Vietnamlotteriet[1]

I 1969 havde utilfredsheden med det amerikanske engagement i Vietnamkrigen også spredt sig til den måde nogle amerikanere tilsyneladende kunne unddrage sig militærtjenesten efter forgodtbefindende. Det blev derfor besluttet at gøre udsendelsen mere retfærdig ved at indkalde de følgende årgange via en national repræsentativ lodtrækning, der altså ikke favoriserede bestemte befolkningsgrupper osv. Udtrækningen skulle afhænge af fødselsdatoen. De 366 mulige fødselsdage blev lagt ind i kapsler og placeret i en stor gennemsigtig tønde, hvorefter kapslerne blev trukket én for én. Den først udtrukne fødselsdato fik udtrækningsnummeret 1, den næste udtrækningsnummeret 2 osv. Jo lavere udtrækningsnummer man trak jo større chance for at blive sendt af sted, idet de udsendte kompagnier blev fyldt op nedefra i forhold til udtrækningsnummeret. Blandt de mulige kandidater ved lodtrækningen var præsidenterne Bill Clinton og George W. Bush, der begge træk frinummer (311 og 327). Alt skulle således sikre retfærdighed i udsendelsen til den ikke altid lige populære krigstjeneste.

The New York Times, 4 Jan. 1970 David E. Rosenbaum[2]

The new draft lottery is being challenged by statisticians and politicians who claim that the drawings for the lottery are not random. A federal district judge in Wisconsin has agreed to hear a test case on the lottery. It comes at a time when hundreds of thousands of men have been assigned a place in the draft sequence and the first men are about to be inducted using the new lottery.

President Richard Nixon signed Nov.26 an executive order to "establish a random selection sequence" for induction. The order stipulated that the lottery would be based on birthdays but did not say how the dates should be chosen.

After a staff meeting it was decided that the 366 dates of a year should be placed in capsules and then be drawn one by one from a large bowl. A man's draft number would then correspond to the order in which his birthday was drawn. For example Sept.14 was the first date drawn and June 8 was the last number drawn. Thus a man with birthday Sept.4 would have draft number 1 and someone born June 8 would have draft number 366. Pentagon manpower specialists believe that those in the last third of the numbers (200 to 366) would escape the draft entirely.

A knowledgeable White House official said this week that "discussions that the lottery was not random are purely speculative." He added that there was no possibility of another drawing.

Senator Edward Kennedy asked the National Sciences last month to analyze the "apparent lack of randomness" in the selection. The Academy has not yet decided whether to do this or not.

The challenge to the randomness is being brought by Mr.Stodosky, a 24 year-old doctorate student in computer planning. The challenge is based on the average numbers for the men in the lottery for each month. If the system were random, each month could be expected to average around 183 or 184. Each of the first six months average above this and each of the last six months average below it. Statisticians, who have studied the lottery, say that this could occur if the capsules with later months were not mixed as thoroughly as those with early months.

[1] Oplysninger om vietnamlotteriets historie, herunder ovenstående billede, kan findes på den amerikanske regerings hjemmeside for **Selective Service System**: www.sss.gov

[2] Citeret efter Chance News 6.10, se hjemmesiden http://www.dartmouth.edu/~chance

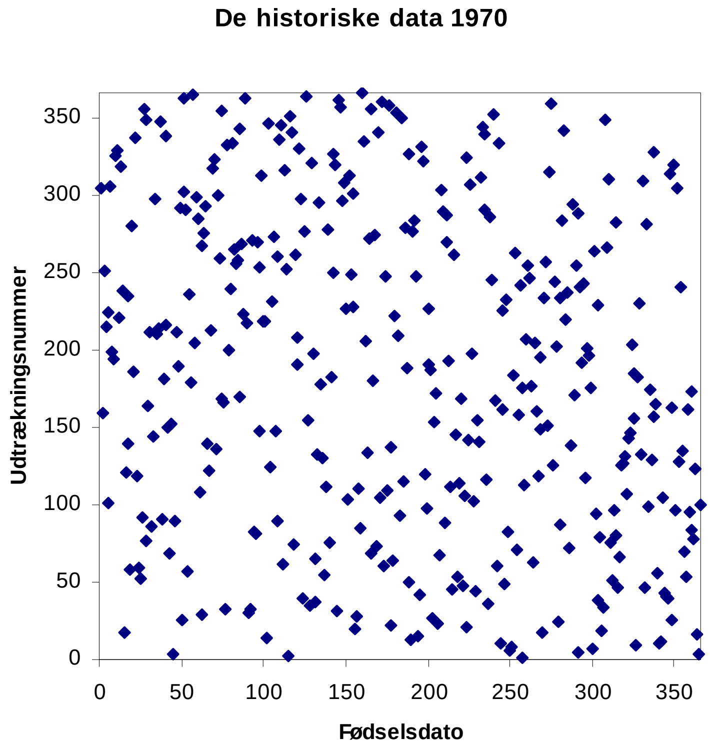
Som det fremgår af det ovenstående avisartikel fremkom der hurtigt protester over at udtrækningen havde favoriseret kandidater med en lav fødselsdato. Hvordan kan man nu undersøge dette problem? Hvis man kigger på listen over de faktisk udtrukne numre[3] bliver det ikke meget klarere:

258 115 365 45 292 250 300 251 327 341 244 342 190 102
194 364 15 270 306 156 223 178 206 279 50 349 203 157 62
91 145 92 77 307 128 237 132 304 346 124 345 195 344 229
215 316 332 221 247 189 312 25 357 218 137 340 54 19 24
173 242 112 264 179 131 317 207 43 165 356 254 286 169
118 140 311 28 362 305 314 95 249 94 360 159 32 280 210
46 109 38 26 183 302 359 351 313 199 334 366 5 228 151
171 343 222 321 61 175 158 214 138 259 219 185 236 296
23 267 198 16 67 363 104 276 318 319 353 336 136 320 330
133 163 355 71 177 287 66 18 231 225 322 33 217 323 98
107 269 42 273 44 204 230 127 326 338 255 2 266 246 358
348 30 339 76 241 220 75 86 289 205 361 335 257 299 263
135 56 167 39 328 141 252 325 21 202 187 48 200 120 294
213 9 268 298 130 227 8 79 297 278 324 265 58 162 260
121 182 35 31 47 68 36 4 41 90 101 100 284 12 180 88 6
245 150 201 154 303 329 105 248 271 281 17 55 285 14 80
354 293 256 295 277 239 262 174 193 153 142 3 114 97 290
261 83 272 84 73 108 216 119 253 301 82 309 63 87 96 211
93 164 106 168 64 125 191 139 186 20 333 315 282 192 60
238 212 291 209 53 234 49 65 288 134 148 34 123 59 72
155 51 208 352 1 7 226 149 331 310 232 99 152 347 274
113 69 13 144 350 129 197 70 224 10 143 188 337 11 122
196 78 243 81 161 110 22 40 235 117 170 283 85 233 111
103 37 308 29 184 116 240 181 74 27 166 147 176 275 172
146 89 52 126 57 160

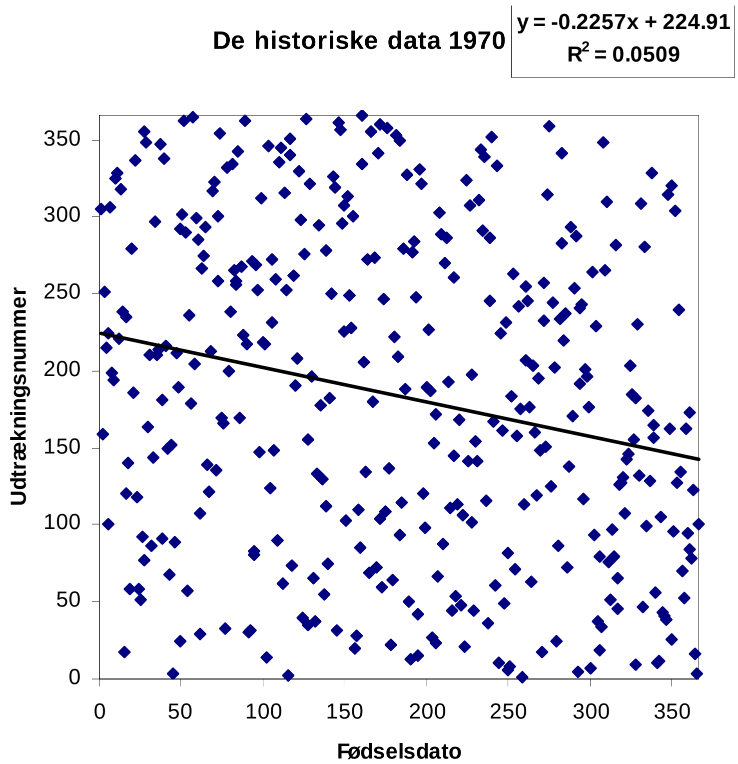
Her står 258 for den først udtrukne fødselsdato (den 14. september) og alle der var født den 14. september var altså sikker på at blive udtrukket. Tilsvarende gælder for alle de følgende udtrukne numre. Dem der blev trukket tidligt var sikre på at komme af sted, dem der blev udtrukket sent var sikre på ikke at blive udsendt. Grænsen regnede man med ville ligge omkring udtrækningsnummeret 200, dvs. de første 200 udtrukne datoer var sikre på at blive sendt af sted, mens resten kunne regne med at trække frinummer. Det er de første 200 udtrukne fødselsdatoer vi har markeret med gråt i skemaet. men det gør det ikke nemmere at afsløre om der er et mønster bag tallene, dvs. en eller anden form for tendens i den måde de blev udtrukket på.

Som sædvanlig vil vi derfor i stedet kigge på grafiske fremstillinger. Fx kan vi afbilde udtrækningsnummeret som funktion af fødselsdatoen. Derved fås den følgende graf (se næste side). Men man skal kigge ret grundigt på en sådan graf for at få en fornemmelse af at der ikke skulle være tale om en tilfældig fordeling, men i stedet en udtynding i øverste højre hjørne svarende til at unge mænd med en høj fødselsdato har en tendens til at trække et lavt nummer – og dermed ryge til Vietnam.

[3] Listen, der findes mange steder på nettet, fx hjemmesiden for Selective Systems Service, kan overføres direkte til et regneark, men er også stillet til rådighed som et regneark på EMU.



Vi prøver derfor at indlægge en tendenslinje (mindste kvadraters linje):



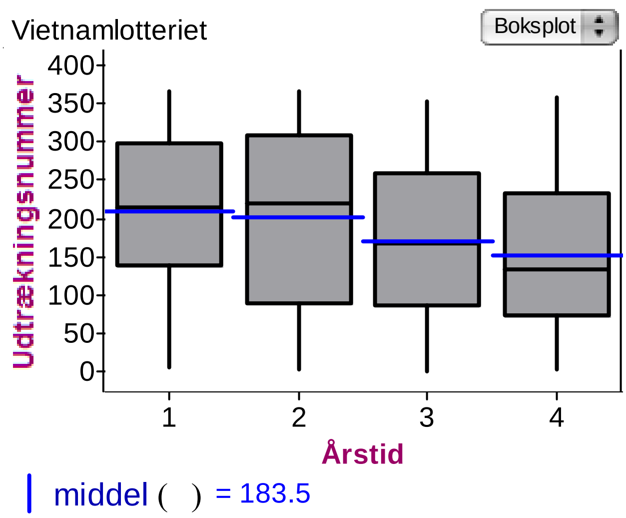
Vi ser da at tendenslinjen klart hælder nedad, men forklaringsgraden er kun ca. 5%. Vi ville selvfølgelig have forventet en stort set vandret linje med en forklaringsgrad tæt på 0%, idet udtrækningsnummeret slet ikke burde afhænge af fødselsdatoen. Her hælder tendenslinjen i stedet nedad, hvilket netop betyder at mænd med en høj fødselsdato ser ud til at have en større tendens til at trække et lavt nummer.

