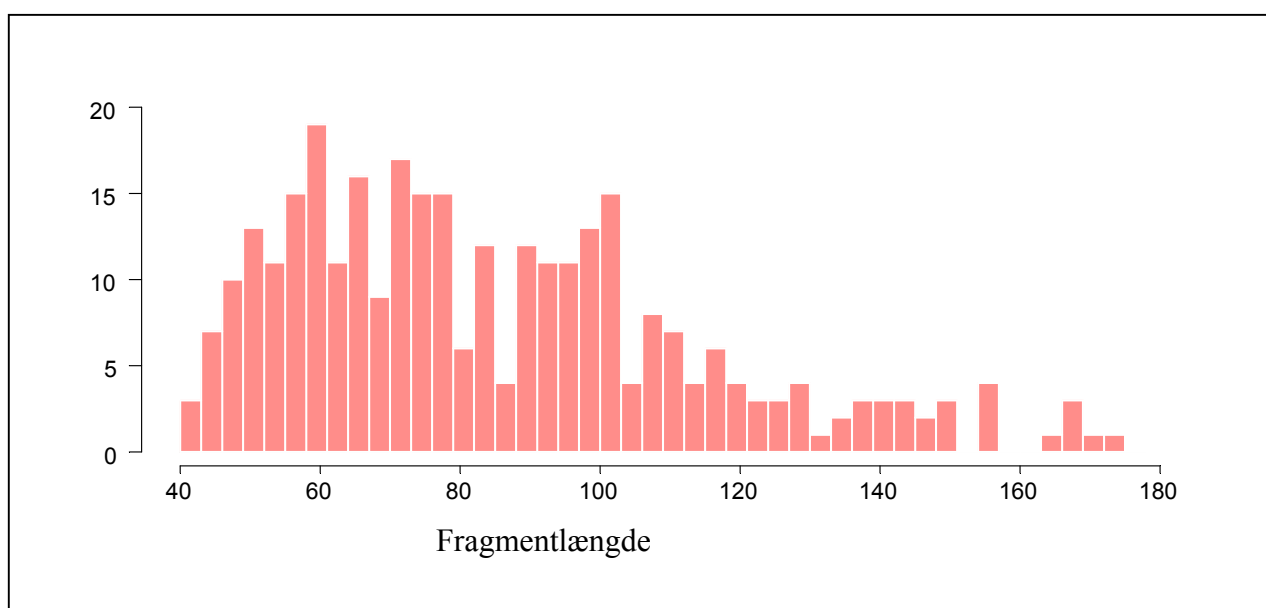


Statistik – viden eller tilfældighed

DNA-analyser

1 Sandsynligheden for at en uskyldig anklages

Følgende histogram viser, hvordan fragmentlængden for et DNA-område varierer inden for befolkningen. Der indgår 315 tilfældigt udvalgte personer og observationerne er grupperet i intervaller af længde 3 mm.



Vi går ud fra at fordelingen er repræsentativ for befolkningen.

Antag nu at et spor har en fragmentlængde på 58 mm. Hvor stor del af befolkningen vil have en fragmentlængde, der er overensstemmende med sporet, hvis der benyttes en acceptgrænse på 3 mm?

Antag at befolkningsvariationen er som ovenfor for alle 8 DNA-områder, der udgør DNA-profilen, og at de otte fragmentlængder er uafhængige af hinanden. Hvor stor en del af befolkningen kan have samme DNA-profil, som et givet spor?

Statistik – viden eller tilfældighed

2 Konstant måleusikkerhed

Punkt-plottet i teksten tyder på at måleusikkerheden ikke er afhængig af fragmentlængden. Man kan undersøge dette nærmere ved at beregne spredningen af $L_1 - L_2$, hvor L_1 og L_2 er to målinger af samme fragmentlængde, for par (L_1, L_2) grupperet efter gennemsnittet af L_1 og L_2 . I den følgende tabel er 1000 par delt op i to grupper; midtpunktet af intervallet fra 40 til 180 er brugt til at dele op efter.

Beregn spredningen af $L_1 - L_2$ i hver af de to grupper.

