Udregning af statistisk sikkerhed med angivelse af mest sandsynlige antal positive prøvebrønde for undersøgelse af virusangreb i kartoffelknolde

Rådgivningsnotat fra DCA – National Center for Fødevarer og Jordbrug

Niels Holst, Institut for Agroøkologi, Aarhus Universitet

AARHUS
UNIVERSITET
DCA - NATIONALT CENTER FOR FØDEVARER OG JORDBRUG

Datablad

Titel:	Udregning af statistisk sikkerhed med angivelse af mest sandsynlige antal positive prøvebrønde for undersøgelse af virusangreb i kartoffelknolde
Forfatter(e):	Senior Forsker Niels Holst, Institut for Agroøkologi
Fagfællebedømmelse:	Lektor René Gislum, Institut for Agroøkologi
Kvalitetssikring, DCA:	AC Fuldmægtig Susanne Hansen, DCA Centerenheden
Rekvirent:	Landbrugsstyrelsen, Ministeriet for Fødevarer, Landbrug og Fiskeri (FVM)
Dato for bestilling/levering:	01.04.2022 / 12.04.2022
Journalnummer:	2022-0355951
Finansiering:	Besvarelsen er udarbejdet som led i "Rammeaftale om forskningsbaseret myndighedsbetjening" indgået mellem Ministeriet for Fødevarer, Landbrug og Fiskeri (FVM) og Aarhus Universitet under ID nr. 1.50 i "Ydelsesaftale Planteproduktion 2022-2025".
Ekstern kommentering:	Nej.
Eksterne bidrag:	Nej.
Kommentarer til besvarelse:	Notatet præsenterer resultater, som ved notatets udgivelse ikke har været i eksternt peer review eller er publiceret andre steder. Ved en evt. senere publicering i tidsskrifter med eksternt peer review vil der derfor kunne forekomme ændringer.
Citeres som:	Holst, N. 2022. Udregning af statistisk sikkerhed med angivelse af mest sandsynlige antal positive prøvebrønde for undersøgelse af virusangreb i kartoffelknolde. Antal sider: 12 Rådgivningsnotat fra DCA – Nationalt Center for Fødevarer og Jordbrug, Aarhus Universitet, leveret: 12.04.2022.
Rådgivning fra DCA:	Læs mere på https://dca.au.dk/raadgivning/

Baggrund

Landbrugsstyrelsen har i en bestilling sendt til DCA – Nationalt Center for Fødevarer og Jordbrug ved Århus Universitet (AU) ønsket en undersøgelse af om den statistiske sikkerhed Landbrugsstyrelsen (LFST) bruger til at at angive virusfund i en knoldprøve på analysebevis fra Fødevarestyrelsens (FVST) laboratorie kan målrettes med virustolerancerne, som er angivet i bekendtgørelse nr. 1050 af 30.05.2021 om læggekartofler og avl af konsumkartofler ved brug af flerer eller færrer prøvebrønde end der bruges i dag.

LFST udtager kartoffelknolde til undersøgelse af angreb af bladrullevirus og kartoffelvirus Y. For hvert høstnummer udtages 110 knolde, som udgør én prøve. Der udtages én knoldprøve pr. påbegyndt syv ha. for hvert høstnummer for udvalgte sorter i i basisavlen og, hvor arealet er mere end 0,1 ha. Prøverne tages repræsentativt i hele marken og fra forskellige planter. Prøven sendes til FVST laboratorie, som tester 10 delprøver i 10 prøvebrønde.

Undersøgelsen af virusangivelserne ønskes med henblik på at have samme grænseværdier på analysebeviset som i bekendtgørelsen, for at øge sammenhængen mellem tolerancer og virusfund. Især tolerancen for certificerede klasse A på 8 % er relevant.

Undersøgelsen skal også omfatte en angivelse af det mest sandsynlige antal af positive prøvebrønde for hver model.

