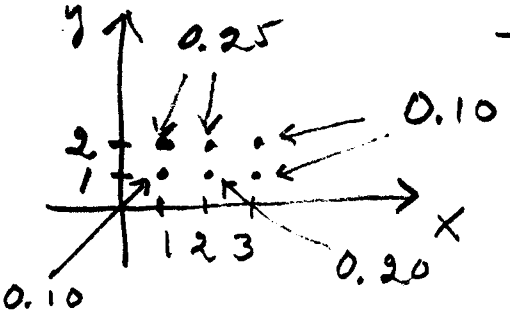
Jfr ex. på sid 177 i NCT. Vi (bl.a.) avser att beräkna korrelationskoefficienten mellan X och Y. $(= \rho(X, Y))$

$$X = \begin{cases} 1 & \text{om personen är mellan 16 och 25 år} \\ 2 & \text{" " " " 26 och 45 år.} \\ 3 & \text{" " " " 46 och 65 år.} \end{cases}$$

$$Y = \begin{cases} 1 & \text{om personen vill köpa varan.} \\ 2 & \text{" " ej " " "} \end{cases}$$



T. ex $p_{X,Y}(1,2) = P((X=1) \cap (Y=2)) = 0.25$

Den simultana sannolikhetsfunktionen för (X, Y)

Vi bestämmer nu des s.k. <u>marginella</u> sannolikhetsf:ter för X och Y:

$$p_X(x) = \begin{cases} 0.35^{*)} & \text{om } x=1 \\ 0.45 & \text{om } x=2 \\ 0.20 & \text{om } x=3 \\ 0 & \text{annars} \end{cases} \qquad p_Y(y) = \begin{cases} 0.40 & \text{om } y=1 \\ 0.60^{*)} & \text{om } y=2 \\ 0 & \text{annars} \end{cases}$$

*) = 0.10 + 0.25 (se figur)

*) = 0.25 + 0.25 + 0.10

<u>Def.</u> X och Y ober $\Leftrightarrow p_{X,Y}(x,y) = p_X(x) \cdot p_Y(x)$ för <u>alla</u> x och y. Här gäller (t.ex.) $p_{X,Y}(1,2) \neq p_X(1) \cdot p_Y(2)$ så X och Y är ej ober.

För att kunna beräkna $\rho(X, Y)$ behöver vi $E(X)$, $E(Y)$, $D(X)$, $D(Y)$ och $E(X \cdot Y)$. Alla dessa beräkningar går tillbaka på principen för att beräkna väntevärde.

$E(X) = 1 \cdot 0.35 + 2 \cdot 0.45 + 3 \cdot 0.20 = \underline{1.85} \ (= \mu_X)$

$E(X^2) = 1^2 \cdot 0.35 + 2^2 \cdot 0.45 + 3^2 \cdot 0.20 = 3.95$

$V(X) = E(X^2) - [E(X)]^2 = 0.5275$. $D(X) = \underline{\sqrt{0.5275}}$

På samma sätt får man $E(Y) = \underline{1.6} \ (= \mu_Y)$, $D(Y) = \underline{\sqrt{0.24}}$

$E(X \cdot Y) = \sum_{x=1}^{3} \sum_{y=1}^{2} x \cdot y \cdot p_{X,Y}(x,y) = 1 \cdot 1 \cdot 0.10 +$

$+ 1 \cdot 2 \cdot 0.20 + 1 \cdot 3 \cdot 0.10 + 2 \cdot 1 \cdot 0.25 + 2 \cdot 2 \cdot 0.25 +$

$+ 2 \cdot 3 \cdot 0.1 = 2.9$. $Cov(X, Y) =$ (se F.S. sid 2) $=$

$= E(X \cdot Y) - \mu_X \cdot \mu_Y = 2.9 - 1.85 \cdot 1.6 = \underline{-0.06}$

$\rho(X, Y) =$ (se NCT) $= \dfrac{Cov(X,Y)}{D(X) \cdot D(Y)} \approx -0.17$.