Interval	Hyppighed af $L_1 - L_2$ når $\frac{L_1 - L_2}{2} < 110$	Hyppighed af $L_1 - L_2$ når $\frac{L_1 - L_2}{2} > 110$
$] - 2,5; - 2,0]$	3	1
$] - 2,0; - 1,5]$	9	6
$] - 1,5; - 1,0]$	49	11
$] - 1,0; - 0,5]$	139	33
$] - 0,5; 0,0]$	191	42
$] 0,0; 0,5]$	196	50
$] 0,5; 1,0]$	131	26
$] 1,0; 1,5]$	56	23
$] 1,5; 2,0]$	21	7
$] 2,0; 2,5]$	2	3
$] 2,5; 3,0]$	1	0

Statistik – viden eller tilfældighed

Sandsynlighedsberegninger i Plejebo-sagen

3

Binomialsandsynligheder

De resultater, der er illustreret på figur 1 i teksten, fik man ved at lave simulationseksperimenter, men ved brug af binomialsandsynligheder kan man få det samme frem. Antallet af dødsfald på en afdeling i løbet af to år kaldes X . Når man går ud fra den model, der lå bag simulationseksperimenterne, gælder der, at X er binomialfordelt med antalsparameter $n = 730$ og sandsynlighedsparemeter $p = 16 / 730$.

a) Find $P(X \geq 40)$ og $P(X \geq 44)$ ved brug af for eksempel TI-83.

Hvis man ser på 2000 afdelinger, vil antallet Y af afdelinger, der har 44 dødsfald eller mere, være binomialfordelt med $n = 2000$ og $p = P(X \geq 44)$.

- b) Find sandsynligheden for at der er mindst én afdeling med mere end 43 dødsfald. Find sandsynligheden for, at der er mindst én afdeling med mere end 39 dødsfald, og sammenlign med resultatet af simulationseksperimenterne i teksten.
- c) Find sandsynligheden for, at der i løbet af 2000 år vil være mindst en afdeling ud af 2000 med flere end 43 dødsfald, når de 2000 år tænkes delt op i 1000 par på hvert 2 år.

Statistik – viden eller tilfældighed

4

Poissonsandsynligheder

Modellen kan ændres så der godt kan ske flere dødsfald pr. dag. I stedet for at opdele de to år i 730 dage kan man opdele i $n = 24 \cdot 730$ timer eller $n = 60 \cdot 24 \cdot 730$ minutter. Den tilsvarende sandsynlighed p , for at der sker et dødsfald i løbet af den pågældende tidsenhed, skal så vælges således at $n \cdot p = 16$. Jo højere n jo mere realistisk er modellen, men der er en grænse for hvor højt n en lommeregner som for eksempel TI-83 kan klare, når man beregner binomialsandsynligheder. Man kan imidlertid vise, at når man lader n gå mod uendelig og samtidig sørger for at $n \cdot p = \lambda$, så gælder der, at binomialsandsynligheden

$$\binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} \text{ går mod } e^{-\lambda} \cdot \frac{\lambda^x}{x!} \quad (1)$$

En sandsynlighedsfordeling af sidstnævnte type, hvor x gennemløber tallene $0, 1, 2, \dots$ kaldes en *poissonfordeling*.

- a) Benyt at $(1 + \frac{x}{n})^n$ går mod e^x , når n går mod uendelig, til at vise (1);

binomialsandsynligheden kan for eksempel først omskrives til

$$\frac{np \cdot (np - p) \cdots (np - px + p)}{x!} \cdot \left(1 - \frac{n \cdot p}{n}\right)^n \cdot \left(\frac{1}{1-p}\right)^x$$

- b) Kumulerede poissonsandsynligheder kan udregnes på TI-83. Antag at Z er poissonfordelt med $\lambda = 16$. Find $P(Z \geq 20)$ og $P(Z \geq 44)$, og sammenlign med de tilsvarende sandsynligheder i en binomialfordeling med $n = 730$ og $p = 16/730$.

Statistik – viden eller tilfældighed

5

Gevinster i Lotto

Vi vil se på, hvordan sandsynligheden for at vinde i lotto afhænger af antallet af rækker, der spilles. En lotto-række består af 7 af tallene fra 1 til 36, og der er *gevinst*, hvis mindst 4 af disse er blandt de 7 vindertal, der senere udtrækkes blandt tallene fra 1 til 36. Antallet af rigtige på en given lotto-række betegnes X .

a) Beregn $P(X = 4)$, $P(X = 5)$, $P(X = 6)$ og $P(X = 7)$.

Hvis man spiller flere rækker, vil hver enkelt række – isoleret set – have samme gevinstsandsynlighed, men samtidig ved man, at jo flere rækker man spiller jo større gevinstsandsynlighed har man.

b) Beregn sandsynligheden for gevinst, når man spiller henholdsvis 1, 10 og 100 rækker.