Besvarelse

1.1 Problemformulering

Der modtages en sæk med kartofler (knolde), som skal testes for forekomsten af virus. Grænseværdier kan fx være defineret som følger (fig. 1):

Virusundersøgelse – tilladt forekomst af bladrullevirus og kartoffelvirus Y (pct. af planter/knolde)

	Præ-basis læggekartofler		Basis læggekartofler			Certificerede læggekartofler
Klasse	PBTC	PB	S	SE	E	A
Virus i alt	0	0,5	1,0	2,0	2,0	8,0

Figur 1: Grænseværdier for tilladt forekomst af bladrullevirus og kartoffelvirus Y (Landbrugsstyrelsen, 2021).

En grænseværdi på fx $p = 8\%$ fortolkes sådan, at der i marken højst må være 8% smittede knolde. En knold anses kun for smittet, såfremt, at *hvis* den blev testet, ville den give et positivt resultat.

Fra sækken udtages N delprøver hver med n knolde. Der udtages altså i alt $N \cdot n$ knolde. Hver delprøve homogeniseres og testes i en prøvebrønd for sig. Resultat består dermed af i alt N tests, som hver er enten positiv eller negativ.

Problemet er at beregne: Hvad indikerer et bestemt antal positive delprøver (N_{pos}; $0 \leq N_{pos} \leq N$) vedrørende forekomsten af virus i marken (p; andelen af knolde i marken, som er smittet, hvor $0 \leq p \leq 100\%$) ?

Spørgsmålet besvares ved at tabellægge to parametre for givne værdier af n og N (og med $N_{pos} = 0..N$):

- $\hat{p}$: den mest sandsynlige værdi for forekomsten af virus i marken.
- P_α: sandsynligheden for at den sande forekomst af virus i marken overskrider grænseværdien α, hvor α antager værdierne 0,5%, 1%, 2% og 8%.

Der fremstilles tre tabeller i rapporten for værdierne:

- $n = 10$ og $N = 10$.
- $n = 5$ og $N = 20$.
- $n = 4$ og $N = 25$.

Der vedlægges desuden et R-script, som kan anvendes til at fremstille tabeller for vilkårlige værdier af n og N (bilag 1).

1.2 Beregning

1.2.1 Metode

Løsningen beror på simpel kombinatorik, som udledes i det følgende. Det vedlagte R-script producerer alle de viste tabeller og figurer.

1.2.2 Udledning

Sandsynligheden, for at en delprøve er positiv, er 1 minus sandsynligheden for, at ingen af de n knolde er smittede:

$$P_{pos}(p,n) = 1 - (1-p)^n \tag{1}$$

Fx $P_{pos}(0{,}08;10) = 0{,}57$. Dvs. hvis den sande forekomst er 8%, så vil en delprøve bestående af 10 knolde være positiv i 57% af tilfældene.

Hvis vi nu gentagne gange udfører en test bestående af N delprøver, hver med n knolde med en virusforekomst p, så vil antallet af positive delprøver (N_{pos}) i en test følge binomialfordelingen,

$$\mathcal{F}(p,n,N) = \binom{N}{N_{pos}} f^{N_{positiv}}(1-f)^{N-N_{pos}}, \tag{2}$$

hvor $f = P_{pos}(p,n)$.

