ABCLakrids

Smilk

13 sep 2018

Introduktion til data og metode

Simon har samplet 4 poser ABC lakrids fra Haribo fra en Kvickly. Poserne er samplet ud af af en population vi kan kalde ABC lakridser som sælges i danske butikker d. 8. september, hvor populationen (N) formodes at være meget stor, hvorved vi må antage at 1/N og n/N er 0, hvilket også betyder at vores estimat for populationstotalen vil være uendelig. I stikprøven har vi oberveret hvilke bogstaver hver enhed har. Således er interessevariablen en indikatorfunktion som kan tage 29 værdier. Vi har dog måttet sande at hverken ø eller å er blevet observeret, og at det er umuligt at skeldne mellem M og W og Z og N, da det ikke fremgår klart hvordan de respektive bogstaver bør se ud. Vi har derfor antaget, at bogstaver som kan kategoriseres i to værdier fordeles til halv halv, og at bogstaver som ikke er observeret ikke forekommer i totalpopulationen. Hver pose betragtes som en sample, af størrelse n (stokastisk værdi afhænger af vægten, da antallet af bogstaver varierer i hver pose). Det gennemsnitlige antal bogstaver pr pose er 80 (mere præcist 79.75 men vi kan ikke observere halve bogstaver).

Overvejelser og problematikker med stikprøven.

Det er en række forskellige eventuelle problemer med vores stikprøve, som vi er nød til at behandle inden vi laver beregninger på data. Som nævnt ovenfor kender vi ikke populationsstørrelse N, men forventer den er stor og antager derfor at N er uendelig men dog deterministisk.

En anden problematik er bogstaver vi ikke kan kategoriserer med sikkerhed samt bogstaver vi ikke har samplet. Som nævnt ovenfor antages derfor, at vi betragter med et alfabet med 27 bogstaver (uden ø og å). Løsningen på dette, at sorterer disse som halv halv, har dog den konsekvens, at vi muligvis introducere bias i vores stikprøve, da vi ikke med sikkerhed kan sige at der findes både z'er og N'er eller W'er og M'er.

Man kan også overveje om det er unbiased alle fire poser er samplet fra samme butik på samme tidspunkt. I stedet burde man have samplet ved SRS med tilbagelægning hvilke 4 butikker i Danmark hvor poserne skulle købes, da vi jo a priori ikke ved om det kunne skabe en bias i vores sample inden vi laver eksperimentet. Grundet problematikken med at finde steder som sælger ABC lakridser, og mobalitet har vi alligevel samplet fra samme butik.

Sammenligning af interessevariable med afhabetet

For at kunne anvende corollar 2.13 i noterne, antages stikprøvestørrelsen at være fast 75, som er minimum i samplet. Hvis antallet af individer overstiger dette, fjernes tilfældigt en passende mængde individer fra stikprøven (koden for dette kan findes i bilag 3). Hvis vi sampler et individ som ikke er i delpopulationen, ville vi benytte redjection sampling og sample et nyt individ fra populationen. Dette gemmes i en ny variable `dat_n`.

For simplificering betragter vi i følgende, en gruppe af bogstaver ad gangen. fx individer med interessevaribel værende en vokal. Hvis vi ønsker at skrive et bestemt ord, kræver det ofte (i alt fald på dansk) at vi har en rimmelig mængde af vokaler til rådighed. Lad derfor $y_v(i)$ betegne indikatorfunktionen for at individ i er en vokal.

i så fald er $\bar{y}$ inden for hver pose 0.2400 0.2400 0.2533 og 0.2667, med standardavfigelse 0.0496 0.0496 0.0506 og 0.0514 igen for hver pose. Konfidens interval for hver pose er som følger,

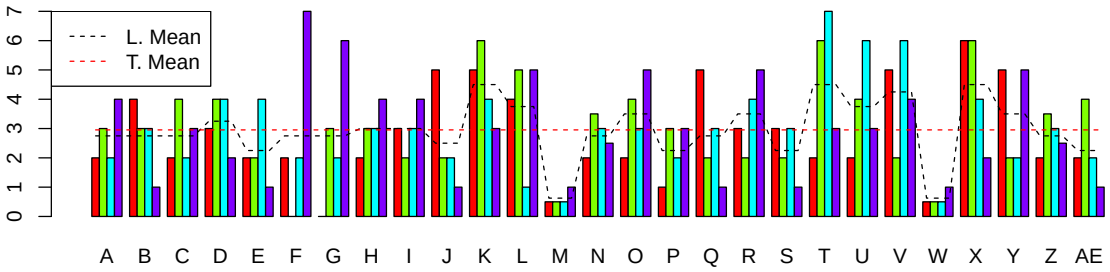


Figure 1: Bogstaver samplet fra de forskellige poser. Hver farvet bar indikerer en pose. Den sorte linje (L. Mean) viser gennemsnittet for hver bogstav over de 4 poser, mens den røde linje angiver gennemsnittet over alle bogstaver på tværs af poser

```
##         [,1]   [,2]
## [1,] 0.1427 0.3373
## [2,] 0.1427 0.3373
## [3,] 0.1542 0.3524
## [4,] 0.1659 0.3674
```