Statistik – viden eller tilfældighed

Fødes der færre og færre drenge i forhold til piger?

6

Test af hypoteser

I 1996 offentliggjorde Henrik Møller en artikel i *The Lancet*, hvor han gjorde opmærksom på, at der er tegn på, at der efter 1951 fødes færre og færre drenge i forhold til piger. Analysen byggede på tallene i nedenstående tabeller, som er offentligt tilgængelige:

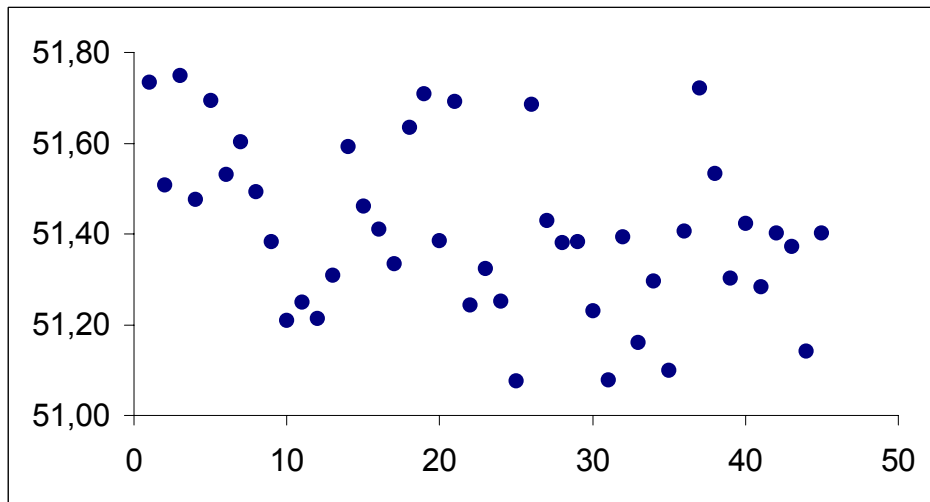
år	Antal drenge	Procent drenge
1951	39608	51,74
1952	39632	51,51
1953	40499	51,75
1954	39309	51,48
1955	39725	51,69
1956	39538	51,53
1957	38839	51,60
1958	38456	51,49
1959	37986	51,38
1960	38959	51,21
1961	39175	51,25
1962	39849	51,21
1963	42286	51,31
1964	43006	51,59
1965	44151	51,46
1966	45412	51,41
1967	41791	51,33
1968	38491	51,64
1969	36867	51,71
1970	36382	51,39
1971	38955	51,69
1972	38692	51,24

år	Antal drenge	Procent drenge
1974	36556	51,25
1975	36811	51,08
1976	33734	51,69
1977	31823	51,43
1978	31875	51,38
1979	30555	51,38
1980	29352	51,23
1981	27117	51,08
1982	27063	51,39
1983	26001	51,16
1984	26572	51,30
1985	27465	51,10
1986	28434	51,41
1987	29079	51,72
1988	30324	51,53
1989	31475	51,30
1990	32620	51,42
1991	33005	51,28
1992	34812	51,40
1993	34609	51,37
1994	35637	51,14

I det følgende betegner x antal år efter 1951, og y procent drengefødsler det pågældende år.

Et plot af punkterne (x, y) viser måske nok en faldende tendens, men ikke mere end at man nok synes, at det kunne skyldes tilfældige udsving.

Statistik – viden eller tilfældighed



Regressionslinien gennem punkterne har hældningen $-0,0054$, og vi skal nu se, hvordan man kan vurdere om denne afvigelse fra 0 kan skyldes tilfældigheder.

Hvis der ikke sker nogen ændring i sandsynligheden for at få en dreng i perioden fra 1951 til 1995, kan de 45 observationer stort set (NB antal fødsler varierer noget) opfattes som resultater af gentagelser af samme eksperiment. Et eksperiment, der giver resultater med en middelværdi på ca. 51,4 og en spredning på ca. 0,2. Den vandrette linie, der går gennem middelværdierne for hvert år, kaldes den *sande* regressionslinie, og den vil altså have hældningen 0.