ρ mäter storleken av det _linjära_ beroendet mellan X och Y. Om X och Y är oberoende så är $\rho(X,Y) = 0$, men omvändningen gäller _inte_. Obs att $-1 \leq \rho \leq 1$ alltid gäller.

Nu kan du räkna tal 5 på tenta 3. (I b. används formeln

$V(aX + bY) = a^2 \cdot V(X) + b^2 V(Y) +$
$+ 2ab \, Cov(X, Y)$ som finns i F.S.

Obs. att $C(X, Y) = 0$ om X och Y är ober vilket ger samma formel som vi har använt tidigare.)

SKATTNINGAR

Vi illustrerar skattning med följande exempel: "Ett okänt avstånd ($=\mu$) skulle bestämmas med hjälp av en avståndsmätare märkt $\sigma = 0.005$." Fyra mätningar av avståndet gav:

1132.155, 1132.158, 1132.145, 1132.163

Vi ser dessa värden som observationer på stokastiska variabler X_1, X_2, X_3, X_4 ($(X_1)_{obs} = 1132.155$ etc) och gör följande (rimliga!) antaganden: X_1, X_2, X_3, X_4 är ober och normalfördelade.

Om mätningarna skett utan systematiskt fel så gäller $E(X_i) = \mu$ $i = 1, 2, 3, 4$ och om vi tolkar 0.005 som deras standardavvikelse så har vi X_i är $N(\mu, 0.005)$. Eftersom X_i:na är ober gäller att $\overline{X} = \frac{1}{4}(X_1 + \cdots + X_4)$ också är N-fördelat. Vi har då (jfr. FS sid 3 längst ner) $\overline{X}$ är $N(\mu, \frac{0.005}{\sqrt{4}})$ d.v.s. $\frac{\overline{X} - \mu}{0.0025}$ är $N(0,1)$. Genom att använda tabell 8 längst ner får vi mer: $P\left(-1.96 < \frac{\overline{X} - \mu}{0.0025} < 1.96\right) = 95\%$

Omformning ger $P\left(\overline{X} - 1.96 \cdot 0.0025 < \mu < \overline{X} + 1.96 \cdot 0.0025\right) = 95\%$. Det vi fått fram är ett s.k. konfidensintervall för μ. 95 % kallas för konfidensgraden.

$\overline{X}$ är den naturliga skattningen av μ och $D(\overline{X}) = \frac{0.005}{\sqrt{4}} = 0.0025$ som kallas för medelfelet mäter noggrannheten i denna skattning. I siffror blir konfidensintervallet $1132.150 < \mu < 1132.160$ (95%). (ty $\overline{X}_{obs} = 1132.155$).

Antag nu att $\sigma = 0.005$ ej funnits angivet. Vi blir då tvungna att skatta det med s. Vi har (se FS sid 1)

$$s^2 = \frac{1}{4-1}\left((1132.155 - 1132.155)^2 + \cdots + (1132.163 - 1132.155)^2\right)$$

varur $s = 0.0076 \approx \sigma$. Med S' betecknande skattningen av σ sedd som stokastisk variabel ser vi nu på $\frac{\overline{X} - \mu}{S'/\sqrt{4}}$. Detta är nu en kvot mellan två stokastiska variabler. När antalet mätningar går mot oändligheten går S mot σ och kvoten blir då $N(0,1)$ men för ett mindre antal mätningar får kvoten en s.k. t-fördelning. Precis som förut får vi nu:

$$P\left(\overline{X} - 3.18 \cdot \frac{0.0076}{\sqrt{4}} < \mu < \overline{X} + 3.18 \cdot \frac{0.0076}{\sqrt{4}}\right) = 95\%$$

se tabell 8 med $\nu = 4-1$

Intervallet blir bredare eftersom vi har mindre information d.v.s. vi visste inte σ utan blev tvungna att uppskatta det.