R-koden kan findes i Bilag 3 del 1 Til sammenligning forventer vi at det danske alphabet med 27 bogstaver (ink. W ekls. ø og å) og 7 vokaler har en forekomst af vokaler på 0.26. Det virker altså ikke urimeligt, at forekomsten af vokaler i stikprøverne svarer til forekomsten af vokaler i afabetet.

Ifølge sproget.dk (https://sproget.dk/temaer/ord-og-bogstaver/hvad-er-de-mest-almindelige-bogstaver-pa-dansk#kilder-kjeld-kristensen-bogstavernes d. 13/9) forekommer bogstaverne e, n, d og r hyppigere end andre bogstaver i alphabetet. Lad os derfor gøre tilsvarende som ovenfor for at undersøge om disse forekommer i en frekvens der minder om alfabetet eller, hvis poserne er designet til at man kan stave ord, vil vi forvente disse forekommer oftere end andre bogstaver. Vi får standard afvigelserne,

```
## [1] 0.0395 0.0419 0.0465 0.0403
```

og følgende konfidensintervaller

```
##         [,1]   [,2]
## [1,] 0.0559 0.2108
## [2,] 0.0712 0.2354
## [3,] 0.1089 0.2911
## [4,] 0.0609 0.2191
```

Al R-koden kan findes i bilag 3 del 2. Bemærk, at 4/27 aproksimativt 0.15 er indeholdt i samtlige konfidensintervaller. Kan vi ikke forkaste at bogstaverne e, n, d og r forekommer med samme hyppighed som i alfabetet, og det virker derfor rimeligt at overveje om fordelingen af bogstaver er ens.

Test for ens fordeling

Vi ønsker at teste, om bogstavfordelingen er den samme i de fire poser HARIBO ABC lakrids. Vi opstiller modellen

$$M_0 : X_i = (X_{i1}, \ldots, X_{ij}, \ldots, X_{i27}) \sim m(n_i, \pi) = m(n_i, (\pi_1, \ldots, \pi_{27})),$$

for $i = 1, \ldots, 4$.

og tester hypotesen

$$H_1 : \pi_1 = \cdots = \pi_4.$$

$-2lnQ(x)$-teststørrelsen regnes til

$$-2lnQ(x) = 2 * \sum_{i=1}^{4} \sum_{j=1}^{9} x_{ij} ln(\frac{x_{ij}}{e_{ij}}) = 18,509.$$

Her er e_{ij} de forventede værdier. Tabeller over disse er vedlagt i Bilag1 og Bilag2. Jævnfør Cochrans regel skal alle de forventede værdier være større eller lig 1, og højst 20% af de forventede værdier må være mindre end 5, før vi kan bruge en χ^2-approximation. I Bilag1 ses de forventede værdier, og vi ser, at de langt fra opfylder Cochrans regel, da ingen af dem er større end 5. Derfor grupperes bogstaverne i grupper af tre, og vi får da 9 grupper istedet for 27. Dette nye data samt de forventede værdier er angivet i Bilag2. Disse forventede værdier opfylder Cochrans regel, da de alle er større end 5.

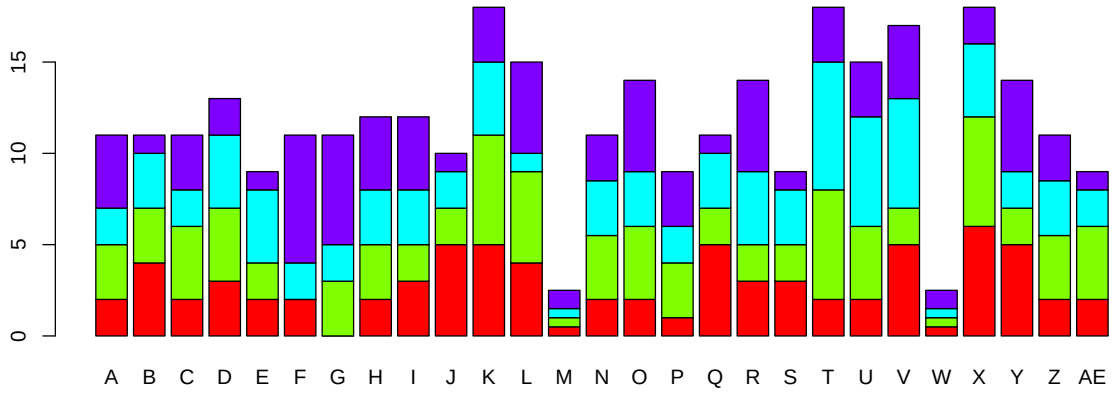
Ovenstående teststørrelse giver p-værdien

$$p_{obs}(x) = 1 - F_{\chi^2(4-1)(9-1)}(18,509) = 0,778.$$

Altså forkastes H_1 ikke. Mao. vi kan ikke forkast hypotesen om at de fire poser HARIBO ABC lakridser har samme fordeling af bogstaver.

