Foto: Colourbox

FODBOLD-LANDSHOLDETS FØDSELSDAGSBULE

Hvis du godt kunne tænke dig at få en søn, der engang kommer på det danske fodboldlandshold, så øger du chancen ved at blive gravid i maj måned. Claus Thorn Ekstrøm tager et statistisk kig på det danske landsholds fødselsdagsfordeling gennem tiden.

»Hvis bare man har evnerne, skal man nok komme på landsholdet«. Det skulle man i det mindste tro var rigtigt, men flere undersøgelser har indikeret, at det måske alligevel ikke gælder. I mange sportsgrene har det vist sig, at der blandt de professionelle spillere er en overrepræsentation af sportsudøvere, der er født i årets første måneder, og færre, der er født sidst på året. Dette fænomen – der på engelsk kaldes "birthday bulge" ("fødselsdagsbule") – er set i mange lande for blandt andet fodbold, amerikansk fodbold og ishockey, og her vil jeg undersøge, om det også gør sig gældende i Danmark.

Argumentet for, at personer, der er født i årets første måneder, har lettere ved at komme på landsholdet, er følgende: Sports- og fritidsaktiviteter er ofte organiseret, så børn og unge inddeles på hold på baggrund af deres fødselsår. Et hold vil derfor bestå af børn med en aldersforskel op til tæt på 1 år, og specielt i børnenes yngste år vil en 10-11 måneders forskel udgøre en markant forskel i vækst og fysiologi. Børn, der er født tidligt på året, vil overvejende være større, stærkere og hurtigere, og disse børn er derfor mere tilbøjelige til at blive udvalgt til at spille kampe, og de har lettere ved at klare sig bedre end deres yngre holdkammerater. Et barn født i årets første måneder vil derfor opleve flere succeser og er derfor mere tilbøjelig til at fortsætte med at dyrke deres sport, hvilket dermed giver dem større sandsynlighed for efterfølgende at blive udtaget til landsholdet.

Test af hypotese

DBU's landsholdsdatabase indeholder oplysninger om samtlige danske landsholdsspillere siden ca. 1908. Pr. 7. maj 2017 var der 799 personer, der har været udtaget til det danske herre A-landshold i fodbold,

Forfatteren

Claus Thorn Ekstrøm er professor ved sektion for biostatistik, Københavns Universitet.
ekstrom@sund.ku.dk

og fordelingen af samtlige spilleres fødselsmåned fremgår af figur 1.

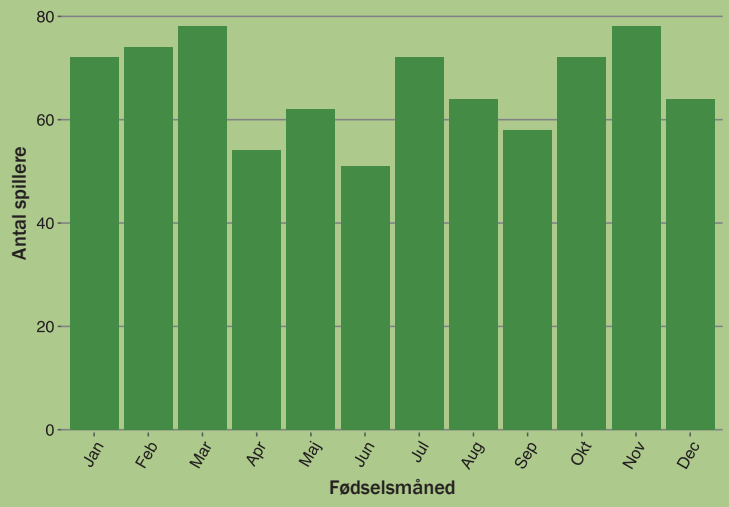
Ved første øjekast antyder figur 1, at der måske *ikke* er noget om snakken. Der lader til at være en ret jævn fordeling af fødselsmåneder henover året, når vi tager samtlige spillere på landsholdet over alle årene i betragtning. Vi kan undersøge hypotesen mere formelt ved statistisk at teste, om fordelingen af fødselsmåneder blandt landsholdsspillerne svarer til fordelingen i baggrundspopulationen. Dette kan gøres ved at lave et såkaldt χ^2 goodness-of-fit test, hvor sandsynligheden for at blive født i en bestemt måned er givet ud fra en på forhånd defineret nulhypotese.

Nulhypotesen kan vælges på lidt forskellig vis alt afhængig af, hvor præcis man ønsker den. Et simpelt bud er at sige, at fødselshyppighederne er ligefordelt henover de 12 måneder svarende til en sandsynlighed på 1/12 for hver måned. Alternativt kan fødselshyppighederne afhænge af, hvor mange dage, der er i hver måned. Endelig kan vi bruge oplysninger om danskernes fødselsmønstre fra Danmarks Statistiks statistikbank for at sammenligne fodboldspillernes fødselsfordeling med den fordeling, som vi ser for resten af danskerne.