Og hvad så? Giver det anledning til bekymring? I en tilfældig lodtrækning vil linjen jo aldrig være helt vandret, men sommetider få en svag stigning, sommetider et svagt fald, ligesom forklaringsgaden ikke helt kan blive nul fordi der typisk vil være en lille men tilfældig ubalance. Så kunne dette ikke bare være tilfældets spil?

Kritikerne hæftede sig ved udtrækningsmåden. Som det tydeligt ses på fotografiet af udtrækningen er det svært at nå bunden af tønden. Hvis kapslerne nu ikke var blevet blandet godt nok op i tønden, så de højeste fødselsdatoer havde en tendens til at ligge øverst, ville de også have en tendens til at blive trukket først. Men hvordan var kapslerne blevet lagt ned i tønden? Her er en samtidig beskrivelse[4]:

> The reporter, Fred Garretson, however, was able to find out exactly how the lottery was held and how the bias likely occurred. He found that the people in charge of the draft lottery placed all of the days of the year in a shoebox, adding the days to the box month by month. As the box became full, the numbers at one end would be pushed back towards the other end and the new month's dates added. When the hand was removed, the old dates would slide back over the new. This was repeated until all of the days in a year were in the box. Then the box would be shaken and turned over, the numbers falling into a second container. As the box was quite full, shaking might not have mixed the things up very much. The result was that the dates from the end of the year could have been enriched in the top of the second container. At some later time someone would pick a number, probably off the top.

Vi må altså kigge grundigere på tallene for at se om vi kan dokumentere det påståede problem. Det kan da typisk betale sig at forenkle datastrukturen ved at samle nøgletal om de indsamlede data. I stedet for at kigge på udtrækningsnummeret som funktion af fødselsdagen kunne vi da forenkle datastrukturen ved at erstatte fødselsdagen med årstiden, dvs. vinteren svarende til de tre første måneder (91 dage), foråret svarende til de næste tre måneder (91 dage), sommeren svarende til de næste tre måneder (92 dage) og efteråret svarende til de sidste tre måneder (92 dage). For hver af de fire årstider kan vi så udregne kvartilsættet og middelværdien og afbilde de tilhørende boksplot[5]:



[4] Citeret efter hjemmesiden http://csee.lbl.gov/asteroid_impact/franksstory.html

[5] I de følgende illustrationer skifter vi regneark til databehandlingsprogrammet DataMeter.

Nu træder strukturen langt tydeligere frem: Middelværdien falder systematisk med årstiden og tilsvarende falder kvartilboksen tydeligt i slutningen af året. Dette bekræftes af en oversigtstabel:

Vietnamlotteriet

	Årstid				Række total
	1	2	3	4	
	210.13187	202.51648	170.91304	150.93478	183.5

R1 = middel (Udtrækningsnummer)

Vi kan også tydeliggøre strukturen ved at kigge på kategoriserede data. Hvis en ung mand møder op til sessionen for at få oplyst sit udtrækningsnummer er det eneste afgørende sådan set hvordan nummeret ligger i forhold til den forventede grænse på 200. Om det er 17 eller 135 er fuldstædigt ligegyldigt for i begge tilfælde bliver han sendt af sted. Og tilsvarende for høje numre. Vi kan derfor kategorisere udtrækningsnumrene alt efter som de ligger under (eller lig med) 200 eller over 200. I det første tilfælde er svaret på spørgsmålet om han bliver sendt til Vietnam 'JA', idet andet tilfælde 'NEJ':

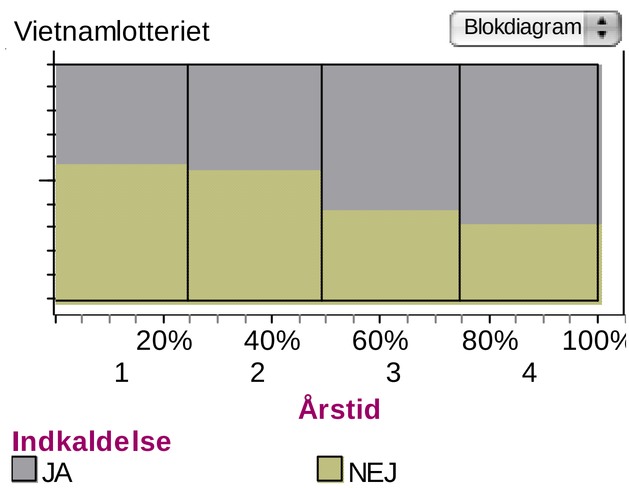
Derved finder man følgende tal:

Vietnamlotteriet

		Årstid				Række total
		1	2	3	4	
Indkaldelse	JA	39	41	57	63	200
	NEJ	52	50	35	29	166
	Søjle total	91	91	92	92	366

R1 = tæl (   )

Antallet af indkaldte stiger altså støt og roligt med årstiden og der er ca. 50% større chance for at blive indkaldt hvis man er født i årets sidste måneder, end hvis man er født i årets første måneder. Det kan også fremstilles med en graf, her hyppighederne for indkaldelserne som funktion af årstiden:



Vi har altså efterhånden samlet en hel del evidens for at der rent faktisk er et problem! Men hvordan dokumenterer vi problemet statistisk? Hvordan sikrer vi os mod at tallene ikke lige så godt kunne være fremkommet ved rene tilfældigheder?

