Udregning af statistisk sikkerhed for undersøgelse af virusangreb i kartoffelknolde

Rådgivningsnotat fra DCA – Nationalt Center for Fødevarer og Jordbrug

Af Niels Holst

Institut for Agroøkologi, Aarhus Universitet

AARHUS
UNIVERSITET
DCA - NATIONALT CENTER FOR FØDEVARER OG JORDBRUG

Datablad

Titel:	Udregning af statistisk sikkerhed for undersøgelse af virusangreb i kartoffelknolde
Forfatter(e):	Seniorforsker Niels Holst, Institut for Agroøkologi, AU
Fagfællebedømmelse:	Lektor René Gislum, Institut for Agroøkologi, AU
Kvalitetssikring, DCA:	Specialkonsulent Lene Hegelund, DCA Centerenheden, AU
Rekvirent:	Landbrugsstyrelsen, Ministeriet for Fødevarer, Landbrug og Fiskeri (FVM)
Dato for bestilling/levering:	02.12.2021 / 16.12.2021
Journalnummer:	2021-0317766
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 "Ydelsesaftale Planteproduktion 2021-2024".
Ekstern kommentering:	Nej.
Eksterne bidrag:	Nej.
Kommentarer til besvarelse:	Der er vedlagt et R-script til besvarelsen, til produktion af alle de viste tabeller og figurer. Scriptet kan udleveres ved henvendelse til forfatteren.
Citeres som:	Holst N. 2021. Udregning af statistisk sikkerhed for undersø-gelse af virusangreb i kartoffelknolde. 11 sider. Rådgivningsnotat fra DCA – Nationalt Center for Fødevarer og Jordbrug, Aarhus Universitet, leveret: 16. december 2021.
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 ønsket en undersøgelse af den statistiske sikkerhed, som Landbrugsstyrelsen bruger til at angive virusfund i knoldprøver på analysebeviser fra Fødevarestyrelsens laboratorium.

Det fremgår af bestillingen, at laboratoriet modtages sække med kartofler (knolde), som skal testes for forekomsten af virus, og at grænseværdierne aktuelt er defineret som følger:

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

Besvarelse

Det ses af Landbrugsstyrelsens tabel, at den tilladte forekomst af virus i læggekartofler læggekartofler ligger mellem 0% for præbasis læggekartofler (PBTC) og 8.0% for certificerede læggekartofler (A).

En grænseværdi på fx $p = 0.08 = 8\%$ fortolkes sådan, at der i marken højst må være 8% smittede knolde. En knold anses kun for smittet, hvis den testes positiv.

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 demed af udkommet fra N prøvebrønde, som hver er enten positiv eller negativ.

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

Da forekomsten (p) skal beregnes ud fra en stikprøve på $N \cdot n$ knolde, kan et given resultat (N_{pos}) ikke omsættes til én præcis værdi for forekomsten; den må udtrykkes som et sikkerhedsinterval $[p_{min}, p_{max}]$, som endvidere afhænger af sikkerhedsniveauet (α). Hvis man fx vælger $\alpha = 2\%$, betyder det, at givne værdier for n, N og N_{pos} indikerer, at forekomsten (p) ligger inden for intervallet $[p_{min}, p_{max}]$ med en sikkerhed på 2%. Dvs. man accepterer, at i 2% af tilfældene vil resultatet blive fejlfortolket, idet den sande forekomst (p) ligger uden for intervallet.

Opgaven besvares ved at opstille tabeller, som med givne værdier for n, N, N_{pos} og α angiver det resulterende sikkerhedsinterval $[p_{min}, p_{max}]$.

Metode for beregning

Opstilling af tabeller med værdier for n, N, N_{pos} og et sikkerhedsinterval $[p_{min}, p_{max}]$ beror på simpel kombinatorik udledt i det følgende. Til notatet er vedlagt et R-script, der kan producerer alle de viste tabeller og figurer.

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

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

Fx $P_{positiv}(0.08{,}10) = 0.57$. 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 udføre en test (en test = en sæk kartofler) bestående af N delprøver, hver med n knolde med samme virusforekomst p, så vil antallet af positive delprøver ($N_{positiv}$) i en test følge binomialfordelingen,

$$\mathcal{F}(p,n,N) = \binom{N}{N_{positiv}} f^{N_{positiv}}(1-f)^{N-N_{positiv}}, \quad \text{hvor } f = P_{positiv}(p,n). \tag{2}$$

