SJÄLVSTÄNDIGA ARBETEN I MATEMATIK

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET

Från Bertrand till Gardner med Bayes sats

av

Martin Nymark

2023 - No K27

Från Bertrand till Gardner med Bayes sats

Martin Nymark

Självständigt arbete i matematik 15 högskolepoäng, grundnivå

Handledare: Alan Sola

2023

Sammanfattning

Denna uppsats handlar om hur sannolikheter uppdateras när vi erhåller mer information. Med hjälp av Bayes sats analyserar vi flera problem, med diskreta utfallsrum, som bygger på problemet *Joseph Bertands lådparadox* från slutet av 1800-talet. Några olika lösningsförslag diskuteras bl.a. ett kriterium för likelihoods som erhålles från Bayes sats, som när uppfyllt medför att två händelser i en betingad sannolikhet blir oberoende.

I mitten av 1900-talet populariserades genom Martin Gardner en variant av Bertrands lådor, mest känt under namnet *The boy or girl paradox*, ett problem som var mångtydigt formulerat. Två olika lösningar till Gardners problem diskuteras, lösningar som beror på hur den nya informationen erhållits.

Några nyare varianter av Gardners problem analyseras. Förutsatt att den nya informationen erhålles på ett visst sätt, så leder lösningarna av dessa till överraskande sannolikhetsmått. Vi finner att resultaten följer av att händelser som är oberoende blir betingat beroende givet den nya informationen.

Abstract

In this thesis we examine how probability changes when we receive more information. With the help of Bayes' theorem, we analyze several problems with discrete sample spaces that originated from the problem of *Joseph Bertrand's box paradox* from the end of the 19th century. We discuss a few different solutions, one of which is a criterion for the likelihoods derived by Bayes' theorem. When fulfilled, this criterion entails that two events in a conditional probability become independent.

In the middle of the 20th century, a variant of Bertrand's boxes was popularized by Martin Gardner, most commonly known as *The Boy or Girl Paradox*, a problem that was ambiguously formulated. Two different solutions to Gardner's problem are discussed, solutions that depend on how the new information was obtained.

Two newer variants of Gardner's problem are analyzed. Assuming that the new information is obtained in a certain way, the solutions to these lead to surprising probabilities. We find that the results follow from events that are independent becoming conditionally dependent given the new information.

Ett stort tack till min handledare Alan för hans vägledning och för den tid
som han lagt ned på att hjäpa mig med denna text.

Innehåll

1 Inledning

Vid en grundkurs i sannolikhetsteori fick jag erfara att sannolikheten att en tvåbarnsfamilj har två döttrar givet att minst ena barnet är en dotter är 1/3. Min första reaktion var att Tryckfelsnisse varit i farten, självklart måste det rätta svaret vara 1/2 eftersom en dotter är given och det andra barnet har ca 50% chans att vara en dotter. När jag sedan läste motiveringen till svaret kunde jag till en början inte acceptera den. Jag hade initialt en stark intuition att det fans två möjligheter som var lika sannolika för det andra barnet. När jag sedan lärde mig att det fans en tolkning av problemet med svaret 1/2 stämde matematiken igen, det som återstod var att visa att 1/2 var det "rätta" svaret (ett tankesätt som kom att förändras under arbetets gång). Mitt intresse för problemet, känt som *The boy or girl paradox*, ledde sedan fram till denna text i vilken vi kommer att betrakta några varianter av problemet och olika lösningar.

2 Några inledande definitioner och satser

Definitioner och satser i denna uppsats bygger huvudsakligen på teorin i Alm & Brittons *Stokastik* (2008). [1] Vad menas när vi säger att sannolikheten att slå en trea vid ett kast med en rättvis sexsidig tärning är 1/6? Ett kast med tärning kallar vi ett slumpförsök och resultatet utfall. Utfall betecknar vi med ω och mängden med alla möjliga utfall kallar vi utfallsrummet Ω. För ett kast med en sexsidig tärning är således utfallsrummet mängden $\Omega = \{1, 2, 3, 4, 5, 6\}$. De problem som avhandlas i detta arbete har diskret utfallsrum, det vill säga består av en ändlig eller uppräkneligt oändlig mängd. Utfallsrummet har då lika många element som någon delmängd till de naturliga talen. Att en tärning är rättvis innebär att alla utfall är lika sannolika och då säger vi att utfallsrummet har en likformig sannolikhetsfördelning.

Händelser är en specificerad mängd utfall och betecknas med versaler. Enskilda utfall är också händelser därför används i praktiken ibland versaler även för utfall. Ett sätt att tilldela en händelse ett numeriskt värde är att använda en stokastisk variabel, en funktion som avbildar utfallsrummet på exempelvis mängden av de reella talen.

Definition 1 Stokastisk variabel

En stokastisk variabel $X(\omega)$ är en funktion definierad på utfallsrummet med de reella talen som målmängd.

$$X : \Omega \mapsto \mathbb{R}$$

Vi säger att en stokastisk variabel är diskret om utfallsrummet är ändligt eller uppräkneligt.

Ett sätt att definiera begreppet sannolikhet är som en funktion, ett sannolikhetsmått, som uppfyller Kolmogorovs axiom.

Definition 2 Kolmogorovs axiom

Ett sannolikhetsmått P tilldelar varje händelse A i utfallrummet Ω ett tal $P(A)$ som uppfyller tre axiom.

1. $0 \leq P(A) \leq 1$ för alla händelser $A \subset \Omega$.
2. $P(\Omega) = 1$
3. Om $A \cap B = \emptyset$ så gäller $P(A \cup B) = P(A) + P(B)$.

   För oändliga utfallsrum ersätts 3 med 4.
4. Om $A_1, A_2, ...$ är en oändlig följd av parvis oförenliga händelser sådana att $A_i \cap A_j = \emptyset$ för alla $i \neq j$ så gäller $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Begreppet sannolikhetfunktion kan ha olika innebörd, för att undvika missförstånd definierar vi i denna text $P(.)$ som sannolikhetsmått medan sannolikheten att en (diskret) stokastisk variabel X antar ett vist värde definieras som frekvensfunktionen $p_X(x) := P(X = x)$.

Den klassiska sannolikhetsdefinitionen säger att om utfallsrummet är likfördelat är sannolikheten för en händelse kvoten av antalet gynsamma utfall och antalet möjliga utfall. Låt $n(\cdot)$ står för antalet element i en mängd, dess kardinalitet.

Definition 3 Klassiska sannolikhetsdefinitionen

För ett slumpexperiment med ändligt utfallsrum och med likformig sannolikhetsfördelning gäller at

$$P(A) = \frac{n(A)}{n(\Omega)}.$$

Ett centralt begrepp inom sannolikhetsteori är oberoende. Informellt säger vi att två händelser är oberoende om utfallet av ena händelsen inte påverkar sannolikheten för utfallet av andra händelsen. Sannolikheten att båda utfallen inträffar är då lika med produkten av sannolikheterna för utfallen var för sig. Ett exempel på oberoende händelser är tärningskast. Om vi kastar en rättvis sexsidig tärning har varje utfall sannolikheten 1/6. Det är också sant om vi kastar den igen. Sannolikheten för att få både det första och det andra utfallet i den ordningen blir då 1/36.

Definition 4 Oberoende händelser

Två händelser A och B är oberoende då

$$P(A \cap B) = P(A)P(B).$$

Sannolikheten att en händelse H inträffar givet att en annan händelse E inträffar benämns den betingade sannolikheten för H givet E. Ibland benämner vi händelse E för observationen och använder begreppet a posteriori, med det menas den betingade sannolikheten, i kontrast till den obetingade sannolikheten - a priori observation. Varför vi använder H och E för händelserna kommer framgå i avsnittet om Bayes sats.

Definition 5 Betingad sannolikhet

För $P(E) > 0$ benämns sannolikheten att H inträffar givet att E inträffar som

$$P(H \mid E) = \frac{P(H \cap E)}{P(E)}.$$

En alternativ definition för oberoende händelser som använder betingade sannolikhet är $P(A \mid B) = P(A)$ för $P(A) > 0$.