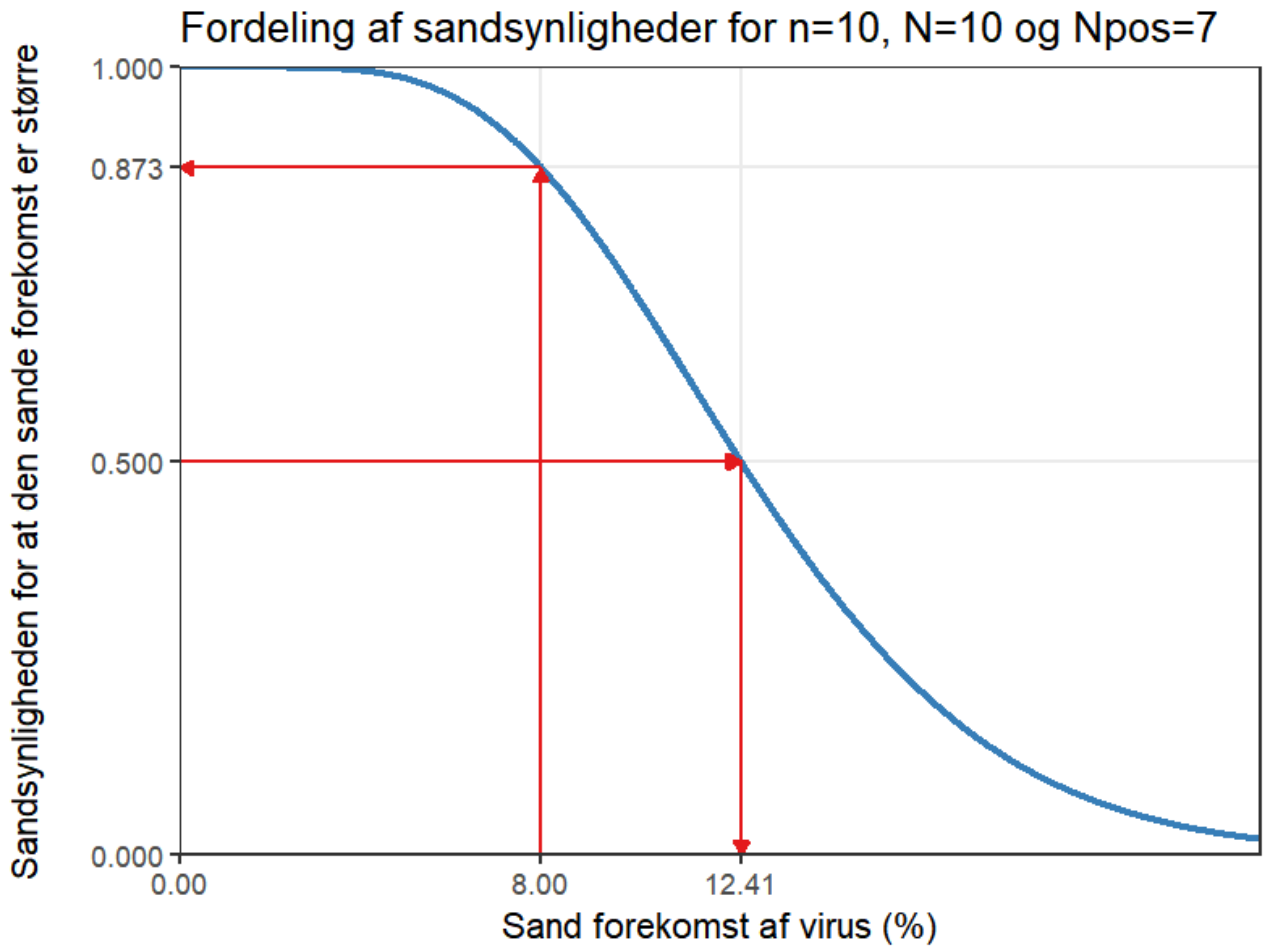
Til eksempel får vi med $\mathcal{F}(0{,}08;10;10)$ følgende fordeling af antallet af positive delprøver (N_{pos}) (tabel 1):

Tabel 1: Fordelingen af antallet af positive delprøver

Antal positive delprøver	Sandsynlighed for udfaldet	Summeret sandsynlighed
0	0,000	0,000
1	0,003	0,003
2	0,018	0,022
3	0,063	0,085
4	0,144	0,229
5	0,226	0,455
6	0,245	0,700
7	0,182	0,882
8	0,089	0,971
9	0,026	0,997
10	0,003	1,000

Vi regner nu denne tabel ud 10.000 gange med alle værdier for den sande virusforekomst, $p = \frac{0}{10000}, \frac{1}{10000}, \frac{2}{10000}, \cdots, \frac{10000}{10000}$. Vi gør det endda flere gange, nemlig for alle de mulige værdier af $N_{pos} = 0..N$.

I den følgende figur 2 vises resultatet for $N_{pos} = 7$ stadig med $n = 10$ og $N = 10$. Altså, hvis vi har 10 delprøver hver bestående af 10 knolde, og vi finder 7 positive delprøver, hvad véd vi da om den sande forekomst (p) i marken?



Figur 2: Sandsynligheden for den sande forekomst af virus på marken ved $N_{pos} = 7$ med $n = 10$ og $N = 10$.