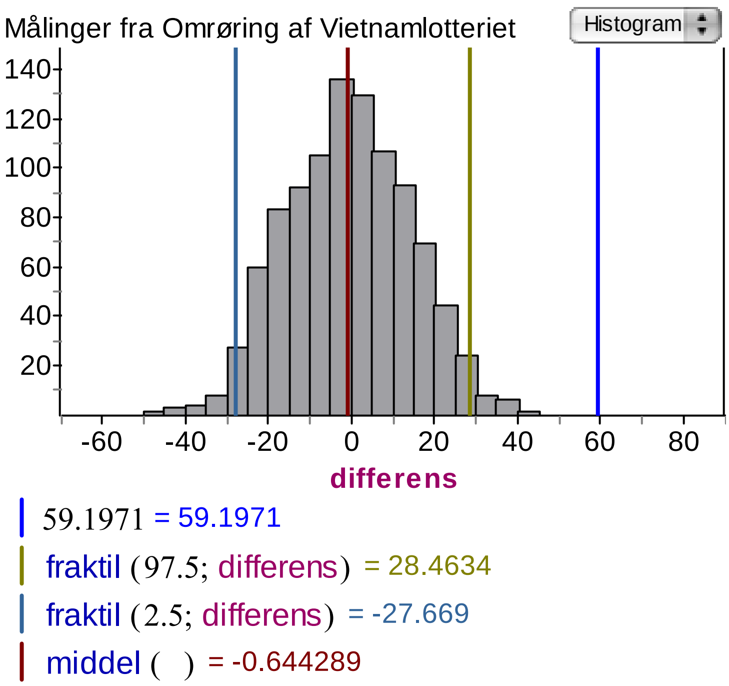
Det kan vi nemmest gøre ved hjælp af en simulering af Vietnamlotteriet hvor vi lader tilfældigheden råde og derved kan undersøge hvilke mønstre man med rimelighed kan tilskrive tilfældigheder og hvilke man må tilskrive systematiske variationer i data.

Det er rimeligt nemt at simulere et vietnamlotteri, hvis blot vi har adgang til en metode til at udtrække 366 tal i en tilfældig rækkefølge. Det skal altså vel at mærke ske som en stikprøve *uden* tilbagelægning. Hvis du ikke har adgang til en sådan uden videre, kan man selv skrive en lille stump programkode, der kan stå for udtrækningen (se indskuddet). Her vil vi benytte Datameter til at simulere problemet. Der er forskellige metoder til at udføre disse simuleringer, men det nemmeste er at benytte 'omrøring af en variabel' til at udføre tilfældige permutationer af udtrækningsnumrene. Derefter kan vi nemt se hvilken indflydelse det har på denne måde at lade tilfældigheden råde. Vi skal da blot blive enige om hvilket mønster det er vi vil undersøge. Vi giver nogle eksempler:

I første omgang hæftede vi os ved at middelværdien aftog gennem de fire årstider. Vi kan da spørge om hvor stor sandsynligheden for dette er hvis vi lader tilfældigheden råde. Her er selv rækkefølgen ikke nok, for der er kun 24 måder at arrangere fire tal i rækkefølge på, så rent tilfældigt vil en faldende rækkefølge opstå en ud af 24 gange og det er jo ikke så usædvanligt igen. Men forskellen mellem middelværdierne i den første og sidste årstid forekom os at være signifikant, så den kunne vi forsøge at måle på:

differens	59.1971	middel (Udtrækningsnummer; Årstid = 1) − middel (Udtrækningsnummer; Årstid = 4)

I det aktuelle tilfælde er den 59.1971. Hvor sandsynligt er det nu at få så stor en afvigelse? Vi omrører derfor udtræksnummeret 1000 gange og kigger på fordelingen af forskellen:

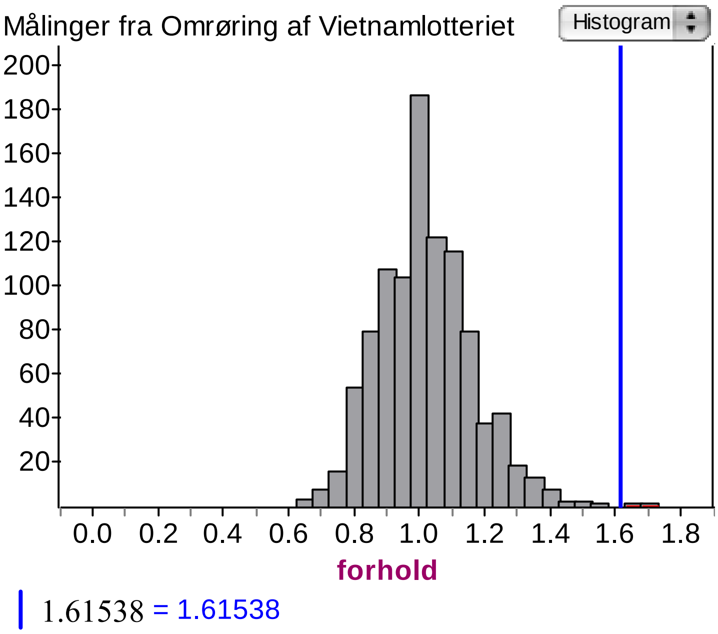


Selv med tusind forsøg lykkedes det os altså ikke en eneste gang at frembringe en tilfældig differens, der var lige så stor. Det er altså en meget sjælden begivenhed og alt tyder derfor på en systematisk variation. De midterste 95%, dvs. de forventede udfald, går fra ca. -30 til ca. 30. Så sessionskomiteen har et forklaringsproblem!