Vi vil finde ud af, hvordan hældningerne fordeler sig for regressionslinier til et stort antal simulerede observationssæt, ved hjælp af et program på TI-83, der er vist nedenfor:

```

: 0 → S : 1 → K
: seq(X,X,1,45) → L1
: Input N
: While K ≤ N
: randNorm(51.4,0.2,45) → L2
: LinReg(a+bx) L1,L2,Y1
: Y1(1) - Y1(0) → H
: (abs(H) ≥ 0.0054) + S → S
: K+1 → K
: End
: Disp " P = " , S/N

```

I hvert af årene 1 til 45 (svarende til 1951 til 1995) vælger lommeregneren tilfældigt et tal fra en normalfordeling med middelværdi 51,4 og spredning 0,2. Disse tal lægges i L_2 og årene lægges i L_1 . Antallet af observationssæt kaldes N , og H er hældningen af en *simuleret* regressionslinie.

Statistik – viden eller tilfældighed

Programmet giver en P -værdi, der siger, hvor ofte den simulerede hældning har været under – 0,0054 eller over 0,0054.

Når man skal vurdere P -værdien bruger man normalt følgende regel:

Hvis P er under 0,05 forkaster man den hypotese man vil teste; i dette tilfælde hypotesen: der sker ikke nogen ændring i andelen af drengefødsler.

- a) Kør programmet med for eksempel N lig med 100 (det tager ca. 10 minutter), og lav en passende konklusion.
- b) Man kan godt lave andre, mere enkle tests: hvis y_x betegner andelen af drengefødsler x år efter 1950, kan man se på forskellene $d_x = y_{x+23} - y_x$. Ved optælling ses, at 7 ud af 22 forskelle er negative. Antallet af negative differenser kaldes D . Hvis det er rigtigt, at der ikke sker nogen ændring i sandsynligheden for at få en dreng, skulle man forvente lige mange negative og positive differenser. Beregn $P(D \leq 7)$ under forudsætning af, at sandsynligheden for en negativ differens er 0,5. Hvad ville man konkludere på baggrund af dette test med hensyn til, om der sker ændring af sandsynligheden for at føde en dreng.
- c) I ovenstående test udnytter man ikke den information, der ligger i differensernes størrelse, man ser kun på fortegnet. Lav et plot af differenserne i for eksempel Excel. Viser dette plot tegn på, at der sker ændring af sandsynligheden for at føde en dreng?

Statistik – viden eller tilfældighed

7 Misvisende sammenligninger

- Lav selv eksempler som de fiktive avisnotitser, hvor man får falske sammenhænge fordi man ikke deler op efter betydningsfulde variable.
- Find autentiske eksempler i aviser eller tidsskrifter
- Diskuter følgende tabel, der viser at man kan “vende sammenhænge om”. Her virker behandling 2 bedst i begge patientgrupper, mens behandling 1 ser ud til at virke bedst, hvis ikke man tager hensyn til, hvor “raske” patienterne er.

Behandling 1

Patient tilstand	Overlevelse		Pct. der overlever
	ja	nej	
Værre	4	6	40%
Bedre	60	40	60%
Alle	64	46	58%

Behandling 2

Patient tilstand	Overlever		Pct. der overlever
	ja	nej	
Værre	50	50	50%
Bedre	8	2	80%
Alle	58	52	53%

Konstruer selv tilsvarende eksempler.

- I det følgende eksempel har man samme overlevelse for de to behandlinger i begge patientgrupper, og samme overlevelse i den samlede gruppe. Prøv at beskrive, hvorfor det går godt at sammenligne i den samlede tabel. Prøv at formulere en generel betingelse for, at man må have lov til at slå grupper sammen.

Behandling 1

Patient tilstand	Overlevelse		Pct der overlever
	ja	nej	
Dårlig	20	30	40%
God	80	20	80%
Alle	100	50	67%

Behandling 2

Patient tilstand	Overlever		Pct med overlever
	ja	nej	
Dårlig	10	15	40%
God	40	10	80%
Alle	50	25	67%

- Spørgsmålet om sammenligne grupper, der har fået forskellige behandlinger spiller fx en meget stor rolle i den kliniske afprøvning af lægemidler. Skal man sammenligne to lægemidler, er det vigtigt, at sikre så godt man nu kan, at det ene ikke bliver givet til gruppe relativt raske patienter, mens det andet bliver givet til nogen der i udgangspunktet er sygere. Dette prøver man at forhindre ved at udføre det, der hedder randomiserede forsøg, hvor man for hver patient trækker lod om behandlingen. Herved har man ingen garanti for at få sammenlignelige grupper, men man forhindrer i hvert fald, at et bestemt middel på forhånd bevidst favoriseres. Diskuter, en sådan fremgangsmåde i lyset af resultaterne fra opgave 4.