Y-aksen viser sandsynligheden for, at den sande forekomst (p) er større end værdien på x-aksen. Den starter med y-værdien 1: Sandsynligheden for at $p > 0\%$, når vi har fundet 7 positive delprøver, er naturligvis 1.

Lad os gå ned ad y-aksen til værdien y=0,50. Den tilsvarende x-værdi er 12,41%. Dvs. at sandsynligheden for at den sande forekomst (p) er større end 12,41% er 0,5. Sandsynligheden for, at p er mindre end 12,41%, er ligeledes 0.5. Således har vi fundet den mest sandsynlige værdi, $\hat{p} = 12{,}41\%$. Altså, med 7 positive delprøver ud af 10 mulige (hver med 10 knolde) er den mest sandsynlige sande forekomst 12,41%.

Grænseværdierne for virusforekomsten findes på x-aksen. I figuren er angivet grænseværdien, $p = 8\%$. Aflæser vi kurven for denne værdi, finder vi på y-aksen værdien 0,873. Dvs. med 7 positive delprøver ud af 10 mulige (hver med 10 knolde), så er der en sandsynlighed på 0,873 for, at grænseværdien på 8% er overskredet. Vi har således bestemt sandsynligheden $P_{8\%} = 0{,}873$.

Vi kan nu efter samme metode fremstille tabeller for $\hat{p}$ og N_α for alle værdier af $N_{pos} = 0..N$ og udvalgte værdier af α og n.

1.3 Tabeller

Ifølge den nuværende procedure udtages i alt 100 knolde, som opdeles i $N = 10$ delprøver á hver $n = 10$ kartofler. Her fremstilles tabeller for denne opdeling af prøven samt for to alternative opdelinger: ($N = 20$, $n = 5$) og ($N = 25, n = 4$).

Tabel 2: Tabel for 10 delprøver á 10 knolde

Antal positive delprøver	Forventet forekomst (%)	Sandsynligheden for at forekomsten er større end			
		0,5%	1%	2%	8%
0	0,68	0,597	0,359	0,129	0
1	1,74	0,914	0,748	0,422	0,003
2	2,92	0,989	0,935	0,722	0,020
3	4,24	0,999	0,988	0,904	0,080
4	5,76	1	0,999	0,976	0,219
5	7,53	1	1	0,996	0,441
6	9,68	1	1	0,999	0,686
7	12,41	1	1	1	0,873
8	16,23	1	1	1	0,967
9	22,76	1	1	1	0,996
10	62,42	1	1	1	1

Tabel 2 aflæses således: Hvis man fx har fundet 1 positive delprøve ud af de 10 mulige, så svarer det til en forventet (eller mest sandsynlig) forekomst på $\hat{p} = 1{,}74\%$. Det er usandsynligt, at grænseværdien på 8% er overskredet idet $P_{8\%} = 0{,}003$, men ganske sandsynligt at grænseværdien på 2% er overskredet da $P_{2\%} = 0{,}422$.

Tabel 3: Tabel for 20 delprøver á 5 knolde

Antal positive delprøver	Forventet forekomst (%)	Sandsynligheden for at forekomsten er større end			
		0,5%	1%	2%	8%
0	0,68	0,597	0,359	0,129	0.000
1	1,69	0,910	0,738	0,408	0,003
2	2,75	0,987	0,925	0,692	0,014
3	3,86	0,999	0,984	0,877	0,050
4	5,03	1	0,997	0,961	0,129
5	6,26	1	1	0,990	0,262
6	7,55	1	1	0,998	0,437
7	8,93	1	1	1	0,621
8	10,39	1	1	1	0,778
9	11,95	1	1	1	0,888
10	13,63	1	1	1	0,952
11	15,46	1	1	1	0,983
12	17,46	1	1	1	0,995
13	19,67	1	1	1	0,999
14	22,16	1	1	1	1
15	25,01	1	1	1	1
16	28,37	1	1	1	1
17	32,52	1	1	1	1
18	38,03	1	1	1	1
19	46,65	1	1	1	1
20	74,85	1	1	1	1

Tabel 3 aflæses således: Hvis man fx har fundet 6 positive delprøver ud af de 20 mulige, så svarer det til en forventet (eller mest sandsynlig) forekomst på $\hat{p} = 7{,}55\%$. Sandsynligheden for, at den sande forekomst er større end 8%, er $P_{8\%} = 0{,}437$, mens det er omtrent sikkert, at den sande forekomst er større end 2%, da $P_{2\%} = 0{,}998$.