Det finns flera sätt att beräkna betingade sannolikheter, ett sätt kan vara att direkt använda definitionen, att annat att resonera kring hur utfallsrummet och dess fördelning uppdateras när vi gör en observation, ett tredje är att använda Bayes sats vilken kommer avhandlas i nästa avsnitt.

En viktig sats för beräkning av sannolikheter är lagen om total sannolikhet:

Sats 1 Lagen om total sannolikhet

Låt $A_1, ..., A_n$ vara disjunkta (oförenliga) händelser sådana att $P(A_i) > 0$, $i = 1, ..., n$, som tillsammans utgör hela utfallsrummet d.v.s. att $A_i \cap A_j = \emptyset, i \neq j$, och $\cup_{i=1}^{n} A = \Omega$. Då gäller

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i).$$

Bevis
Vi delar upp händelsen B på de händelser A_i som utfallen ligger i så att även mängderna $B \cap A_i$ blir disjunkta. Vi har

$$B = \cup_{i=1}^{n}(B \cap A_i).$$

Vi använder Kolmogorovs axiom 3 på sannolikhetsmåttet $P(B)$,

$$P(B) = P(\cup_{i=1}^{n}(B \cap A_1))) = \sum_{i=1}^{n} P(B \cap A_i). \quad (1)$$

Användning av definitionen för betingad sannolikhet på högerledet i ekva-

tion (1) ger

$$\sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i). \tag{2}$$

Eftersom vänsterledet i (1) då är lika med högerledet i (2) är beviset klart.

Vi avslutar detta avsnitt med att definiera en viktig diskret fördelning, nämligen binomialfördelningen. I termer tärningskasten i inledningen kan vi använda binomialfördelningen för att beräkna sannolikheten att få k stycken av ett specifikt utfall på n tärningskast.

Definition 6 Binomialfördelning

Låt $0 \leq p \leq 1$ och $q := (1-p)$. Vi säger att en stokastisk variabel är X binomialfördelad och har beteckningen $X \sim \text{Bin}(n, p)$ om den har frekvensfunktionen

$$p_X(k) = P(X = k) = \binom{n}{k} p^k q^{n-k} \quad k = 0, ..., n.$$

3 Bayes sats

Den kanske viktigaste satsen om betingad sannolikhet formulerades av Tomas Bayes i *An Essay Toward Solving a Problem in the Doctrine of Chances* (1764), så betydelsefull att den har gett upphov till en egen gren inom statistik, Bayesiansk statistik.

Sats 2 Bayes Sats

Under samma villkor som för lagen om total sannolikhet, det vill säga om Om $H_1, ..., H_n$ är disjunkta händelser sådana att $P(H_i) > 0, \quad i = 1, ..., n$, som tillsammans utgör hela utfallsrummet så att $H_i \cap H_j = \emptyset, i \neq j$, och $\cup_{i=1}^{n} H = \Omega$ och dessutom $E \cap H_j \neq \emptyset$ för något j så gäller

$$P(H_i \mid E) = \frac{P(H_i)P(E \mid H_i)}{\sum_{j=1}^{n} P(H_j)P(E \mid H_j)}. \tag{3}$$

Bevis
Insättning av definitionen för betingad sannolikhet,

$$P(E \mid H_i) = \frac{P(E \cap H_i)}{P(H_i)},$$

i högerledets täljare i (3) ger

$$\frac{P(H_i)P(E \mid H_i)}{\sum_{j=1}^{n} P(H_j)P(E \mid H_j)} = \frac{P(E \cap H_i)}{\sum_{j=1}^{n} P(H_j)P(E \mid H_j)}. \tag{4}$$

Användning av Lagen om total sannolikhet på nämnaren i (4) ger

$$\frac{P(E \cap H_i)}{\sum_{j=1}^{n} P(H_j)P(E \mid H_j)} = \frac{P(E \cap H_i)}{P(E)}$$

och av definitionen av betingad sannolikhet följer slutligen

$$\frac{P(E \cap H_i)}{P(E)} = P(H_i \mid E). \tag{5}$$

Eftersom högerledet i (5) då är lika med vänsterledet i (4) så är beviset klart.

Bayes sats handlar om att uppdatera vad vi tror när vi gör nya observationer. I definitionen för betingad sannolikhet införde vi händelserna H och E som i kontext av Bayes Sats står för hypotes respektive evidens. Det enklaste fallet för beräkning av betingad sannolikhet med Bayes sats är då vi har ett sannolikhetsmått som sedan uppdateras vid en observation. Låt os säga att vi vet att en tärning är rättvis så sannolikheten att kasta en trea är 1/6. Får vi veta att ett kast med tärningen ger ett udda utfall uppdateras sannolikheten att utfallet blev en trea från 1/6 till 1/3 enligt

$$P(\text{trea} \mid \text{udda}) = \frac{P(\text{udda} \mid \text{trea})P(\text{trea})}{P(\text{udda})} = \frac{1(\frac{1}{6})}{\frac{1}{2}} = 1/3.$$

Satsen låter oss invertera betingade sannolikheter i den meningen att den ger ett uttryck för sannoliksmåttet för evidensen givet hypotesen. I viss meningen handlar Bayes sats om proportioner. Betrakta de tre sannolikhetsmåtten:

- $P(H)$ - Sannolikheten att hypotesen är sann före observation av evidens. Ett sätt att beskriva denna sannolikhet är med en stokastisk variabel Θ definierad på utfallsrummet så att $\Theta : \Omega \to \mathbb{R}$ och som har en så kallad a-priori-fördelning. Denna fördelning kan vara grundad på tidigare observationer eller subjektiva gissningar. Om man inte har fog för att göra någon gissning före observation av data kan man låta den ha likformig fördelning. En frekvensfunktion (i det diskreta fallet) $p_\Theta(\theta) = P(\Theta = \theta)$ representerar då sannolikheterna för hypoteserna.

- $P(H \mid E)$ - Sannolikheten att hypotesen är sann efter observation av evidens. Denna sannolikhet beskrivs av a-posteriori-fördelningen som kan betraktas som en uppdatering av a-priori-fördelningen givet observation av data.

- $P(E \mid H)$ - Den betingade sannolikheten för evidensen givet att hypotesen är sann. Inom Bayesiansk statistik kallas denna sannolikhet för likelihoodfunktionen. Likelihoodfunktionens fördelning skiljer sig från a-priori-fördelningen och a-posteriori-fördelningen genom att Θ är fixerat, vi betingar här på $\Theta = \theta$. Likelihoodfunktionerna hör till hypoteserna men kan vara onormerade, de behöver alltså inte summeras till 1 även fast hypoteserna utgörs av uttömmande och disjunkta händelser.

Låt oss betrakta ett exempel som visar hur fördelningen för Θ förändras när vi observerar data, en så kallad Bayesiansk uppdatering. Exemplet bygger på teorin i texten *Bayesian Updating with Discrete Priors* (2014) [3] av Bloom & Orloff, medan notationerna delvis är tagna från texten *Bayesianska metoder* från KTH. [12] Vi betraktar återigen ett tärningsexempel:

> Ponera att vi har fyra sexsidiga tärningar som ser likadana ut och känns likadana. Låt oss säga att vi vet att tre av tärningar är rättvisa och en är viktad så att den visar utfallet 3 dubbelt så ofta. Om vi observerar utfallet 3 vid ett kast med en slumpmässigt vald tärning, hur stora är sannolikheterna att den är en av de rättvisa respektive den viktade tärningen?