I anden omgang hæftede vi os ved den store forskel i hyppigheder. Her kunne det være forholdet mellem antallet af indkaldte fra den fjerde årstid og antallet af indkaldte fra den første årstid vi fandt signifikant. Denne gang måler vi derfor størrelsen:

forhold	1.61538	$\dfrac{\text{tæl}(\text{Udtrækningsnummer} \leq 200;\ \text{Årstid} = 4)}{\text{tæl}(\text{Udtrækningsnummer} \leq 200;\ \text{Årstid} = 1)}$

I det aktuelle tilfælde er forholdet 1.61538.



Konklusionen er stort set den samme: I tusind tilfældige udtrækninger lykkedes det i dette forsøg kun to gange at opnå et forhold mellem de udvalgte fra sidst på året og de udvalgte fra først på året, der var mindst lige så stort som det observerede forhold 1.61538. Sessionen har altså stadigvæk et alvorligt forklaringsproblem med at fastholde at udtrækningen var tilfældig.

Pointen med at gentage simuleringen på to forskellige måder er selvfølgelig ikke at man altid skal gøre det på en hel masse forskellige måder, men mere at der er en betydelig valgfrihed i hvilken måde man vil gøre det på. Netop fordi vi bestemmer sandsynligheden eksperimentelt kan vi frit vælge vores teststørrelse. På den anden side kan vi selvfølgelig også udføre en kanonisk test. En af favoritterne er chi-i-anden-testen. Vi kan fx vælge at teste om chancen for at blive indkaldt er uafhængig af årstiden. I så fald hæfter vi os altså ved fordelingen af hyppighederne for indkaldelserne:

Vietnamlotteriet

		Årstid				Række total
		1	2	3	4	
Indkaldelse	JA	39	41	57	63	200
	NEJ	52	50	35	29	166
	Søjle total	91	91	92	92	366

R1 = tæl ( )

Vi udfører derfor et uafhængighedstest for at undersøge om det er lykkedes for sessionen at fordele udtrækningsnumrene uafhængigt af hvilken årstid man er født:

Test fra Vietnamlotteriet

Test af uafhængighed

Første variabel (kategoriseret): Årstid

Anden variabel (kategoriseret): Indkaldelse

		Årstid				Række total
		1	2	3	4	
Indkaldelse	JA	39 (49.7)	41 (49.7)	57 (50.3)	63 (50.3)	200
	NEJ	52 (41.3)	50 (41.3)	35 (41.7)	29 (41.7)	166
	Søjle total	91	91	92	92	366

Første variabel: **Årstid**
Antal kategorier: **4**
Anden variabel: **Indkaldelse**
Antal kategorier: **2**
Alternativ hypotese: Der er en sammenhæng mellem **Årstid** og **Indkaldelse**

Teststørrelsen, chi-i-anden, er **17.57**. Der er **3** frihedsgrader (antallet af rækker minus én ganget med antallet af søjler minus én).

Hvis det var sandt at **Årstid** var uafhængig af **Indkaldelse** (nulhypotesen), og stikprøven blev gentaget mange gange, så ville sandsynligheden for at få en værdi for chi-i-anden, der var mindst lige så stor være **0.00054**.

Tallene i parentes i tabellen er de forventede antal.

Vi indsætter derfor de to kategoriserede variable i testskabelonen. Vi får da oplyst at teststørrelsen, dvs. chi-i-anden, er 17.57, samt at sandsynligheden for at få et chi-i-anden, der er mindst lige så stort er 0.054% – hvis vi altså vel at mærke lader tilfældigheden råde. Det er lige så usandsynligt som før og tyder igen på et klart signifikant resultat. Også chi-i-anden-testen viser altså tydelige tegn på en systematisk skævvridning af lodtrækningen.

Fordelen ved de kanoniske test er at der findes en omfattende teoretisk underbygning af testen, ligesom den er velbeskrevet i litteraturen og der findes mange udførlige vejledninger i brug af testen. Ulempen er at man altså netop skal kende til en hel del teori for at forstå testen: En begrundelse for valget af teststørrelsen

$$\sum \frac{(\text{observeret-forventet})^2}{\text{forventet}},$$

centrale begreber som antallet af frihedsgrader, fordelingsfunktionen for teststørrelsen, kriterierne for anvendelsen af testen osv.

Man kan dog også gå frem rent eksperimentelt med udgangspunkt i det ovenstående skema, idet man kan oprette en måling over chi-i-anden-teststørrelsen og så undersøge dens eksperimentelle fordeling. Vi rører derfor rundt i indkaldelsen (dvs. trækker de 200 første numre helt tilfældigt) og tvinger den derved til at være uafhængig af årstiden. Derved kan vi udføre en test på de tilfældige omrørte data. Den har selvfølgelig typisk en meget mindre teststørrelse og en sandsynlighed for at ramme tilsvarende skævt, der er væsentligt højere, og derfor slet ikke spor overraskende. Derefter samler vi gentagne målinger af chi-i-anden-teststørrelsen fx 1000 gange for at kunne undersøge fordelingen. Derved kan vi ikke blot få afgjort rent eksperimentelt hvor sandsynligt det er at vi finder så stor en teststørrelse som 17.57, men også i hvilket interval de forventede teststørrelser rent faktisk befinder sig i.