Tabel 4: Tabel for 25 delprøver á 4 knolde

Antal positive delprøver	Forventet forekomst (%)	Sandsynligheden for at forekomsten er større end			
		0,5%	1%	2%	8%
0	0,68	0,597	0,359	0,129	0
1	1,68	0,909	0,736	0,405	0,002
2	2,72	0,986	0,924	0,687	0,013
3	3,80	0,998	0,983	0,871	0,045
4	4,91	1	0,997	0,958	0,116
5	6,07	1	1	0,989	0,236
6	7,27	1	1	0,997	0,396
7	8,52	1	1	1	0,571
8	9,83	1	1	1	0,729
9	11,19	1	1	1	0,849
10	12,62	1	1	1	0,926
11	14,12	1	1	1	0,968
12	15,71	1	1	1	0,988
13	17,39	1	1	1	0,996
14	19,18	1	1	1	0,999
15	21,09	1	1	1	1
16	23,16	1	1	1	1
17	25,41	1	1	1	1
18	27,89	1	1	1	1
19	30,65	1	1	1	1
20	33,78	1	1	1	1
21	37,43	1	1	1	1
22	41,86	1	1	1	1
23	47,62	1	1	1	1
24	56,26	1	1	1	1
25	79,69	1	1	1	1

Referencer

Landbrugsstyrrelsen, 2021: Bilag 9, *Bekendtgørelse om læggekartofler og avl af konsumkartofler*, nr 1050 af 30.05.2021. https://www.retsinformation.dk/eli/lta/2021/1050