Låt H_1 stå för händelsen att tärningen är rättvis, H_2 för händelsen att tärningen är den viktade och E för utfallet vid ett kast med tärningen. Vi kan jämföra de betingade sannolikheterna för hypoteserna $P(H_1 \mid E)$ och $P(H_2 \mid E)$ genom att använda Bayes sats på dem enligt

$$P(H_1 \mid E) = \frac{P(H_1)P(E \mid H_1)}{P(E)} = \frac{(3/4)(1/6)}{5/24} = \frac{3}{5} \tag{6}$$

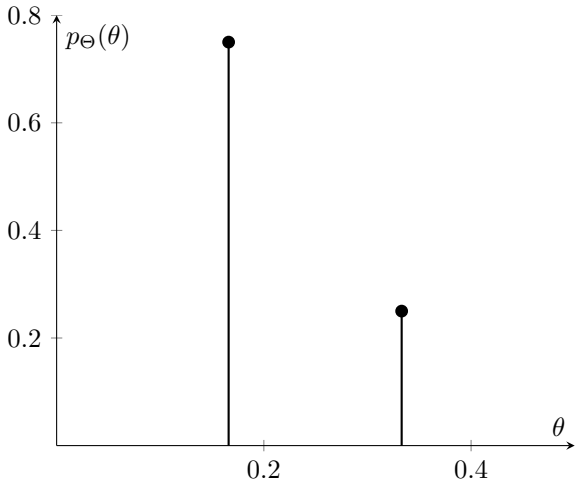
och

$$P(H_2 \mid E) = \frac{P(H_2)P(E \mid H_2)}{P(E)} = \frac{(1/4)(1/3)}{5/24} = \frac{2}{5}. \tag{7}$$

Vi ser att nämnaren som erhålles med satsen om total sannolikhet enbart beror på evidensen och är densamma för båda ekvationerna. Den fungerar som en normaliseringskonstant som gör att de betingade sannolikheterna summeras till ett och det följer att ($\propto$ betyder proportionerlig mot)

$$P(H \mid E) \propto P(H) \cdot P(E \mid H).$$

Likelihoodfunktionen hör till hypoteserna och här använder vi deras värde när vi tilldelar numeriska värden $\theta_i, \quad i = 1, 2$, till H_1 och H_2. Låt $X \sim \text{Bin}(n, \theta_i)$ vara en stokastisk variabel och x en observation av ett utfall från ett kast med tärningen så att $x = 1$ står för utfallet tre och $x = 0$ för inte tre. Före observationen beskrivs sannolikheterna av den a-priori-fördelade frekvensfunktionen $p_\Theta(\theta) := P(\Theta = \theta)$ illustrerad i Figur 1.



Figur 1: A-priori-fördelningen som representerar att det är 75% chans att tärningen är rättvis innan vi kastat den.

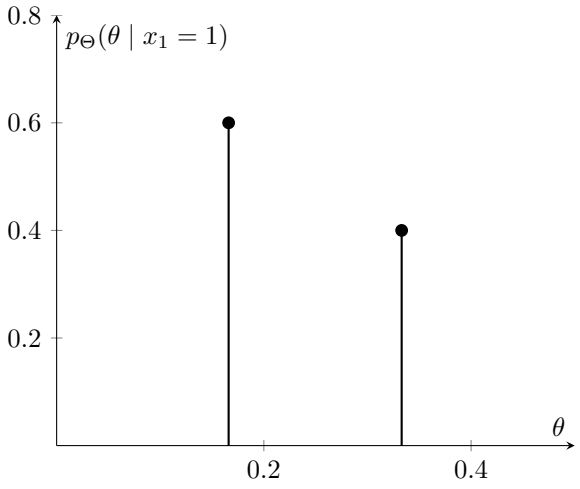
När vi sedan observerar ett utfallet $x = 1$ från ett tärningskast (en trea) uppdateras fördelningen för Θ enligt Bayes sats på samma vis som för sannolikhetsmåtten i ekvation (6) och (7). Likelihoodfunktionen har här beteckningen $p_{X|\Theta=\theta}(x)$ för att förtydliga att den inte är en funktion av Θ utan av X givet $\Theta = \theta_i$.

Figur 2 Visar a-posteriori-fördelningen efter första kastet som updateras enligt (här onormerade)

$$p_\Theta(\theta_1 \mid x_1 = 1) \propto p_\Theta(\theta_1) p_{X|\Theta=\theta}(x_1 = 1 \mid \theta_1) = (3/4)(1/6) = 1/8$$

och

$$p_\Theta(\theta_2 \mid x_1 = 1) \propto p_\Theta(\theta_2) p_{X|\Theta=\theta}(x_1 = 1 \mid \theta_2) = (1/4)(1/3) = 1/24.$$



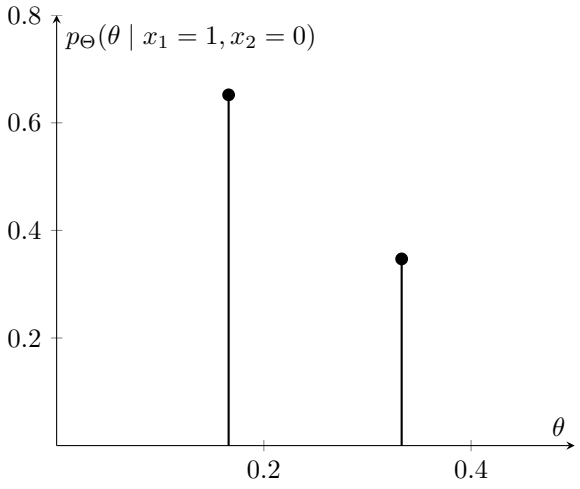
Figur 2: A-posteriori-fördelningen som representerar att sannolikheten minskat till 60% att tärningen är rättvis efter att vi observerat ett kast.

Vi kan uppdatera igen och igen med flera tärningskast. Låt oss säga att nästa tärningskast (med samma tärning) ger utfallet blir $x_2 = 0$ (inte trea). Vi uppdaterar genom att betrakta a-posteriori-fördelningen som vår nya a-priori-fördelning. Eftersom tärningskastens utfall är oberoende händelser kan vi multiplicera täljarna från första uppdateringen med likelehoodfunktionen $p_{X|\Theta=\theta}(x_2 = 0 \mid \theta_i)$ som symboliserar sannolikheten att andra kastet inte blev en trea givet hypoteserna. Den nya a-posteriori fördelningen illustreras i Figur 3 och uppdateras enligt

$$\begin{aligned} p_\Theta(\theta_1 \mid x_1 = 1, x_2 = 0) &\propto p_\Theta(\theta_1) p_{X|\Theta=\theta}(x_1 = 1 \mid \theta_1) p_{X|\Theta=\theta}(x_2 = 0 \mid \theta_1) \\ &= (3/4)(1/6)(5/6) = 5/48 \end{aligned}$$

och

$$\begin{aligned} p_\Theta(\theta_2 \mid x_1 = 1, x_2 = 0) &\propto p_\Theta(\theta_2) p_{X|\Theta=\theta}(x_1 = 1 \mid \theta_2) p_{X|\Theta=\theta}(x_2 = 0 \mid \theta_2) \\ &= (1/4)(1/3)(2/3) = 5/144. \end{aligned}$$



Figur 3: A-posteriori-fördelningen som representerar att sannolikheten ändrats till 65% att tärningen är rättvis efter att vi observerat två utfall från kast med tärningen.

På samma vis kan vi uppdatera fördelningen med n kast med tärningen enligt

$$p_\Theta(\theta \mid x_1, x_2, ..., x_n) \propto p_\Theta(\theta) \prod_{i=1}^{n} p_{X|\Theta=\theta}(x_i \mid \theta).$$

Vi avslutar exemplet med den formella definitionen av likelihoodfunktionen som en produkt, vilken kanske inte hade varit så meningsfull utan kontexten.

Definition 7 Likelihoodfunktionen

Låt $x_1, x_2, ..., x_n$ vara ett slumpmässigt stickprov från en diskret stokastisk variabel X med fördelning $F(x \mid \theta)$. Likelihoodfunktionen definieras som produkten

$$L(\theta) := \prod_{i=1}^{n} p(x_i \mid \theta).$$

Det skall tilläggas att vi fortsättningsvis kommer benämna den betingade sannolikheten för E givet H som likelihooden.

En reflektion vi kan göra från exemplet är att om vi har mycket mätdata blir valet av a-priori-fördelning mindre viktigt. Om vi kastar tärningen tillräckligt många gånger bör vi få reda på om den är rättvis eller viktad oavsett hur den första a-priori-fördelningen var förskaffad.