Om man gör mer än 30 mätningar behöver man inte anta att X_i:na är N-fördelade utan kan använda cgs istället. Denna ger att

$\bar{X} \approx N(\mu, \frac{\sigma}{\sqrt{n}})$ när n är ≥ 30

och man kan fortsätta på samma sätt som tidigare fast nu blir det <u>aldrig</u> någon t-fördelning.

<u>Ex.</u> Ur ett stort varuparti tar man ett stickprov om 600 enh.

X: ant. felaktiga enh. av de 600.

Bilda ett $\approx 95\%$ igt k.i. för p (= prop felaktiga enh. i varupartiet)

Om man inför $U_j = 1$ om j:te enh är felaktiga och $= 0$ annars gäller:

$X = U_1 + \cdots + U_{600}$. Också här kan cgs användas men $n \cdot p(1-p) > 5$ måste gälla för att approximationen ska bli bra. Antag att $X_{obs} = 24$. Då skattas p med $24/600 = 0.04$ och vi får $np(1-p) \approx 600 \cdot 0.04 \cdot 0.96 > 5$.

d.v.s. $X \approx N(np, \sqrt{np(1-p)})$

Betecknar vi nu skattningen av p (sedd som stokastisk variabel) med $\hat{p}$ så gäller $\hat{p} = \frac{X}{n} \approx N(p, \sqrt{\frac{p(1-p)}{n}}) \approx$

$\approx N(p, \sqrt{\frac{0.04 \cdot 0.96}{600}}) = N(p, 0.008)$ och

vi får (jfr. FS sid 4 längst ner)

$I_p: 0.04 \pm 1.96 \cdot 0.008$ $(\approx 95\%)$

d.v.s. $0.024 < p < 0.056$ $(\approx 95\%)$

Räkningarna ovan förutsatte att varupartiet var stort jämfört med stickprovsstorleken. Med "stort" menas vanligen $\frac{n}{N} \leq \frac{1}{10}$ där N är partistorleken. Om detta inte är uppfyllt blir medelfelet inte $\sqrt{\frac{p(1-p)}{n}}$ som ovan utan

istället $= \sqrt{\frac{p(1-p)}{n}} \cdot \sqrt{1-\frac{n}{N}}$.

Den sista faktorn kallas för "ändlighetskorrektion". Denna ändlighetskorrektion används inte bara när det gäller att skatta proportioner utan även i följande fall:

STRATIFIERAT URVAL

Antag att vi vill bestämma medellängden (μ) av alla ($= N$ st) som går på Universitetet. Om vi då tar ett stickprov utan vidare (från N) kanske vi har otur och får med för många män och därmed kommer att överskatta medellängden.

Bättre är därför att dela upp populationen i män och kvinnor och ta två stickprov som sedan vägs ihop. Säg att det går N_1 kvinnor och N_2 män på universitetet och att vi tar ett stickprov av n_1 kvinnor och n_2 män. (Vi kan t.ex. välja n_1 och n_2 prop mot N_1 och N_2, dvs. göra ett s.k. prop urval)

Med $\bar{x}_j$ som medellängden i stickprov j.

Skattar vi μ med $\frac{N_1}{N}\bar{X}_1 + \frac{N_2}{N}\cdot\bar{X}_2$.

Vi har $\hat{\mu} = \frac{N_1}{N}\bar{X}_1 + \frac{N_2}{N}\bar{X}_2$ varur

$$V(\hat{\mu}) = \left(\frac{N_1}{N}\right)^2 \cdot V(\bar{X}_1) + \left(\frac{N_2}{N}\right)^2 \cdot V(\bar{X}_2)$$

där $V(\bar{X}_j) = \frac{\sigma_j^2}{n_j}\cdot\left(1 - \frac{n_j}{N_j}\right)$ om ändlighetskorrektion används.

(Jfr. FS sid 4 överst)