Marginellt kontra betingat oberoende

I samband med analys av korstabeller diskuterades begreppen **marginellt** och **betingat oberoende**. Skillnaden mellan dessa är inte nödvändigtvis enkel att förstå och därför betraktar vi dem med ett ytterligare exempel.

Låt oss studera en familj som består av mamman, pappan och ett barn. Vi är intresserade av dessa tre individers *genotyper* vid ett specifikt lokus i arvsmassan. För enkelhetens skull, betecknar vi de möjliga genotyperna enligt $\{AA, BB, AB\}$. Genotypen motsvarar ett oordnat par av *alleler* som individen ärvt av sina föräldrar, en från modern och en från fadern.

Vi antar nu att mamman och pappan kommer från en stor population av individer, så att deras genotyper kan var för sig betraktas som slumpmässigt dragna ur en sannolikhetsfördelning över genotyper för denna population. Om X betecknar en individs genotyp, kan X nu anta något av värdena $\{AA, BB, AB\}$, med motsvarande sannolikheter $P(AA), P(BB), P(AB)$, som representerar den enskilda genotypens relativa andel i populationen vid en viss tidpunkt.

Föräldrarna är enligt antagandet inte på något sätt 'närbesläktade', de kommer t. ex. inte från samma isolerad by på en avlägsen ort. Då kan vi betrakta föräldrarnas genotyper **marginellt oberoende** från varandra, dvs. det att vi vet att mamman har en genotyp $X_M = AA$, hjälper inte till om vi vill giss pappans genotyp X_P. Uttryckt med sannolikheter, innebär detta att $P(X_M, X_P) = P(X_M)P(X_P)$, dvs om vi nu vet att mammans genotyp X_M är lika med AA, är den betingade sannolikheten för alla tänkbara genotyper för pappan givet X_M densamma som den marginella sannolikheten $P(X_P)$. Alltså, t. ex. sannolikheten $P(X_P = AB|X_M = AA) = P(X_P = AB)$. Vi noterar här att definitionen på den betingade sannolikheten är i sammanhanget

$$P(X_P = AB|X_M = AA) = \frac{P(X_P = AB \ \& \ X_M = AA)}{P(X_M = AA)},$$

där '&' betecknar 'och'. Det marginella oberoendet gör att täljaren blir $P(X_P = AB)P(X_M = AA)$ och därmed förenklas uttrycket till $P(X_P = AB)$ då $P(X_M = AA)$ i täljaren och nämnaren tar ut varandra.

Situationen är annorlunda om vi studerar föräldrarnas genotyper, givet att vi vet barnets genotyp, säg X_B. Låt denna nu vara t. ex. lika med AB. Givet denna information, är föräldrarnas genotyper **inte betingat oberoende** från varandra. Vi ser att detta stämmer om vi tittar närmare på problemet. Vi vet nu att $X_B = AB$, och låt vidare $X_M = AA$. Då vet vi exempelvis att den betingade sannolikheten att pappan också har AA lyder

$$\begin{aligned} P(X_P &= AA|X_M = AA \ \&\ X_B = AB) = \\ &= \frac{P(X_P = AA \ \&\ X_M = AA \ \&\ X_B = AB)}{P(X_M = AA \ \&\ X_B = AB)} = 0, \end{aligned}$$

dvs. det är omöjligt att även pappan har AA då mamman har AA och barnet AB. Ändå är det i allmänhet möjligt för pappor att ha AA om deras barn har AB. Alltså, det att vi får veta mammans genotyp, efter att vi fått kännedom om barnets genotyp, hjälper oss att gissa bättre pappans genotyp jämfört med situationen där vi endast vet barnets genotyp. **Om** genotyperna X_M och X_P skulle vara **betingat oberoende** givet barnets genotyp X_B, kunde vi förenkla de betingade sannolikheterna enligt följande

$$P(X_P = AA|X_M = AA \ \&\ X_B = AB) = P(X_P = AA|X_B = AB), \text{ eller}$$

$$P(X_M = AA|X_P = AA \ \&\ X_B = AB) = P(X_M = AA|X_B = AB),$$

dvs. vi kunde kasta borta en del information från betingelsen, eftersom den inte är till nytta i sammanhanget. Men vi vet nu att föräldrarnas genotyper **inte är betingat oberoende**, givet barnets genotyp, trots att de är **marginellt oberoende**.

Summan av kardemumman är att marginellt och betingat oberoende betyder olika saker och kan inte allmänt härledas från varandra, dvs. betingat oberoende saker **behöver inte** vara marginellt oberoende och vice versa. Marginellt oberoende saker **kan dock** även vara betingat oberoende (eller vice versa) under vissa omständigheter. Förvirrande, eller hur?