Test for uniform fordeling af bogstaverne.

Vi vil nu teste hypotesen, om bogstaverne er uniformt fordelte. Vi vil hertil begynde med at betragte et histogram, som viser en fordeling af vores stikprøve. Et barplot summer over antallet af de forskellige bogstaver i `Dat_new` og viser således et histogram af de observerede bogstaver.



Ud fra histogrammet kan vi ikke udelukke en uniform fordeling af bogstaverne. Vi vil nu bruge Kolomogorov-Smirnov testen til at finde en sandsynlighed for, at bogstaverne er uniformt fordelte. Hertil finder vi en vektor, der indeholder tallene fra 1 til 27, som henholdsvis repræsenterer der observerede bogstaver fra `A` til `Æ`. Se billag 4 for R kode.

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  let
```

```
## D = 0.061016, p-value = 0.1791
## alternative hypothesis: two-sided
```

Sandsynligheden $p = 0.179$ forkaster ikke antagelsen om en uniform fordeling, dog argumenterer den ikke særlig stærk for en uniform fordeling af tallene.
Som nævnt, så er der ingen forskel mellem bogstaverne `M` og `W`. Vi vil derfor gentage Kolomogorov-Smirnov testen, hvor vi først lægger antallene af observerede `M`'er og `W`'er sammen. I den følgende kodestump bliver `W` transformeret til `M`. Derefter kører vi `ks.test` (Se bilag 4 del 1 for R-Kode).

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  let
## D = 0.04867, p-value = 0.4266
## alternative hypothesis: two-sided
```

Sandsynligheden $p = 0.4266$ er allerede et meget stærkere argument for en uniform fordeling. Hvis `M` og `W` antages at være den samme bogstav, vil vi altså ikke forkaste hypotesen med en større p-værdi end før.

Konklusion

Ved SRS sampling af ABClakrids i 5 poser har vi fortaget fire undersøgelser. Undersøgelse in vokalerne forekommer oftere end konsonanterne i poserne i sammenligning med fordelingen i alfabetet. Ved denne undersøgelse blev der konkluderet, at de 7 vokaler som er inkluderet i ABClakrids ikke forekommer oftere end i alfabetet. Den næste undersøgelse omhandlet om e, r, d og n forekom oftere end andre bogstaver i ABClakrids. Her gav 95 % konfidensintervallet, at e, d, r og n kunne konkluderes, at forekomme oftere i ABClakrids end andre bogstaver. Til gengæld gav test af homogenitet, at der kunne antages at bogstaverne har samme fordeling, hvilket sammen med et histogram af data førte videre til test uniform fordeling. Her blev konklusionen, at bogstaverne var uniformt fordelt.

Hvis samplings metoden skulle ændres skulle der være tages højde, at butikkerne blev random valgt og ligeledes med poserne i butikken. Derudover skulle overvejelserne omkring M og W samt N og Z, havde medført spørgsmålet: Er Z og W overhovedet i poserne, når Å og Ø ikke er? Derud over skulle overvejelserne omkring, at Å og Ø ikke forekom og om dette skulle fremkomme i datasættet.

Bilag 1

Forventede værdier				
	Pose 1	Pose 2	Pose 3	Pose 4
A	2,6	2,8	2,8	2,8
B	2,6	2,8	2,8	2,8
C	2,6	2,8	2,8	2,8
D	3,1	3,3	3,3	3,3
E	2,1	2,3	2,3	2,3
F	2,6	2,8	2,8	2,8
G	2,6	2,8	2,8	2,8
H	2,8	3,1	3,0	3,0
I	2,8	3,1	3,0	3,0
J	2,4	2,6	2,5	2,5
K	4,2	4,6	4,6	4,6
L	3,5	3,9	3,8	3,8
M	0,6	0,6	0,6	0,6
N	2,6	2,8	2,8	2,8
O	3,3	3,6	3,6	3,6
P	2,1	2,3	2,3	2,3
Q	2,6	2,8	2,8	2,8
R	3,3	3,6	3,6	3,6
S	2,1	2,3	2,3	2,3
T	4,2	4,6	4,6	4,6
U	3,5	3,9	3,8	3,8
V	4,0	4,4	4,3	4,3
W	0,6	0,6	0,6	0,6
X	4,2	4,6	4,6	4,6
Y	3,3	3,6	3,6	3,6
Z	2,6	2,8	2,8	2,8
Æ	2,1	2,3	2,3	2,3