Tabellen viser forskellen i sandsynligheder for de tre forskellige måder at definere nulhypotesen på. Den sidste søjle i tabellen viser, at der i den danske population i al almindelighed er stor sandsynlighed for at være født i perioden fra marts til maj. Ved at sammenligne med den empiriske fordeling i den sidste søjle sikrer vi, at vi ikke fejlagtigt konkluderer, at der er en overrepræsentation hos landsholdsspillerne, der skyldes, at danskerne bare i al almindelighed er mere tilbøjelige til at få børn i årets første måneder.

Ligner landsholdsspillerne resten af befolkningen?

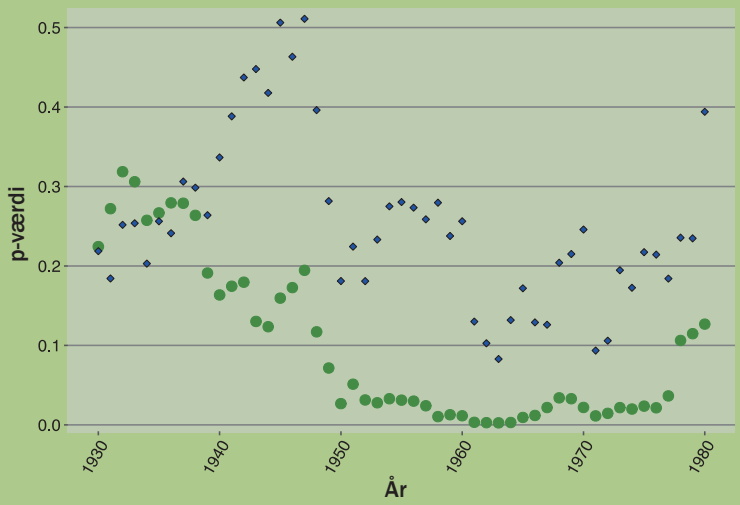
Som nævnt kan vi bruge et goodness-of-fit χ^2-test til at sammenligne fordelingen af landsholdsspillernes



Figur 1. Fødselsmåned for samtlige 799 spillere på A-landsholdet siden landsholdsdatabasen startede. Ældste person i databasen er Johannes Gandil fra 1873 og yngste er Kasper Dolberg fra 1997.

	Ligefordeling	Månedslængde	Befolkningen
Januar	8,33	8,49	7,96
Februar	8,33	7,67	7,74
Marts	8,33	8,49	8,91
April	8,33	8,22	8,86
Maj	8,33	8,49	8,89
Juni	8,33	8,22	8,62
Juli	8,33	8,49	8,82
August	8,33	8,49	8,62
September	8,33	8,22	8,43
Oktober	8,33	8,49	8,04
November	8,33	8,22	7,55
December	8,33	8,49	7,58

Nulhypotesen kan enten baseres på, at hver måned er lige hyppig (sandsynligheden for at være født i en måned er 1/12 = 8,33%), at fødselsmønstrene afspejler længden af månederne, eller at fødselsfordelingen ligner den danske baggrundsbefolkning (her vist på baggrund af alle danske fødsler fra 1970 og frem til og med 1997).



Figur 2. P-værdier for goodness-of-fit, hvis man udelukkende betragter spillere, der er født i et givent år eller senere. De blå værdier er fra det sædvanlige χ^2-test, mens de grønne værdier er fra det stærkere Kolmogorov-Smirnov-test, der tager ordningen med i betragtning. Små værdier indikerer, at landsholdsspillerne har et andet fødselsmønster end baggrundsbefolkningen.

Sammenligning af de to test

χ^2-testet sammenligner den observerede fordeling af spillernes fødselsmåneder med den fordeling af spillere, man ville forvente, hvis nulhypotesen var korrekt. Har vi eksempelvis 100 spillere som fordeler sig som vist i tabellen nedenfor, så kan vi let udregne det forventede antal spillere ud fra befolkningens frekvenser og sammenligne de to sæt tal.