Test fra Omrøring 2 af Vietnamlotteriet

Test af uafhængighed

Første variabel (kategoriseret): Årstid
Anden variabel (kategoriseret): Indkaldelse

		Årstid				Række total
		1	2	3	4	
Indkaldelse	JA	55 (49.7)	46 (49.7)	48 (50.3)	51 (50.3)	200
	NEJ	36 (41.3)	45 (41.3)	44 (41.7)	41 (41.7)	166
Søjle total		91	91	92	92	366

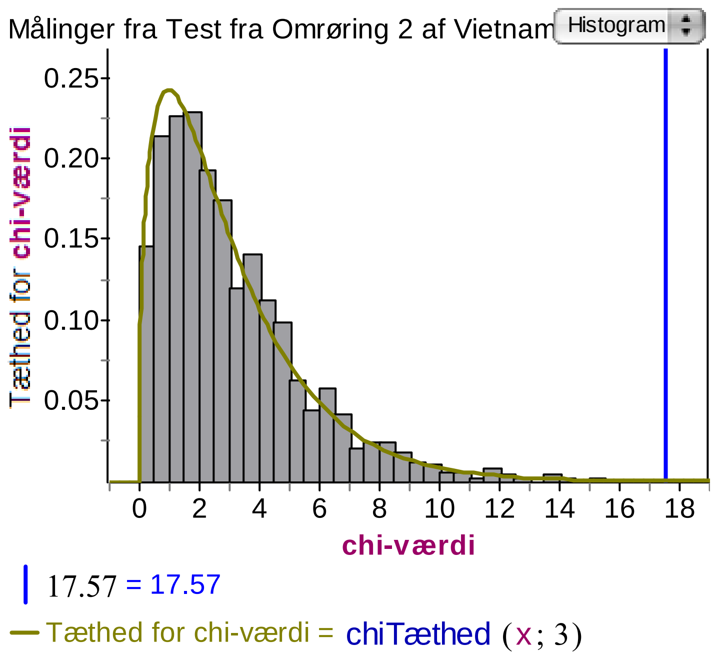
Første variabel: **Årstid**
Antal kategorier: **4**
Anden variabel: **Indkaldelse**
Antal kategorier: **2**
Alternativ hypotese: Der er en sammenhæng mellem **Årstid** og **Indkaldelse**

Teststørrelsen, chi-i-anden, er **2.099**. Der er **3** frihedsgrader (antallet af rækker minus én ganget med antallet af søjler minus én).

Hvis det var sandt at **Årstid** var uafhængig af **Indkaldelse** (nulhypotesen), og stikprøven blev gentaget mange gange, så ville sandsynligheden for at få en værdi for chi-i-anden, der var mindst lige så stor være **0.55**.

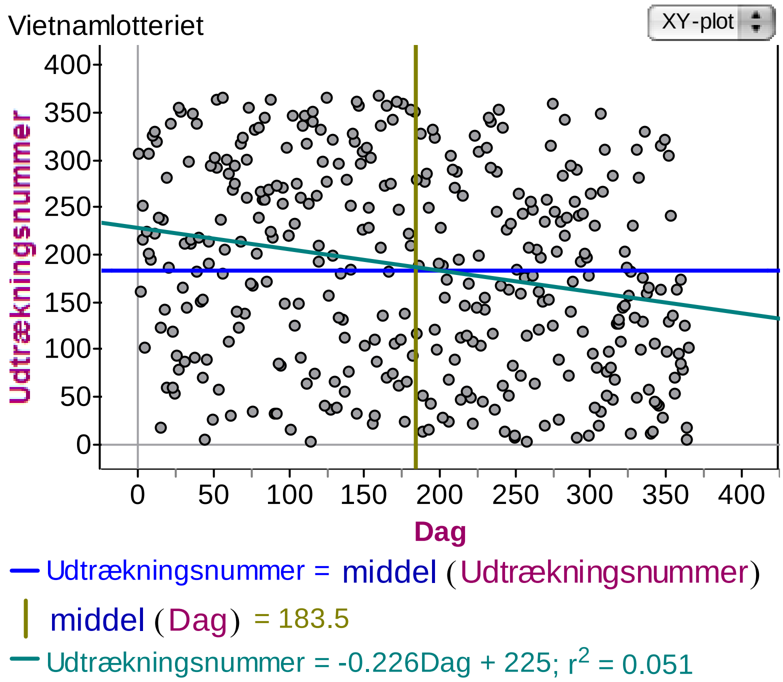
Tallene i parentes i tabellen er de forventede antal.

Vi finder da den følgende eksperimentelle fordeling:



Ud af de 1000 tilfælde var der altså ikke en eneste, der var lige så skæv som den observerede. Sandsynligheden for at få så skævt et testresultat er altså (stort set) under 0.1% i overensstemmelse med den kanoniske test.

Vi slutter med at diskutere den lineære regressionsmodel fra starten. Her kan resultatet godt virke lidt uoverskueligt:



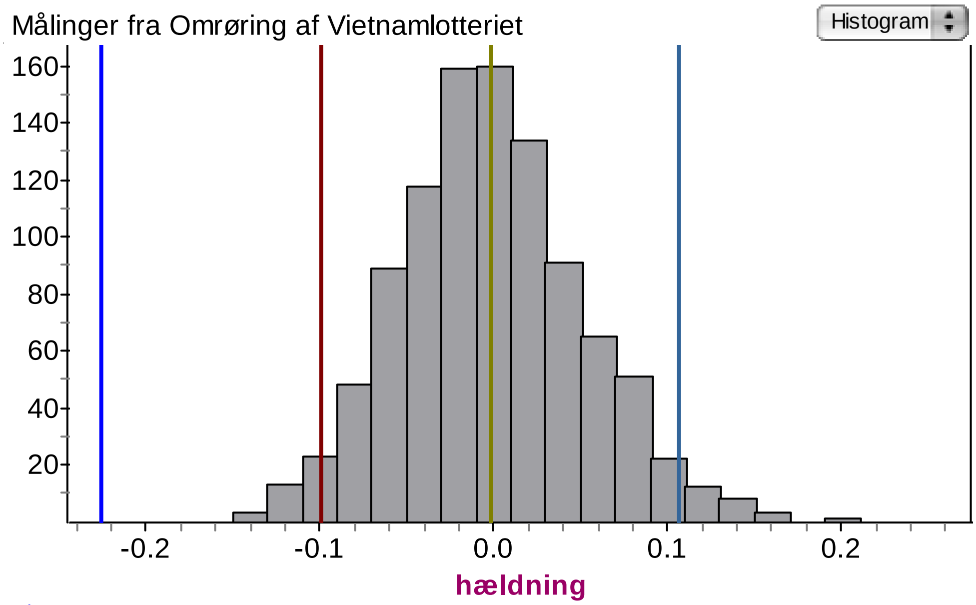
Kan man virkelig slutte noget ud fra en lineær model med en forklaringsgrad så lav som 5.1%, dvs. hvor vi kun kan forklare ca. 5% af variationen ved hjælp af den lineære model, mens de resterende 95% er tilfældige fluktuationer? Det afhænger selvfølgelig af hvad vi vil med modellen. Det er klart at en sådan model ikke har stor forudsigelseskraft! Givet en tilfældig fødselsdag vil der være en endog meget stor usikkerhed i det udtrækningsnummer, som modellen forudsiger. Modellen kan derfor helt oplagt ikke bruges til at forudsige brugbare enkeltresultater med. Men pointen er jo også en anden: Forudsætningen for at lotteriet er retfærdigt er at udtrækningsnummeret er uafhængigt af fødselsdatoen. Altså er den forventede forklaringsgrad 0% og den forventede bedste rette linje en vandret linje svarende til middelværdien, dvs. 367/2 = 183.5.

Det er svært at se visuelt at det ser *så* tosset ud, men splitter vi datasættet i fire kvadranter ud fra middelværdierne ser vi at der er en vis ubalance i tallene:

109	74
74	109

Det er denne ubalance der trækker linjen skævt nedad, så vi får en negativ hældning. Men er ubalancen signifikant, dvs. tyder den på en svag men tydelig skævhed i udtrækningsmønstret? For at afgøre det statistisk udfører vi en omrøring af udtrækningsnummeret og måler hældningen for den fremkomne bedste rette linje fx 1000 gange. Derved kan vi benytte fordelingen af de tilfældigt frembragte hældninger til at vurdere om vi med rimelighed kan forklare den observerede hældning som værende fremkommet ved rene tilfældigheder.

Vi finder da at vi overhovedet ikke kommer i nærheden af den observerede hældning i løbet af tusind forsøg. Vi ser også at en typisk hældning (svarende til de midterste 95%) ligger mellem ca. -0.10 og 0.10.



-0.225686 = -0.225686

middel ( ) = -0.00155066

fraktil (97.5; hældning) = 0.106628

fraktil (2.5; hældning) = -0.0987682

Igen findes der selvfølgelig kanoniske vurderinger af konfidensintervaller for en sådan lineær model:

Model for Vietnamlotteriet — Simpel lineær regression

Responsvariabel (numerisk): Udtrækningsnummer
Forklaringsvariabel (numerisk): Dag

Stikprøvens størrelse: **366**

Ligningen for mindste kvadraters linje: **Udtrækningsnummer** = **-0.225686 Dag + 224.91**
Korrelationskoefficienten: r = **-0.225686**
R-kvadrat = **0.050934**, hvilket vil sige at **5.0934**% af variationen i **Udtrækningsnummer** stammer fra **Dag**.

Ved et konfidensniveau på **95** % er det bedste skøn over hældningen **-0.225686** +/- **0.100413** . (Standardfejlen for hældningen er **0.051062**.)

Når **Dag** = **0** , vil den skønnede værdi for en fremtidig observation af **Udtrækningsnummer** være **224.913** +/- **204.076**.

Vi ser da at konfidensintervallet for hældningen ved et konfidensniveau på 95% er -0.226 ± 0.100 i overensstemmelse med det eksperimentelt bestemte usikkerhedsinterval. Den negative hældning kan altså ikke bare forklares som et tilfældigt udsving. Der er virkelig grund til tro at det observerede mønster er reelt. Men læg også mærke til det enorme usikkerhedsniveau, der er hvis man vil benytte modellen til at forudsige fremtidige udtrækninger. Usikkerhedsintervallet er stort set smurt ud over samtlige numre!

Benytter man den færdige model kræver det altså igen en vis teoretisk modenhed at forstå baggrunden for de forskellige oplysninger i tabellen.

Appendiks 1:
Simuleringer af Vietnamlotteriet.

Hvis man vil simulere Vietnamlotteriet skal man være opmærksom på at der er tale om en lodtrækning uden tilbagelægning. men kan derfor ikke uden videre benytte en tilfældighedsgenerator og blot fremstille en serie på 366 tilfældige tal fra 1 til 366. De enkelte udtrækninger er altså ikke uafhængige fordi der bliver færre og færre numre og fx er den allersidste udtrækning fuldstændigt bestemt af de forrige, fordi der nu kun er et tal tilbage. Vi skal derfor i stedet have frembragt en tilfældig stikprøve. Det kræver en lille programstump baseret på den følgende ide:

Der oprettes en liste med tallene fra 1 til 366.
Første gange anvender vi et tilfældigt tal mellem 1 og 366 til at finde indeks for det første tal i udtrækningen. Dette tal fjernes fra listen og resten af tallene i listen rykkes én frem for at 'lukke hullet'.
Næste gang anvender vi et tilfældigt tal mellem 1 og 365 til at finde indeks for det andet tal i udtrækningen. Dette tal fjernes fra listen osv..
Næste gang anvender vi et tilfældigt tal mellem 1 og 364 til at finde indeks for det tredje tal i udtrækningen. Dette tal fjernes fra listen osv.
Således fortsætter vi til alle tal er udtrukne.