Bilag 2

Data - indelt i grupper af 3					
	Pose 1	Pose 2	Pose 3	Pose 4	Rækkesum
A-C	8	10	7	8	33
D-F	7	6	10	10	33
G-I	5	8	8	14	35
J-L	14	13	7	9	43
M-O	4,5	8	6,5	8,5	27,5
P-R	9	7	9	9	34
S-U	7	12	16	7	42
V-X	11,5	8,5	10,5	7	37,5
Y-Æ	9	9,5	7	8,5	34
Søjlesum	75	82	81	81	319

Forventede værdier					
	Pose 1	Pose 2	Pose 3	Pose 4	Rækkesum
A-C	7,8	8,5	8,4	8,4	33,0
D-F	7,8	8,5	8,4	8,4	33,0
G-I	8,2	9,0	8,9	8,9	35,0
J-L	10,1	11,1	10,9	10,9	43,0
M-O	6,5	7,1	7,0	7,0	27,5
P-R	8,0	8,7	8,6	8,6	34,0
S-U	9,9	10,8	10,7	10,7	42,0
V-X	8,8	9,6	9,5	9,5	37,5
Y-Æ	8,0	8,7	8,6	8,6	34,0
Søjlesum	75,0	82,0	81,0	81,0	319,0

Bilag 3

R kode for at hente data og lave figur 1

R kode til sikre fast stikprøvestørrelsetil beregningerne

```
set.seed(10503) # For others to get same result
s <- sample(c(LETTERS,"AE"),sum(n-min(n)),replace = TRUE) #Sample values to be removed

n-min(n) #antal enheder for meget
```

```
## [1] 0 7 6 6
```

```
dat_n <- Dat_new # create new dataset
dat_n[s[1:((n-min(n))[2])],2] <- dat_n[s[1:((n-min(n))[2])],2]-1
dat_n[s[1:((n-min(n))[3])],3] <- dat_n[s[1:((n-min(n))[3])],3]-1
dat_n[s[1:((n-min(n))[4])],4] <- dat_n[s[1:((n-min(n))[4])],4]-1

# test if we need to sample any values again
any(dat_n<0)
```

```
## [1] FALSE
```

Del 1

Beregninger for gennemsnittet af interessevariablen vokal eller ej.

```
k <- 75
id <- ifelse(rownames(dat_n)%in% c("A","E","I",
                                   "O","U","Y",
                                   "AE"), TRUE,FALSE)
y_bar <- colSums(dat_n[id,])/k
var_emp <- (1-0)*y_bar*(1-y_bar)/(k-1)

uu <- qnorm(.975)
CI <- cbind(y_bar-uu*sqrt(var_emp),y_bar+uu*sqrt(var_emp))
#CI
round(CI,4)
```

```
##        [,1]   [,2]
## [1,] 0.1427 0.3373
## [2,] 0.1427 0.3373
## [3,] 0.1542 0.3524
## [4,] 0.1659 0.3674
```

#Del 2 Beregninger for hyppigt forekomne bogstaver.

```
id2 <- ifelse(rownames(dat_n)%in% c("E","N","D","R"),
                TRUE,FALSE)
y_endr <- colSums(dat_n[id2,])/k

var_endr <- y_endr*(1-y_endr)/(k-1)
round(sqrt(var_endr),4)
```

```
## [1] 0.0395 0.0419 0.0465 0.0403
```

```r
CI <- cbind(y_endr-uu*sqrt(var_endr),y_endr+uu*sqrt(var_endr))
round(CI,2)
```

```
##      [,1] [,2]
## [1,] 0.06 0.21
## [2,] 0.07 0.24
## [3,] 0.11 0.29
## [4,] 0.06 0.22
```

Bilag 4

```r
Dat_new2 <- Dat_new[-(28:29),]
let <- integer(sum(Dat_new2))
i = 1

for (k in 1:4){
  for (j in 1:27){
    while (Dat_new2[j,k] > 0){
      let[i] = j
      Dat_new2[j,k] = Dat_new2[j,k] - 1
      i = i+1
    }
  }
}

suppressWarnings(ks.test(let, "punif", 1, 27))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  let
## D = 0.061016, p-value = 0.1791
## alternative hypothesis: two-sided
```

del 1

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  let
## D = 0.04867, p-value = 0.4266
## alternative hypothesis: two-sided
```