	Jan	Feb	Mar	Apr	Maj	Jun	Jul	Aug	Sep	Okt	Nov	Dec
Observeret	14	12	11	10	8	8	8	6	6	6	6	5
Forventet	7,96	7,74	8,91	8,86	8,89	8,62	8,82	8,62	8,43	8,04	7,55	7,58

Teststørrelsen for χ^2-testet er defineret som vist nedenfor, hvor vi sammenligner og summerer over alle 12 måneder:

$$X^2 = \sum_{i=1}^{12} \frac{(\text{Obs}_i - \text{Forvent}_i)^2}{\text{Forvent}_i}$$

Hvis teststørrelsen er stor, stemmer hypotesen (de forventede tal under nulhypotesen om, at spillernes fordeling burde ligne baggrundsbefolkningen) dårligt overens med de reelle observerede tal, men hvis X^2 er tæt på 0 må det betyde, at de observerede tal og de forventede tal stemmer godt overens.

For vores situation har det sædvanlige χ^2-test den ulempe, at det ikke bruger informationen omkring månedernes rækkefølge. Som regel er det en god egenskab ved χ^2-testet, at det lægger lige stor vægt på alle kategorierne, men her betyder det, at bytter vi rundt på rækkefølgen af månederne, så får vi helt den samme teststørrelse og konklusion.

Den oprindelige påstand var, at der blandt fodboldspillerne ville være en ophobning af fødsler i starten af året. Man kan derfor med fordel vælge en anden teststørrelse, der har lettere ved at finde disse specifikke afvigelser (og som til gengæld så har svære-re ved at identificere andre typer afvigelser). En sådan teststørrelse er Kolmogorov-Smirnov-teststørrelsen, der betragter den absolutte kumulative afvigelse over månederne mellem de observerede data og de forventede data:

$$KS = \max_{m\in\{1,\ldots,12\}} \left| \sum_{i=1}^{m} (\text{Obs}_i - \text{Forvent}_i) \right|$$

Fordi vi summerer over månederne i rækkefølge vil en systematisk afvigelse fra befolkningens fordeling have lettere ved at træde igennem, hvilket præcis svarer til den oprindelige påstand om, at fødslerne var ophobet i årets første måneder.

Kolmogorov-Smirnov-teststørrelsen følger ikke nogen simpel fordeling, så for at vurdere vores fund må vi simulere data under nulhypotesen om, at spillernes fordeling ligner baggrundsbefolkningen, og så sammenligne teststørrelsen fra landsholdets data med teststørrelserne fra de simulerede data.

Fra et statistisk synspunkt er det interessant, hvor meget styrke, der går tabt, når man bruger det sædvanlige χ^2-test, hvor månedernes rækkefølge ikke bruges. Afvigelserne fra baggrundsbefolkningens fordeling er så lille, at effekten ikke kan identificeres med det sædvanlige χ^2-test, hvilket kan ses på de store forskelle i figur 2.

Links
www.sandsynligvis.dk
www.badmintonpeople.dk
www.dbu.dk/landshold/landsholdsdatabasen/
www.statistikbanken.dk

fødselsmåneder med danskernes generelle fødselsmønster, men χ^2-testet bruger ikke informationen omkring rækkefølgen på månederne. I stedet kan vi bruge et Kolmogorov-Smirnov test baseret på de kumulerede gruppefrekvenser, hvor kategoriernes ordning tages i betragtning (se boksen).

Figur 2 viser p-værdier for goodness-of-fit for både Kolmogorov-Smirnov-testet, og det sædvanlige χ^2-test. For hvert kalenderår vises resultatet baseret på alle landsholdsspillere født i det pågældende år eller senere. De Kolmogorov-Smirnov-baserede p-værdier viser tydeligt, at for spillere født efter anden verdenskrig forkaster man nulhypotesen om, at fødselsfordelingen blandt landsholdsspillerne følger fordelingen i den danske population.

Tilsyneladende er det danske fodboldlandshold også præget af birthday-bulge effekten – der er i hvert fald en tilsyneladende sammenhæng. Figur 2 viser også, at p-værdierne stiger hen mod slutningen af perioden. Det skyldes primært, at antallet af unge personer på landsholdet er lille, og derfor bliver styrken for goodness-of-fit-testet meget lav.

Ønsker man sig en landsholdsspiller i familien skal man altså overveje at blive gravid i starten af maj, så man kan få sig et barn i starten af det nye år. Så har man givet sit barn et forspring i livet.

Badminton er en anden sag!

Til forskel fra fodbold er klubbadminton organiseret i 2-års grupper, og badminton indebærer ikke samme fysiske nærkampe som fodbold. Som et kuriosum kan det nævnes, at hvis man laver samme vurdering med de 500 bedste danske badmintonspillere (listen kan udtrækkes fra www.badmintonpeople.dk) så finder man ikke en birthday-bulge-effekt – fordelingen af de bedste danske badmintonspillere ser ikke ud til at adskille sig fra resten af befolkningen. Faktisk er nummer et på Danmarks badmintonrangliste Jan Ø. Jørgensen, der er født 31. december! ■