En programstump skrevet i pseudokode kan fx se således ud:

```
randperm(liste)
Func
   n:=dim(liste)                   Find længden af listen
   For k = 1 to n
       indeks:=randint(1,n-k+1)    Træk et tilfældigt indeks
       NyListe[k]:=liste[indeks]   Flyt tallet over i ny liste
           For j = indeks to n-k
               liste[j]:=liste[j+1] Opdater gammel liste
           EndFor j
   EndFor k
   Return NyListe
EndFunc
```

Du behøver ikke forstå koden i detaljer, men kan måske få hjælp fra din lærer til at implementere den i et passende program til din lommeregner, TI-Interactive eller lignende. Så kan du selv simulere tilfældige Vietnamlotterier.

Appendiks 2: *Afsluttende bemærkning:*

The Writing on the Wall[6]
Paul M. Sommers

This is an occasion of great dignity and some solemnity. It represents the first application of a principle believed by many of us to be thoroughly democratic, equal and fair in selecting soldiers to defend the national honor abroad and at home.

— Secretary Baker, War Department
July 12, 1917 draft lottery

Introduction

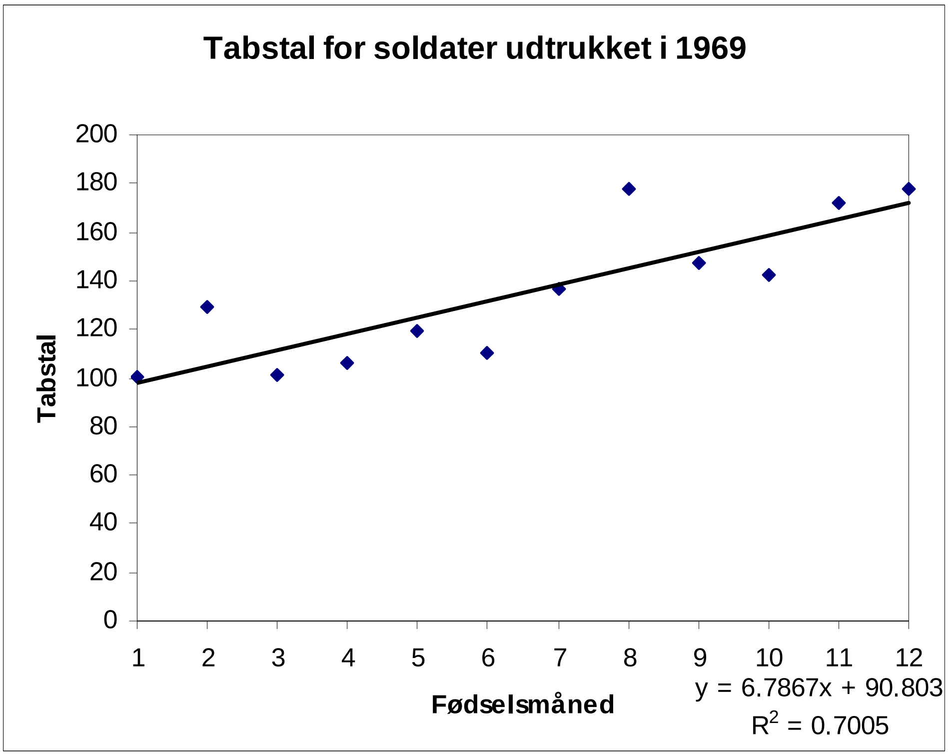
Following charges of unfairness in the procedures by which some local draft boards decided who should be drafted during the Vietnam War, in December 1969 young men were chosen for military service at random, by a "draft lottery." Soon after the drawing statisticians charged that the draft lottery was anything but random. In particular, men with birthdays in later months (lower lottery numbers) tended to be drafted before men with birthdays in earlier months (higher lottery numbers). The then prevailing belief was that the capsules (representing each of the 366 days of the year) were put into a two-foot deep bowl in monthly order and were not thoroughly mixed.

Vietnamkrigen er stadig et åbent sår for amerikanerne. Vietnamlotteriet er selvfølgelig kun en lille brik i dette spil, men brikken bliver stadig vendt på nye måder. For nylig offentliggjordes en liste over tabstallene for de soldater der var blevet indkaldt til tjeneste via den famøse 1969 udtrækning. Det er svært at finde simple statistikker over tabstal fra Vietnamkrigen som kobler dem direkte sammen med soldaternes fødselsår, men samtlige amerikanske tab er registreret elektronisk på web-siden www.the-wall-usa.com, som er en elektronisk pendant til det amerikanske mindesmærke for de faldne i Vietnamkrigen. Ud fra oplysningerne på denne mindeside lykkedes det Paul M. Sommers at stykke en tabel sammen over sammenhængen mellem fødselsår og tabstal. Da 1969 lodtrækningen var den første samlede den en række årgange fra 1944 til 1950:

Fødselsmåned	**Tabstal**
Januar	100
Februar	129
Marts	101
April	106
Maj	119
Juni	110
Juli	137
August	178
September	147
Oktober	142
November	172
December	178

Som det ses er der en svag, men tydeligt voksende tendens ned gennem tabellen. Det afspejles også i tendenslinjen for tabstallene som funktion af fødselsmånedens nummer:

[6] Paul M. Sommers er professor i økonomi ved Middlebury College. Artiklen blev trykt i tidsskriftet Chance, nr. 16.1 i 2003. Den kan downloades flere steder fra Internettet.



Det er heller ikke svært at lave statistiske tests, der viser at fordelingen af tabstal overhovedet ikke er tilfældig. Den uheldige lodtrækning trak altså sine blodige spor videre i forløbet.