4 Bertrands lådor

Det första problemet detta arbeta ska avhandla på temat betingad sannolikhet är en föregångare som gett upphov till ett stort antal varianter. Bertands lådor formulerades av Joseph Bertrand (1822-1900) år 1899 i boken *Calcul des probabilités*. En översättning från franska till engelska av publicerades av Brown & Wagenmakers (2021) [4] och följer här:

> "Three boxes have an identical appearance. Each has two drawers and each drawer contains one coin. The coins of the first box are of gold; those of the second box are of silver; the third box contains one gold coin and one silver coin. One chooses a box; what is the probability of finding, in its drawers, one gold coin and one silver coin? There are three cases and these are equally possible because the three boxes have an identical appearance. Only one case is favorable. The probability is 1/3.
>
> A box is chosen. A drawer is opened. Whatever coin one finds, only two cases remain possible. The drawer that remains closed could contain a coin of which the metal either differs or not from that of the first. Of these two cases, one is favorable for the box of which the coins are different. The probability of laying one's hand on that box is therefore 1/2. However, how can it be that opening a drawer suffices to change the probability and increase it from 1/3 to 1/2? The reasoning cannot be correct. And in fact it is not. After opening the first drawer two cases remain possible. Out of these two cases, only one is favorable, that is true, but the two cases are not equally probable. If the coin that one has seen is of gold, the other one can be of silver, but one stands to gain by betting that it is of gold".

Bokstavlig översättning av Bertrands Lådor

Problemet är alltså slumpexperiment där vi har tre utvändigt lika lådor, tre guldmynt och tre silvermynt. Lådorna har två skilda fack vardera, innehållande ett mynt vardera, enligt uppställningen nedan.

En av lådorna har alltså ett mynt av vardera valör och kallas fortsättningsvis den blandade lådan eller GS. Det som söks är sannolikheten att dra den blandade lådan i två fall, obetingat respektive betingat.

- Fråga 1: Vad är sannolikheten att på måfå dra den blandade lådan?
- Fråga 2: Om vi öppnar ett fack och däri observerar ett guldmynt, vad blir då sannolikheten att vi dragit den blandade lådan?

Eftersom de tre lådorna ser likadana ut och ett utfall är gynnsamt följer att svaret till fråga ett blir $P = 1/3$. Ett annat sätt att säga samma sak är att att

slumpexperimentet a priori har en diskret, likformig fördelning. Vi kan betrakta lådorna som utfall i utfallsrummet $\Omega = \{GG, GS, SS\}$. Låt $A_i \subseteq \Omega, i = 1, 2, 3$, stå för de parvis disjunkta händelserna att dra en av de tre lådorna och A_1 stå för händelsen att dra den blandade lådan. Elementet GS representerar då det gynnsamma utfallet. Sannolikheten för händelsen A_1 beräknas genom den klassiska sannolikhetsdefinitionen

$$P(A_1) = \frac{n(A_1)}{n(\Omega)} = \frac{1}{3}.$$

I fråga två handlar det om att beräkna den betingade sannolikheten $P(A_1|B)$ där $B \subseteq \Omega$ står för händelsen att vi får veta att lådan innehåller minst ett guldmynt. De möjliga utfallen för B är $\{GG, GS\}$. Antar vi att båda utfallen är lika sannolika följer av definitionen för betingad sannolikhet att

$$P(A_1|B) = \frac{n(A_1 \cap B)}{n(B)} = \frac{1}{2},$$

men detta är fel, för att a posteriori har slumpexperimentet inte längre en likformig fördelning. Vi observerar valören guld dubbelt så ofta från lådan GG jämfört med från lådan GS men aldrig från lådan SS. Det korrekta svaret blir därför $P(A_1|B) = 1/3$. Vi kan även beräkna sannolikheten med Bayes Sats där nämnaren $P(B)$ erhålles genom användning av Lagen om total sannolikhet.

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{\Sigma_{i=1}^{3} P(A_i)P(B|A_i)} = \frac{\frac{1}{3}(\frac{1}{2})}{\frac{1}{3}(0 + \frac{1}{2} + 1)} = \frac{1}{3} \tag{8}$$

Det är inte svårt att förstå lösningen men att den inte är intuitiv visades av Bar-Hill & Falk (1982) [2] som lät 53 studenter svara på en variant av problemet där lådorna bytts mot hattar och mynten mot spelkort. 35 av studenterna gav det felaktiga svaret $P(A_1|B) = 1/2$.

På grund av symmetrin i problemet spelar det ingen roll vilken valör vi observerar på myntet, så vad betingar vi på egentligen? Bertrand var mycket insiktsfull då han formulerade problemet noggrant. Ponera nu att det inte fanns åtskilda fack i lådorna och att vi råkade se båda mynt när vi observerade ett guldmynt. Då hade sannolikheten för den blandade lådan blivit noll eller ett. Resonemanget leder till insikten att *vilken* valör vi observerade inte påverkar sannolikheten för den blandade lådan medan *hur* vi erhöll informationen om valören kan göra det. Om vi i formuleringen av problemet lämnar utrymme för tolkning riskerar vi att få flera motstridiga lösningar.

Vi avslutar avsnittet med att diskutera den nämnda symmetrin. Låt händelserna A_2 och A_3 stå för att dra lådan GG respektive SS. Man kan tolka Bertrands text som att observationen B inte påverkar sannolikheten för A_1, vilket är sant, men han angav ingen tydlig matematisk motivering. En sådan motivering kan vara att händelserna A_1 och B måste vara oberoende så att $P(A_1 \mid B) = P(A_1)$.

Vi har en symmetri runt den blandade lådan nämligen $P(A_2) = P(A_3)$ och $P(B \mid A_1) = 1/2$, det visar sig att dessa vilkor är tillräckliga men inte är nödvändiga. Från Bayes sats kan vi härleda de både tillräckliga och nödvändiga vilkoren för oberoendet. Från ekvation (8) ser vi att oberoendet uppstår då

$$\frac{P(B \mid A_1)}{\Sigma_{i=1}^{3} P(A_i)P(B|A_i)} = 1. \tag{9}$$

Villkoret för oberoendet blir med omskrivning av (9)

$$P(B \mid A_1) = P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + P(A_3)P(B \mid A_3). \tag{10}$$

Om vi i ekvation (10) isolerar $P(B \mid A_1)$ i vänsterledet och ersätter $1 - P(A_1)$ med $P(A_2)+P(A_3)$ erhåller vi en form som liknar ett viktat medelvärde med en likelihood i vänsterledet och övriga i högerledet, med a priori sannolikheterna som vikter.

$$P(B \mid A_1) = \frac{P(A_2)P(B \mid A_2) + P(A_3)P(B \mid A_3)}{P(A_2) + P(A_3)} \tag{11}$$

Beräkning av högerledet i ekvation(11) ger med $P(B \mid A_2) = 1$ och $P(B \mid A_3) = 0$ (de likelihoods som representerar sannolikheten att vi observerar guld då vi dragit lådan GG eller SS)

$$\frac{P(A_2)P(B \mid A_2) + P(A_3)P(B \mid A_3)}{P(A_2) + P(A_3)} = \frac{\frac{1}{3}(1) + (\frac{1}{3})0}{2(\frac{1}{3})} = \frac{1}{2}.$$

Falk utförde en liknande räkning i texten *A closer look at the probabilities of the notorius three prisoners* (1992) [7] och benämnde samband (11) *The weighted average criterion.*

Vi kan om vi så önskar använda att samband (11) är uppfyllt för att återigen besvara fråga två:

Eftersom

$$P(B \mid A_1) = \frac{P(A_2)P(B \mid A_2) + P(A_3)P(B \mid A_3)}{P(A_2) + P(A_3)}$$

medför att

$$P(A_1 \mid B) = P(A_1)$$

förändras inte sannolikheten att vi dragit den blandade lådan vid observationen av ett guldmynt (eller av symmetri ett silvermynt).

Värt att notera är att oberoendet inte implicerar att vi inte erhållit någon ny information vid observationen ty sannolikheten för komplementhändelsen, att

vi dragit lådan med två guldmynt, ökar till 2/3. Att uppmärksamma detta är en del av lösningen till varianter av problemet, såsom det ökända *Monty Hall problemet* samt *De tre fångarnas problem*. Dessa problem är mindre väldefinierade än Bertrands lådor och sannolikheter och likelihoods återspeglar i dem olika karaktärers bias vilket ger utrymme för olika tolkningar. En variant som liknar Bertrands lådor men med fyra lådor istället för tre kallas i denna text Syskonproblemet och kommer avhandlas i nästa avsnitt.

