt{n_*} = s\sqrt{\frac{n-1}{n-\Omega}}\sqrt{\frac{\Omega}{n}} = \frac{s}{\sqrt{n}}\sqrt{\frac{\Omega(n-1)}{n-\Omega}} = u(\bar{x})\sqrt{\frac{\Omega(n-1)}{n-\Omega}}$$

och ur t-fördelningen får vi täckningsfaktorn

$$k_\alpha^* = t_\alpha(n - \Omega)$$

där $\alpha$ är konfidensnivån. Ett $\alpha$ %-igt konfidensintervall för väntevärdet ges sedan av uttrycket

$$P\left\{\mu \in \bar{x} \pm k_\alpha^* * u_*(\bar{x})\right\} = P\left\{\mu \in \bar{x} \pm t_\alpha(n-\Omega) * s_* / \sqrt{n_*}\right\} = \alpha\%$$

Det är samma sak som att säga att den utvidgade mätosäkerheten för medeltalet är

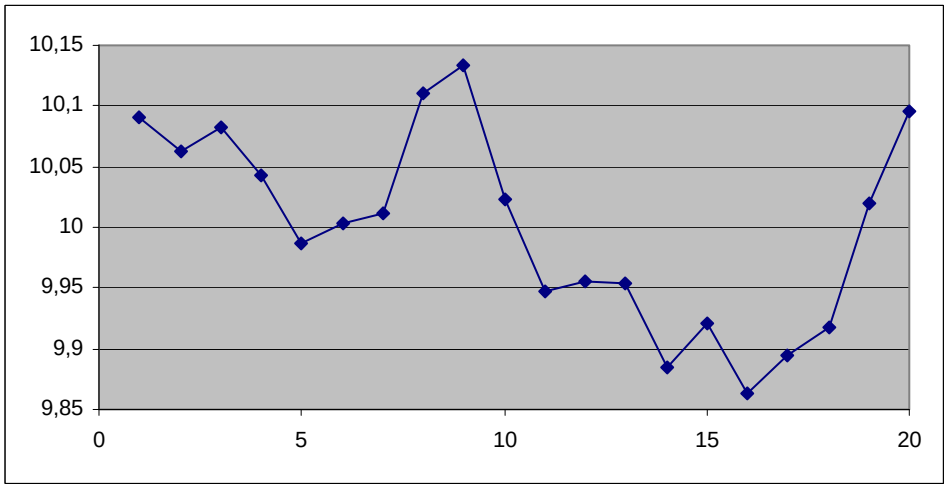
$$U_\alpha(\bar{x}) = k_\alpha^* * u_*(\bar{x}) = t_\alpha(n-\Omega) * s_* / \sqrt{n_*}$$

$\Omega = 1$ innebär okorrelerade data. Då blir samtliga formler identiska med dem i föregående avsnitt.

<u>Exempel:</u> Analysera korrelationen i föregående exempel och modifiera vid behov osäkerhetsskattningarna.

- 

Om vi studerar nedanstående grafiska redovisning av mätningarna så kan vi skönja en viss regelbundenhet – en ”vandring” upp och ned som inte ser helt slumpmässig ut.



Vi börjar med att beräkna korrelationskoefficienten mellan närliggande mätningar (ett alternativt sätt att skatta korrelationens räckvidd).

Vi väljer att göra detta i Excel med funktionen KORREL (se appendix för detaljer). Det ger:

$$\rho_1 = 0,73$$

$$\rho_2 = 0,37$$

$$\rho_3 = 0,11$$

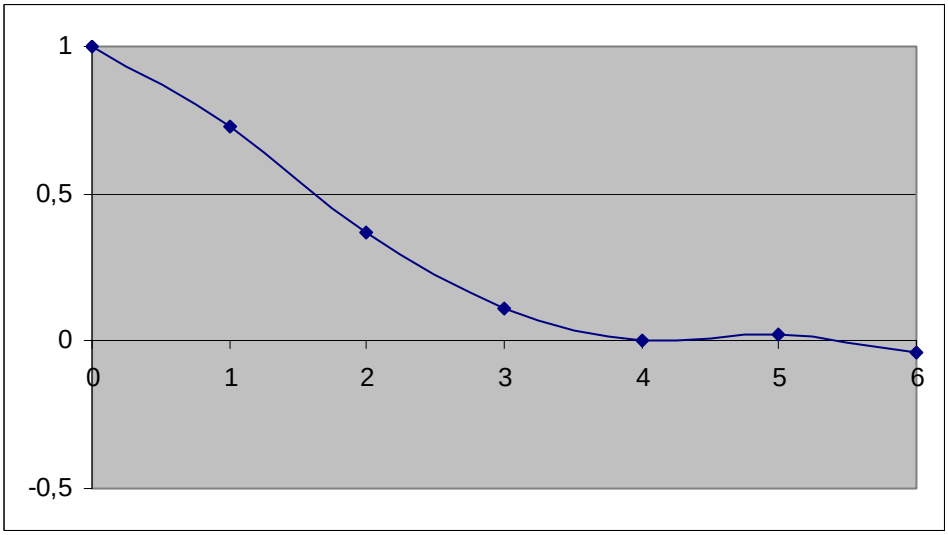
$$\rho_4 = 0,00$$

$$\rho_5 = 0,02$$

$$\rho_6 = -0,04$$

osv.

där indexet anger tidsavståndet mellan mätningarna (1 tidsenhet, 2 tidsenheter etc.). $\rho_0$, autokorrelationen, är definitionsmässigt = 1. Om vi plottar dessa data får vi en approximation av den s.k. korrelationsfunktionen:




Vi ser att korrelationen går ned mot noll vid ungefär 3,5-4 tidsenheter. Låt oss säga 4, dvs. vi får följande uttryck för korrelationens räckvidd

$$\Omega = 4$$

Det i sin tur ger oss $n_* = n / \Omega = 20 / 4 = 5$ effektiva mätningar och $n - \Omega = 20 - 4 = 16$ effektiva frihetsgrader. Vi får vidare

$$s_* = s\sqrt{\frac{19}{16}} = 0,088$$

och

$$u_*(\bar{x}) = s_* / \sqrt{n_*} = u(\bar{x})\sqrt{\frac{4*19}{16}} = 0,039$$

Ur t-fördelningen får vi täckningsfaktorn

$$k_{95}^{*} = t_{95}(16) = 2{,}120$$

Ett 95%-igt konfidensintervall för väntevärdet ges sedan av uttrycket

$$P\left\{\mu \in \overline{x} \pm k_{\alpha}^{*} * u_{*}(\overline{x})\right\} = P\left\{\mu \in 10{,}000 \pm 0{,}083\right\} = 95\%$$

D.v.s den utvidgade mätosäkerheten för medeltalet är

$$U_{\alpha}(\overline{x}) = k_{\alpha}^{*} * u_{*}(\overline{x}) = 0{,}083$$

Den utvidgade mätosäkerheten, liksom konfidensintervallet, förstoras alltså 2,2 gånger p.g.a. korrelationen. Huvuddelen av detta härrör sig från att antalet effektiva mätningar minskar. Det påverkar storleken med en faktor $\sqrt{n / n_{*}} = \sqrt{\Omega} = \sqrt{4} = 2$.

## *Slutord*

Även ganska måttliga korrelationer kan alltså påverka mätosäkerheten betydligt. Det förefaller därför självklart att göra någon form av korrelationsanalys i samband med rapporteringen av mätosäkerhet – inte minst som merarbetet är överkomligt. Endast om man kan visa att någon korrelation inte föreligger kan man utesluta den delen av analysen.

## ***Appendix:* Funktionen KORREL i Excel**

Med funktionen KORREL i Excel kan man beräkna korrelationskoefficienten för olika tidsavstånd.

Lägg de $n$ st. mätningarna i cellerna A1:An i ett excelark. Med hjälp av ”Infoga funktion” anropas funktionen KORREL, och korrelationskoefficienterna för olika tidsavstånd kan beräknas enligt:

$$\rho_i = KORREL(A1 : A(n-i); A(1+i) : An), \quad i = 0, n-1$$

där $i$ är aktuellt tidsavstånd och $\rho_i$ tillhörande korrelationskoefficient. I exemplet ovan var $n = 20$, dvs.

$$\rho_0 = KORREL(A1 : A20; A1 : A20)$$

$$\rho_1 = KORREL(A1 : A19; A2 : A20)$$

$$\rho_2 = KORREL(A1 : A18; A3 : A20)$$

$$\rho_3 = KORREL(A1 : A17; A4 : A20)$$

osv.

En plot av dessa ger en approximation av korrelationsfunktionen och ur denna kan korrelationens räckvidd beräknas – om det nu finns någon korrelation överhuvudtaget. Approximationen är tillfyllest om antalet mätningar är någorlunda stort, dvs. på korta tidsavstånd. Ju större tidsavståndet är desto sämre blir resultatet – vilket för övrigt gäller all korrelationsanalys om mätmaterialet är begränsat.

/Clas-Göran Persson