Bilag

Bilag 1: R-script

```
# Statistisk sikkerhed for undersøgelse af virusangreb i kartoffelknolde
# Niels Holst (niels.holst@agro.au.dk)
# 7. april 2022

# Ryd op
graphics.off()
rm(list=ls(all=TRUE))

# Indlæs biblioteker; disse skal være installeret
library(ggpubr)
library(ggplot2)
library(plyr)

# Farver
red = '#e41a1c'
blue = '#377eb8'

# Parametre:
#
# Sand frekvens af virussmittede knolde (p)
# Antal knolde i én prøve (n)
# Antal delprøver (én delprøve i hver prøvebrønd) (N)

# Angiv mappen hvor tabelfilerne skrives i ("~" angiver mappen, "Dokumenter")
setwd("~/ARKIV/Myndighedsbetjening/2022-04-06 kartoffeltest")

# Sandsynlighed for en positiv delprøve
prob_positiv = function(p, n) {
  # 1 minus sandsynligheden for at ingen knolde i prøven var smittede
  1 - (1-p)^n
}
# Eksempel
prob_positiv(0.08, 10)

# Fordeling af prøveudfald fra N prøvebrønde
prob_fordeling = function(p, n, N) {
  x = 0:N
  M = data.frame(
    AntalPositiveBroende = x,
    Sandsynlighed = dbinom(x, size=N, prob=prob_positiv(p,n))
  )
```

```
  M$Sum = cumsum(M$Sandsynlighed)
  M = round(M, 3)
  write.table(M,paste0("tabel-fordeling.txt"), row.names=FALSE, sep="\t")
  M
}
# Eksempel
prob_fordeling(0.08, 10, 10)

# Hvis vi kender antallet af positive brønde (Npos),
# hvad er sandsynligheden for at få dette resultat,
# givet den sande frekvens af virussmittede knolde (p) ?
prob_resultat = function(Npos, p, n, N) {
  dbinom(Npos, size=N, prob=prob_positiv(p,n))
}
# Eksempel
prob_resultat(7, 0.08, 10, 10)

# Hvis vi kender antallet af positive brønde (Npos),
# hvad er sandsynlighedsfordeling blandt virussmittede knolde (p),
# som fører til dette resultat ?
prob_resultat_fordeling = function(Npos, n, N) {
  p = (0:10000)/10000
  M = data.frame(
    SandFrekvens = 100*p,
    ProbPositiv  = prob_resultat(Npos, p, n, N)
  )
  M$SumProbPositiv = cumsum(M$ProbPositiv)
  M$SumProbPositiv = M$SumProbPositiv/M$SumProbPositiv[10001]
  M
}

# Vis en figur over sandsynlighedsfordelingen
plot_resultat_fordeling = function(Npos, n, N) {

  # Generér fordeling
  M = prob_resultat_fordeling(Npos, n, N)

  # Find punkter på kurve
  i08 = 1
  while (M$SandFrekvens[i08] < 8.0) i08 = i08 + 1
  i50 = 1
  while (M$SumProbPositiv[i50] < 0.50) i50 = i50 + 1
  xfinal = 10001
  while (M$SumProbPositiv[xfinal] > 0.98) xfinal = xfinal - 1

  # Plot kurve med pile på punkter
ggplot(M[1:xfinal,], aes(SandFrekvens)) +
    geom_line(aes(y=1-SumProbPositiv), colour=blue, size=1) +
    annotate("segment", y=0.50, yend=0.50, x=0, xend=M$SandFrekvens[i50], colour=red,
      arrow=arrow(length=unit(0.02, "npc"), type="closed")) +
    annotate("segment", y=0.50, yend=0,    x=M$SandFrekvens[i50], xend=M$SandFrekvens[i50], colour=red,
      arrow=arrow(length=unit(0.02, "npc"), type="closed")) +
    annotate("segment", y=0, yend=1-M$SumProbPositiv[i08], x=M$SandFrekvens[i08], xend=M$SandFrekvens[i08],
colour=red,
      arrow=arrow(length=unit(0.02, "npc"), type="closed")) +
    annotate("segment", y=1-M$SumProbPositiv[i08], yend=1-M$SumProbPositiv[i08], x=M$SandFrekvens[i08], xend=0,
colour=red,
```

```
    arrow=arrow(length=unit(0.02, "npc"), type="closed")) +
  scale_x_continuous(expand=c(0,0), limits=c(0,NA), breaks=c(0,round(M$SandFrekvens[c(i08, i50)], 4))) +
  scale_y_continuous(expand=c(0,NA), limits=c(0,1), breaks=c(0, 0.5, 1, round(1-M$SumProbPositiv[i08], 3))) +
  labs(
    x="Sand forekomst af virus (%)",
    y="Sandsynligheden for at den sande forekomst er større\n",
    title=paste0("Fordeling af sandsynligheder for n=",n, ", N=", N, " og Npos=", Npos)
  ) +
  theme_bw() +
  theme(
    panel.grid.minor = element_blank()
  )
}
# Eksempel
plot_resultat_fordeling(7,10,10)

# Lav en linie til tabel
tabel_linie = function(Npos, n, N) {
 M = prob_resultat_fordeling(Npos, n, N)
 i50 = 1
 while (M$SumProbPositiv[i50] < 0.50) i50 = i50 + 1
 i005 = 1
 while (M$SandFrekvens[i005]< 0.5) i005 = i005 + 1
 i01 = i005
 while (M$SandFrekvens[i01] < 1.0) i01  = i01 + 1
 i02 = i01
 while (M$SandFrekvens[i02] < 2.0) i02  = i02 + 1
 i08 = i02
 while (M$SandFrekvens[i08] < 8.0) i08  = i08 + 1
 c(
   M$SandFrekvens[i50],
   1-M$SumProbPositiv[i005],
   1-M$SumProbPositiv[i01],
   1-M$SumProbPositiv[i02],
   1-M$SumProbPositiv[i08])
}
# Eksempel
tabel_linie(7,10,10)

# Vis tabel og gem i fil
tabel = function(n,N) {
 M = adply(0:N, 1, tabel_linie, n, N)
 M[,1] = 0:N
 M[,3:6] = round(M[,3:6], 3)
 colnames(M) = c("Npos", "p_exp", "P0.5", "P1.0", "P2.0", "P8.0")
 write.table(M,paste0("tabel-",n,"-",N, ".txt"), row.names=FALSE, sep="\t")
 M
}

# Fremstil tre tabeller
tabel(10,10)
tabel( 5,20)
tabel( 4,25)
```