5 Syskonproblemet

5.1 Klassiska syskonproblemet

Martin Gardner (1914-2010) var en skapare av matematiska pussel som under många år hade en kolumn i tidskriften Scientific American. År 1959 ställde han i tidskriften två frågor som kommit att kallas *The boy or girl paradox*. [8]

- Fråga 1: Herr Smith har två barn. Minst en barnet är en pojke. Vad är sannolikheten att båda barnen är pojkar?
- Fråga 2: Herr Jones har två barn. Det äldsta barnet är en flicka. Vad är sannolikheten att båda barnen är flickor?

När man löser denna typ av problem problem behöver man göra visa underförstådda antaganden, exempelvis brukar man inte ta hänsyn till kulturella eller biologiska aspekter som påverkar fördelningen av pojkar och flickor, det vill säga att vi antar att sannolikheten för båda kön a priori är likfördelade. Gardner publicerade sina lösningar i nästkommande månads utgåva av tidskriften [9]. I svaret till fråga 2 menade Gardner att (låt B står för pojke och G för flicka) utfallen $\{GG, GB\}$ är lika sannolika och därför blir svaret 1/2. Det är av lösningen underförstått att det äldsta syskonet står först i varje utfall. Låt oss införa händelserna C_1 för det äldre och C_2 för det yngre syskonets utfall samt begreppet sannolikhetsmått i Gardners lösning. Vi har ett a priorio likformigt fördelat utfallsrum $\Omega = \{GG, GB, BG, BB\}$ och ett a posteriori likformigt fördelat utfallsrum som reducerats till $\Omega' = \{GG, GB\}$ och ett sannolikhetsmått $P(\{\omega \in \Omega' : C_1(\omega) = G\})$ vars värde beräknas med den klassiska sannolikhetsdefinitionen. Det vanliga skrivsättet för samma sannolikhetsmått är den betingade sannolikheten

$$P(GG \mid C_1 = G) = 1/2.$$

Fråga ett är mer intressant och fortsättningsvis diskuteras endast den. Vid en första anblick kan vi tro, eftersom könen på två barn är oberoende, att svaret blir $P(C_2 = B \mid C_1 = B) = P(C_2 = B) = 1/2$. Undertecknad hade inledningsvis mycket svårt att acceptera att detta är ett felaktigt tankesätt som inte tar hänsyn till informationen om att ena barnet är en pojke. Vi ska titta på två situationer som belyser några intressanta egenskaper hos problemet innan vi diskuterar Gardners svar.

- Situation 1: Antag att vi möter Herr Smith som berättar att han har två barn. Vi frågar Herr Smith om han har någon son och han svarar ja.

Vi vet nu att Herr Smith tillhör populationen tvåbarnsfamiljer utan två döttrar och att ett barn C_i är en son (vi vet inte om det är yngsta eller äldsta barnet). Vi beräknar den betingade sannolikheten att båda barnen är pojkar genom att betrakta det a posteriori reducerade utfallsrummet $\Omega' = \{BB, BG, GB\}$ som vi kan anta ha likformig fördelning så att

$$P(BB \mid C_i = B) = \frac{1}{3}. \tag{12}$$

Det är nu sant att Herr Smith har två barn varav minst ena barnet är en pojke och sannolikheten att det andra barnet är en pojke är $P = 1/3$.

- Situation 2: Antag att vi möter Herr Smith som berättar att han har två barn samt typen på ett av barnen.

Låt I_B vara händelsen att Herr Smith berättar att ena barnet är en pojke. Sannolikheten att han har två pojkar beräknas med Bayes sats enligt

$$P(C_1 = C_2 = B \mid I_B) = \frac{P(I_B \mid C_1 = C_2 = B)P(C_1 = C_2 = B)}{P(I_B)} = \frac{1(\frac{1}{4})}{\frac{1}{2}} = \frac{1}{2}. \tag{13}$$

Om Herr Smith berättar att han har en pojke är det sant att han har två barn varav minst ena barnet är en pojke och sannolikheten att det andra barnet är en pojke är $P = 1/2$.

Vilket svar är då det korrekta till fråga ett? Gardner gav svaret $P = 1/3$ som i situation ett ovan, men han publicerade senare, i oktoberupplagan 1959 [10], efter att många läsare skrivit in till tidningen, ett förtydligande avseende lösningen, han medgav att frågan var mångtydig och att svaret beror på *hur* informationen ”minst ena barnet är en pojke” erhållits. Tänk nu tillbaka på problemet med Bertrands lådor, där var det noggrant formulerat hur informationen erhölls. I syskonproblemet å andra sidan får vi ingen information alls om hur informationen erhölls, så då vi jämför problemen är det föga förvånande att mångtydighet föreligger. Problemet kallas ibland paradoxalt men idag råder konsensus om att frågan är mångtydig på grund av vad Gardner kallade ospecificerad slumpprocedur. [11]

Gardner formulerade två procedurer som han menade genererar informationen ”minst ena barnet är en pojke”, det är dessa som använts för att generera informationen om barnets typ i situation 1 och 2 ovan:

- Procedur 1: Från alla familjer med två barn varav minst ena är en pojke, dras en familj på måfå. Då blir sannolikheten $P = 1/3$.
- Procedur 2: Från alla familjer med två barn dras en på måfå. Om familjen har två pojkar säger informanten ”minst ena barnet är en pojke”. Om

familjen har två flickor säger informanten ”minst ena barnet är en flicka” och om barnen är av olika typ drar han ena typen på måfå och säger att ”minst ena barnet är en ...” och deklarerar typen. Då blir sannolikheten $P = 1/2$.

Notera att i Procedur 1 är ”minst ena barnet är en pojke” inte ett tillräckligt vilkor för svaret $P = 1/3$. I situation 1 gjorde vi dessutom antagande att vi inte vet vilket barn C_1 eller C_2 som är en pojke. Om vi istället definierar sonen som första (eller andra) reducerars utfallsrummet till $\{BB, BG\}$ vilket svarar mot fallet med Herr Jones och sannolikheten blir $P=1/2$. Det är också viktigt att noterar att sannolikheten $P = 1/3$ uppstår då vi ställer en fråga om informationen är sann (Dan 2018) [5] exempelvis genom en enkätundersökning, i kontrast till att få informationen slumpmässigt.

Den berömda ”Paradoxen” i problemet uppstår såhär: Om vi löser problemet i fråga ett under tolkningen där vi frågar om informationen är sann genom att reducerar utfallsrummet får vi svaret 1/3. Sedan noterar vi att a priori är sannolikheten 1/2 att ha två barn av samma typ. Om vi får veta att Herr Smith har minst en son så blir sannolikheten 1/3 för två söner. Hade vi fått veta att Herr Smith har minst en dotter hade sannolikheten för två döttrar på samma vis blivit 1/3. Men då måste sannolikheten ändras från 1/2 till 1/3 bara genom att vi inser att Herr Smith har minst ett barn av någon typ. ”Paradoxen” uppstår alltså om vi tar svaret associerat med procedur ett men informationen om barnet erhålles enligt procedur två.

Syskonproblemet är ett fantastiskt pussel som fortfarande är intressant att diskutera 60 år senare. Den betingade sannolikheten att ha två söner givet att man har minst en son kan alltså tolkas på olika sätt beroende hur informationen erhållits. För att sammanfatta så är $P = 1/3$ tolkningen icke intuitiv i den meningen att vi först tänker på könen som oberoende, vilket gör problemet till ett roligt exempel på betingad sannolikhet. Problemet är paradoxalt i den meningen att det råder diskrepans mellan två lösningar, vilket gör det till ett intressant exempel på vikten att specificera hur information erhålles, begreppens innebörd samt vilka antagen det är rimligt att göra i sannolikhetsproblem.

Martin Gardner var en mycket uppskattad författare och pusselskapare och varje år hålls sammankomsten Gathering 4 Gardner för pussel entusiaster. Sammankomsten går till så att en talare går upp på scenen och får sedan en dollar för varje minut som inte används av den avsatta tiden. En talare på 2010 års GFG håvade in ansenlig mäng endollarsmynt genom att presentera en ny variant av Syskonproblemet (Torrence 2010) [18], då han gick upp scenen och kortfattat sa:

“I have two children. One is a boy born on a Tuesday. What is the probability I have two boys? The first thing you think is ‘What has Tuesday got to do with

it?' Well, it has everything to do with it... "

Vi ska i följande avsnitt betrakta några varianter där Syskonproblemets egenskaper får överraskande konsekvenser.

5.2 Ett exempel med tärningar

Idéen att det fins ett statistiskt samband mellan kön och veckodagar är så absurd att det är något de flesta av oss direkt skulle förkasta. Låt oss återkomma till den efter att vi betraktat följande exempel med tärningar.

- Vi har en urna med lika många blåa som gröna sexsidiga tärningar. En spelare drar en tärning och kastar och lägger sedan tillbaks den i urnan och upprepar en gång. Givet att minst ett av tärningskasten blev en blå tvåa, vad blir sannolikheten båda dragningar gav en blå tärning?

På samma vis som i tidigare problem är hur vi får informationen avgörande för sannolikheten. Om vi frågar en spelare om han slog en blå tvåa och denne svarar "-ja" blir utfallsrummet likfördelat. Enligt den klassiska sannolikhetsdefinitionen söker vi kvoten av antalet gynnsamma utfall och totala antal utfall. Så en ansats är att göra en tabell över utfallsrummet (Marian 2016). [13]

B_2G_1	G_1B_2	B_1B_2	B_2B_1
B_2G_2	G_2B_2	B_2B_2	
B_2G_3	G_3B_2	B_3B_2	B_2B_3
B_2G_4	G_4B_2	B_4B_2	B_2B_4
B_2G_5	G_5B_2	B_5B_2	B_2B_5
B_2G_6	G_6B_2	B_6B_2	B_2B_6

Tabell 1: Utfallsrummet för två tärningskast varav minst ena blev en blå tvåa.

Anledningen till att det fins ett tomt fällt på rad 2 kolumn 4 är för att utfallet B_2B_2 redan fins på rad 2 kolumn 3. De gynnsamma utfallen är gråmarkerade. Sannolikheten blir genom att räkna elementen $P = 11/23$ vilket kan vara överraskande då vi tänker på två tärningskast som oberoende, men kan förklaras med begreppet *betingat oberoende* [19][16].

Definition 8 Betingat oberoende

Två händelser A och B sägs vara betingat oberoende givet händelse C om

$$P(A \cap B \mid C) = P(A \mid C)P(B \mid C).$$

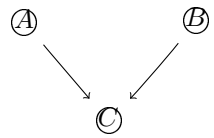
Låt händelserna A, B och C stå för:

A="Första tärningen blev blå"

B="Andra tärningen blev blå"

C ="Minst ena tärningskastet blev en blå tvåa"

Händelserna A och B är oberoende eftersom $P(A)P(B) = P(A \cap B) = 1/4$, men a posteriori, när vi observerat C har vi $P(A \mid C) = P(B \mid C) = 17/23$ och $P(A \mid C)P(B \mid C) \neq P(A \cap B \mid C)$. Definition 8 är inte uppfylld och vi säger A och B är betingat beroende givet C. Händelserna A, B och C bildar ett så kallat Bayesianskt nätverk [17] och illustreras nedan med en riktad acyclisk graf.



Noderna A och B kallas föräldrarna till nod C och att de är oberoende av varandra illustreras med avsaknaden av direkta pilar mellan dem, men det fins ett informationsflöde via C. Resonemanget för det betingade beroendet kallas *explaing away* och med det menas att A och B konkurrerar om att orsaka observationen C. Om vi observerar C och får veta att A är sant minskar sannolikheten att B är sant, och på samma vis om vi observerar C och får veta B är sant minskar sannolikheten att A är sant. Vi har

$$P(A \mid C \cap B) = P(B \mid C \cap A) = 11/17$$

och

$$P(A \mid C) = P(B \mid C) = 17/23.$$

Så att

$$P(A \mid C \cap B) < P(A \mid C)$$

och

$$P(B \mid C \cap A) < P(B \mid C).$$

På samma vis som i tidigare problem är hur vi får informationen avgörande för sannolikheten. Om en spelare gör två kast med tärningar från urnan och sedan istället avslöjar ett av utfallen slumpmässigt och det är exempelvis en blå tvåa blir sannolikheten istället

$$\begin{aligned} P(A \cap B \mid C) &= \frac{P(C \mid A \cap B)P(A \cap B)}{P(C)} \\ &= \frac{\frac{1}{6} \cdot \frac{1}{4}}{\frac{1}{12}} = \frac{1}{2}. \end{aligned}$$

Förklaringen till skillnaden mellan fallen "spelaren nämner" och "vi frågar"är att fördelningen av elementen i utfallsrummet a posteriori är olika. Betrakta utfallsrummet i Figur 1. Om en spelare slumpmässigt nämner att den fått utfallet en blå tvåa är det dubbelt så stor chans att utfallet för båda tärningarna B_2B_2 än för någon enskild av de övriga elementen som innehåller B_2, för vi vet inte om spelaren uppger det första eller andra kastet.

5.3 En son född på en tisdag

I föregående avsnitt beräknade vi sannolikheten för två blåa tärningar givet en blå tvåa och kom fram till $P = 11/23$. Ponera nu att vi istället använde sjusidiga tärningar, då blir problemet isomorft (Zazkis & Wijeratne 2015) [20] med frågan nedan, i den meningen att lösningen har samma struktur fast med andra händelser.

- Herr Smith har två barn varav minst ena är en pojke född på en tisdag, vad är sannolikheten att Herr Smith har två pojkar?

Med samma lösningsprocedur som gav P=11/23 för tärningarna får vi - om vi byter tärningarnas färg mot kön och sidor mot veckodagar

$$P(\text{Två söner} \mid \text{Minst en son född på en tisdag}) = 13/27.$$

Det är klart att samma mångtydighet som präglar det ursprungliga problemet om Herr Smiths söner föreligger här och det fins alltså en tolkning av problemet med svaret P =13/27. Utfallsrummet blir likadant som i Tabell 1 men med sju rader istället för sex. Låt oss kalla den sexsidiga tärningens antal sidor en egenskap av kardinalitet $n = 6$ och en antalet dagar på en vecka en egenskap med kardinalitet $n = 7$. Antalet rader i utfallsrummet blir också n. Genom att betrakta utfallsrummet kan vi notera sambandet [20]

$$P(n) = \frac{2n-1}{4n-1}, \quad n \in \mathbb{Z}^+. \tag{14}$$

I Tabell 2 nedan visas sannolikheter för några egenskaper av olika kardinalitet. Det är inte nödvändigt att egenskapen är av kronologisk karaktär men låt oss börja så, då blir egenskaperna approximativt symmetriska likt i exemplet med tärningen.

Egenskap (Minst en pojke född på)	Kardinalitet	Sannolikhet
Halvår (Sommar/vinter)	$n = 2$	$P = 3/7$
Årstid	$n = 4$	$P = 7/15$
Veckodag	$n = 7$	$P = 13/27$
Månad	$n = 12$	$P = 23/47$
Datum	$n = 365$	$P = 729/1459$

Tabell 2: Sannolikheter för två söner givet att minst en son är född under en viss tidsperiod.

Vi kan formulera följande lemma om sambandet mellan egenskapens kardinalitet och den betingade sannolikheten samt härleda likhet (14) algebraiskt (Rehak 2018). [16]

Lemma 1 Generalisering av fallet då minst en son tilldelas en viss egenskap och informationen erhålles genom att vi frågar om den är sann

Låt $C_i = B$ stå för utfallet att ett barn är en son och B_E för antalet söner med en viss egenskap (utöver kön) som vaje son har sannolikheten $1/n$ att besitta och som båda söner kan besitta samtidigt. Då gäller:

$$P(\{C_1 = C_2 = B\} \mid \{B_E \geq 1\}) = \frac{2n-1}{4n-1}, \quad n \in \mathbb{Z}^+.$$

Bevis
Låt $r \in \mathbb{Q}$ sådant att $0 < r \leq 1$ vara andelen söner som (eller chansen att en son) besitter egenskapen.

Sannolikheten att en tvåbarnsfamilj har två söner varav exakt en besitter egenskapen kan beräknas som produkten av sannolikheten att ha två pojkar med sannolikheten för ett gynnsamt och ett ogynnsamt Bernoulliförsök som avser egenskapen (ett slumpförsök med två utfall). Det gynsamma utfallet kan ske på två sätt - den äldsta eller den yngsta besitter egenskapen. Vi kan uttrycka det algebraiskt som

$$P(\{C_1 = C_2 = B\} \cap \{B_E = 1\}) = 2 \cdot \frac{1}{4} r(1-r).$$

Sannolikheten att en tvåbarnsfamilj har två söner varav båda besitter egenskapen kan skrivas

$$P(\{C_1 = C_2 = B\} \cap \{B_E = 2\}) = P(\{B_E = 2\}) = \frac{1}{4} r^2.$$

Sannolikheten att en tvåbarnsfamilj har exakt en son med egenskapen kan på samma vis uttryckas som ett gynnsamt och ett ogynnsamt Bernoulliförsök som avser kön och egenskapen. Det gynnsamma utfallet kan igen ske på två vis vilket kan uttryckas som

$$P(\{B_E = 1\}) = 2 \cdot \frac{1}{2} r(1 - \frac{1}{2} r).$$

Av definitionen för betingad sannolikhet följer sedan:

$$
\begin{aligned}
P(\{C_1 = C_2 = B\} \mid \{B_E \geq 1\}) &= \frac{P(\{C_1 = C_2 = B\} \cap \{B_E \geq 1\})}{P(\{B_E \geq 1\}))} \\
&= \frac{P(\{C_1 = C_2 = B\} \cap \{B_E = 1\}) + P(\{C_1 = C_2 = B\} \cap \{B_E = 2\})}{P(\{B_E = 1\}) + P(\{B_E = 2\})} \\
&= \frac{2 \cdot \frac{1}{4} r(1-r) + \frac{1}{4} r^2}{2 \cdot \frac{1}{2} r(1 - \frac{1}{2} r) + \frac{1}{4} r^2} \\
&= \frac{r(\frac{1}{2} - \frac{r}{4})}{r(1 - \frac{r}{4})} \\
&= \frac{\frac{2}{r} - 1}{\frac{4}{r} - 1}
\end{aligned}
$$

Insättning av $r = \frac{1}{n}$ ger avslutningsvis

$$P(\{C_1 = C_2 = B\} \mid \{B_E \geq 1\}) = \frac{2n-1}{4n-1}.$$

Vi ser att extremfallen är $P(1) = 1/3$ (när alla söner besitter egenskapen) och $\lim_{n\to\infty} P(n) = 1/2$.

Anledningen till att vi uttryckt P som funktion av n är för att sambandet härleddes ur utfallsrummet i Tabell 1. Det är egentligen mer naturligt att tala om andelen barn med en viss egenskap r, än att tala om en egenskaps kardinalitet, det är inte heller nödvändigt att egenskapen är symmetrisk som sidorna på en tärning eller en veckodag, nästa problem kommer att handla om en mer osymmetrisk egenskap.

5.4 En dotter som heter Florida

En version av syskonproblemet som populariserades genom Leonards Mlodinows bästsäljare *The Drunkards walk: How Randomness Rules Our Life* (2008) [15] handlar om sannolikheten att en familj har två döttrar givet att en dotter heter Florida:

"In a family with two children, what are the chances, if one of the children is a girl named Florida, that both children are girls? Yes, I said a girl named Florida. The name might sound random, but it is not, for in addition to being the name of a state known for Cuban immigrants, oranges, and old people who traded their large homes up north for the joys of palm trees and organized bingo, it is a real name. In fact, it was in the top 1,000 female American names for the first thirty or so years of the last century. I picked it rather carefully, because part of the riddle is the question, what, if anything, about the name Florida affects the odds? But I am getting ahead of myself. Before we move on, please consider this question: in the girl-named-Florida problem, are the chances of two girls still 1 in 3 (as they are in the two-daughter problem)? I will shortly show that the answer is no. The fact that one of the girls is named Florida changes the chances to 1 in 2"

The girl named Florida - Från boken The Drunkard's walk

Det handlar alltså om att besvara de två frågorna:

- Om en tvåbarnsfamilj har minst en dotter, vad är sannolikheten att den har två döttrar?
- Om en tvåbarnsfamilj har en dotter som heter Florida vad är sannolikheten att den har två döttrar?

Mlodinow löser först det klassiska syskonproblemet (i Mlodinows versionen döttrar) med den ansats som leder till sannolikheten $P = 1/3$ (det är underförstått att i Mlodinows modell erhålls informationen genom att vi frågar om den är sann) och visar sedan att sannolikheten ändras till $P = 1/2$ när vi får reda på att en dotter heter Florida. Låt G_F stå för en flicka som heter Florida och G_{F^*} stå för en flicka med annat namn. Det finns a priori 9 möjliga konfigurationer för en tvåbarnsfamilj där flickor definieras som Florida eller inte Florida.

$$\{G_F G_F, G_F G_{F^*}, G_F B, G_{F^*} G_F, G_{F^*} G_{F^*}, G_{F^*} B, BB, BG_F, BG_{F^*}\}$$

Eftersom det är vanligare att flickor inte heter Florida är utfallsrummet inte likfördelat som i tidigare versioner av problemet. Mlodinow menar att eftersom namnet är ovanligt är utfallet $G_F G_F$ approximativt noll, ett argument som kan leda läsaren att tro att resultatet beror på hur ovanligt namnet är, att en tvåbarnsfamilj vanligen inte döper sina barn till samma tilltalsnamn hade kanske varit ett tillräckligt argument - oberoende av namnet.

"The chances of both girls' being named Florida (even if we ignore the fact that parents tend to shy away from giving their children identical names) are therefore so small we are justified in ignoring that possibility."

I Mlodinows modell där två barn aldrig har samma namn blir utfallsrummet a posteriori

$$\{G_F G_{F^*}, G_F B, G_{F^*} G_F, BG_F\}$$

som är likfördelat med två gynnsamma utfall av fyra möjliga så att $P = 1/2$. Mlodinow ger en ytterligare förklaring:

"One way to understand this, if it still seems puzzling, is to imagine that we gather into a very large room 75 million families that have two children, at least one of whom is a girl. As the two-daughter problem taught us, there will be about 25 million two-girl families in that room and 50 million one- girl families (25 million in which the girl is the older child and an equal number in which she is the younger). Next comes the pruning: we ask that only the families that include a girl named Florida remain. Since Florida is a 1 in 1 million name, about 50 of the 50 million one-girl families will remain. And of the 25 million two-girl families, 50 of them will also get to stay, 25 because their firstborn is named Florida and another 25 because their younger girl has that name. It's as if the girls are lottery tickets and the girls named Florida are the winning tickets. Although there are twice as many one-girl families as two-girl families, the two-girl families each have two tickets, so the one- girl families and the two-girl families will be about equally represented among the winners."

Informationen i förklaringen, att sannolikheten för namnet Florida är en på miljonen, kan leda läsaren till att tro att valet av namn har betydelse. I själva verket är andelarna detsamma oavsett valet av namn, vilket uppmärksammades av Marks & Smith (2011) [14]. Mlodinows problem verkar vara detsamma som i Gardners fråga om Herr Jones där döttrarna definierades med en egenskap som båda inte kunde besitta samtidigt, yngst och äldst. Att sannolikheten ökar från $P = 1/3$ till $P = 1/2$ när vi får reda ett namn verkar bero på att det betingade beroendet som skapas från observationen "minst en barnet är dotter" upphör då vi tillskriver något av barnen en egenskapen som båda inte får besitta samtidigt.

I en modell där båda syskon tillåts heta Florida och vi frågar om informationen är sann finns det däremot ett samband mellan sannolikheten P och andelen flickor med namnet r. I En sådan variant visade D'Agostini (2010) [6] att sannolikheten för två döttrar givet en dotter med ett namn som döttrar besitter med sannolikhet r blir

$$P(\text{Två döttrar} \mid \text{Minst en dotter som heter Florida}) \approx \frac{1}{2} - \frac{r}{8} \quad \text{för } r \ll 1.$$

Detta resultat överensstämmer väl Lemma 1 där vi på sista raden i beviset erhöll $P(r) = (\frac{2}{r} - 1)/(\frac{4}{r} - 1)$. En Maclaurinutvecklingen av $P(r)$ ger just

$$P(r) = P(0) + \frac{P'(0)}{1!}(r - 0) + \mathcal{O}(r^2) \approx \frac{1}{2} - \frac{r}{8}$$

då högre ordningens termer kan förkastas om r är litet. Vi har P=1/3 då alla döttrar har samma namn samt $\lim_{r\to 0} P(r) = 1/2$. Sannolikheten ändras till $\approx 1/2$ trots att det betingade beroendet inte upphör, men vi får komma ihåg att $P = 1/3$ bara uppstår i en något konstgjord situation där vi frågar efter informationen, exempelvis en enkätundersökning. $P \approx 1/2$ bör i den här modellen

tolkas som andelen av tvåbarnsfamiljer som svarat ja på frågan ”-Är något av dina barn en flicka som heter Florida?” som har två döttrar och P=1/3 andelen av tvåbarnsfamiljer som svarat ja på frågan ”-Har du någon dotter?” som har två döttrar.

6 Några avslutande kommentarer

Hur bör vi nu tolka resultaten i de föregående avsnitten? Naturligtsvis har vi **inte** funnit ett samband mellan veckodagar och kön. Vi (matematikstudenter) är vana vid entydiga svar, men när det kommer till Syskonproblemets lösningar får vi finna oss i att det handlar om olika tolkningar som kan vara svåra att värdera som rätt eller fel.

I analysen av det klassiska syskonproblemet diskuterade vi mångtydighet och syftade då på två sätt att erhålla information. Men mångtydighet skulle kunna innefatta mycket mer, exempelvis ords innebörd, eller självaste sannolikhetsbegreppet. Förvisso definierar vi ett sannolikhetsmått entydligt m.h.a. Kolmogorovs axiom men det fins ändå utrymme för olika perpektiv och olika metoder. När man gissar oddsen att en händelse ska inträffa talar man om subjektiva sannolikheter, men är det rimligt att diskutera sannolikheten för något som redan har inträffat, så som i Syskonproblemet? Det är egentligen inte sannolikheten för utfallet (Herr Smiths familjekonfiguration) som diskuteras, utan frågan om hur bra gissning vi kan göra beroende på hur mycket information vi har. En tolkning av sannolikhetsbegreppet är den frekventistiska, men utfallet för Herr Smits familj är knappast något som vi kan upprepa flera gånger. I kontrast till den frekventistiska tolkningen har vi Bayesiansk statistik. Den korrekta metoden i en Bayesiansk analys är att använda Bayes sats utan att reducera utfallsrummet (Marks & Smith 2011) [14]. Är Bayesiansk analys då den bästa metoden? En datavetare skulle å andra sidan kanske invända mot Gardners procedur två som ett konstigt sätt att tolka informationen (Marian 2016) [13], d.v.s. om informanten säger pojke är det inte självklart att tolka det som att det förelåg en slumpprocedur med 50% chans att informanten skulle säga flicka. En språkvetare skulle kanske påtala att föräldrar inte beskriver sin familj med fraser som ”-Jag har minst en dotter. ” och att problemet därför har liten verklighetsförankring, eller att det inte går att översätta problemen helt exakt mellan olika språk utan att förändra innebörden. En psykolog är kanske mer intresserad av hur svarsfrekvens förändras när man byter ut ord i problemen. Och när är någons dotter definierad som den ena eller den andra i filosofisk mening?

Med det sagt bör vi kanske inte betrakta lösningarna i denna text med så stort allvar, det handlar ju trots allt om matematiska tankenötter. Problemen har dock varit otroligt intressanta att studera.

Referenser

[1] Alm, S.E. & Britton, T., (2008). *Stokastik* 1:a upplaga 3. Liber.

[2] Bar-Hillel, M. & Falk, R., (1982). Some teasers concerning conditional probabilities. *Cognition*, 11(2), 109–122. https://doi.org/10.1016/0010-0277(82)90021-X

[3] Bloom, j. & Orloff, J. (2014). Bayesian Updating with Discrete Priors. https://ocw.mit.edu/courses/18-05-introduction-to-probability-and-statistics-spring-2014/139aff133c83fe5e4335f942b0ea353c_MIT18_05S14_Reading11.pdf

[4] Brown, N. & Wagenmakers, E. (2021). Literal and Liberal Translations of Bertrand's Box Paradox. *Bayesian-Spectacles.* https://www.bayesianspectacles.org/literal-and-liberal-translations-of-bertrands-box-paradox/ (Hämtad 2023-10-10).

[5] Dan (2021). The Two-Child Problem (when one is a girl named Florida born on a Tuesday). *Untrammeled mind.* https://www.untrammeledmind.com/2017/12/two-child-problem-when-one-is-a-girl-named-florida-born-on-a-tuesday/ (Hämtad 2023-10-24).

[6] D'Agostini, G. (2010). On the so called Boy or Girl Paradox. Universitá "La Sapienza" and INFN, Roma, Italia. https://arxiv.org/pdf/1001.0708.pdf

[7] Falk, R. (1992). A closer look at the probabilities of the notorious three prisoners. Cognition, 43(3), 197–223. https://doi.org/10.1016/0010-0277(92)90012-7

[8] Gardner, M. (1959). Mathematical games. *Scientific American*, 200(5), 164–174. http://www.jstor.org/stable/24940308

[9] Gardner, M. (1959). Mathematical games. *Scientific American*, 200(6), 160–172. http://www.jstor.org/stable/26309515

[10] Gardner, M. (1959). Mathematical games. *Scientific American*, 201(4), 174–184. http://www.jstor.org/stable/24940425

[11] Gardner, M. (2001). *Collosal book of mathematics.* (s.277). W. W. Norton & Company.

[12] KTH. *Bayesianska metoder.* (Författare och årtal saknas.) https://www.math.kth.se/matstat/gru/godis/bayes.pdf (Hämtad 2023-10-25).

[13] Marian, J. (2016). The 'day of the week boy or girl' paradox explained. *Jakub Marian's Language learning, science & art.* https://jakubmarian.com/the-day-of-the-week-boy-or-girl-paradox-explained/ (Hämtad 2023-10-15).

[14] Marks, S. & Smith, G. (2011). The Two-Child Paradox Reborn? *Chance*, 24:1, 54-59, DOI: 10.1080/09332480.2011.10739852

[15] Mlodinow, L. (2008). *The Drunkard's Walk: How Randomness Rules Our Lives* (s.106) https://pdfroom.com/books/the-drunkards-walk-how-randomness-rules-our-lives/bWx5aXeD2BJ (Hämtad 2023-10-25).

[16] Rehak, L. (2018). TuesdayBirthday Problem. *Towards Data Science* https://towardsdatascience.com/tuesday-birthday-problem-2927e83e5af3 (Hämtad 2023-1025).

[17] The Cthaeh (2016). What Are Bayesian Belief Networks? (Part 2). *Probabilistic world.* https://www.probabilisticworld.com/bayesian-belief-networks-part-2/ (Hämtad 2023-10-25).

[18] Torrence, B. (2010). Gathering for Gardner. *Math Horizons*, 18:1, 9-12, DOI: 10.4169/194762110X525593

[19] Wikipedia contributors (2022). Conditional dependence. *Wikipedia, The Free Encyclopedia.* https://en.wikipedia.org/w/index.php?title=Conditional_dependence&oldid=1106881817 (Hämtad 2023-10-25).

[20] Zazkis, R. & Wijeratne, C. (2015). Two Boys problem revisited. Simon Fraser University, Burnaby, British Columbia, Canada. https://www.researchgate.net/publication/273773334_Two_Boys_problem_revisited_or_What_has_Tuesday_got_to_do_with_it