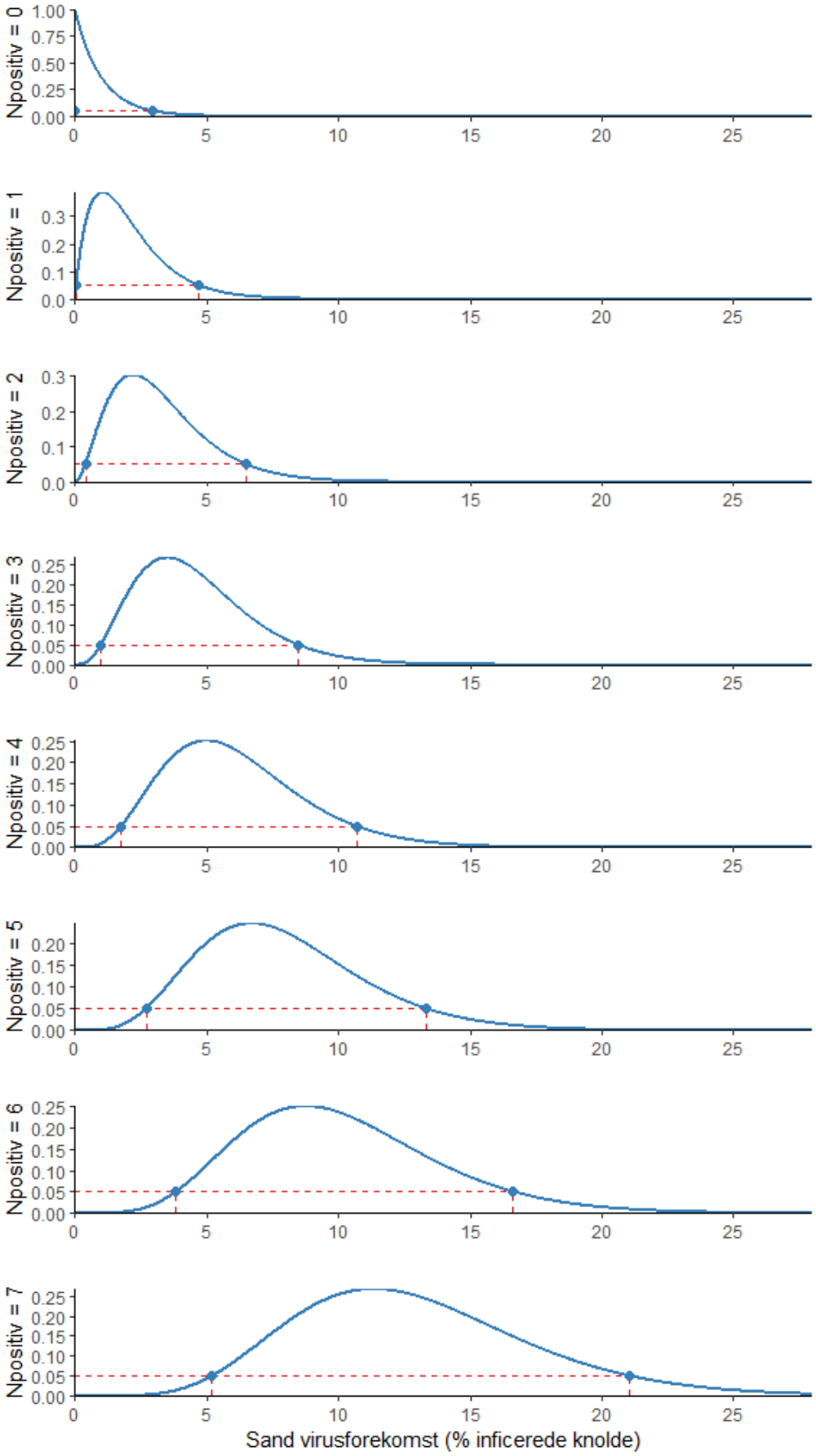
Til eksempel får vi med $\mathcal{F}(0.08{,}10{,}10)$ følgende fordeling af antallet af positive delprøver ($N_{positiv}$):

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

Således ses, at der i 14.4% af tilfældene vil være 4 positive delprøver, når materialet har en sand forekomst på 8%, og der udtages 10 delprøver hver bestående af 10 knolde. I gennemsnit vil man forvente $P_{positiv}(p,n) \cdot N$ positive delprøver, hvilket passer med, at tabellens mest sandsynlige udfald er 6 positive delprøver, idet $P_{positiv}(0.08{,}10) \cdot 10 = 5.66$.

Hvis vi fx fandt 7 positive delprøver, så ville vi forvente dette resultat (hvis vi kendte den sande forekomst, $p = 8\%$) 18.2% af de gange, vi udførte testen i følge binomialfordelingen i tabellen oven for. Men vi kender ikke p; det er den, vi skal bestemme. Vi regner derfor tabellen ud 10,000 gange med alle værdier for den sande virusforekomst, $p = \frac{0}{10000}, \frac{1}{10000}, \frac{2}{10000}, \cdots, \frac{10000}{10000}$. På næste side vises resultatet for 10 delprøver ($N = 10$) á 10 knolde ($n = 10$). For overskuelighedens skyld er kun vist $N_{positiv} \leq 7$.

Nederst i figuren på næste side kan vi se, hvad p sandsynligvis ville være, hvis vi havde fundet 7 positive delprøver. Allermest sandsynligt ville p være ca. 12% (toppen af kurven). Hvis vi udelukker de mindst sandsynlige værdier, defineret som dem, der er mindre end $\alpha = 5\%$ sandsynlige (den vandrette stiplede linie), så finder vi, at den sande virusforekomst med 7 positive delprøver mest sandsynligt ligger i intervallet, $5.2\% < p < 21.0\%$ (afgrænset af de lodrette stiplede linier).


Npositiv = 0
Npositiv = 1
Npositiv = 2
Npositiv = 3
Npositiv = 4
Npositiv = 5
Npositiv = 6
Npositiv = 7
Sand virusforekomst (% inficerede knolde)

Sammenligning med tidligere sikkershedsintervaller

Det fremgår af bestillingens bilag 1, at Landbrugsstyrelsen aktuelt anvender nedenstående tabel med grænseværdier, som er baseret på 10 delprøver ($N = 10$) á 10 knolde ($n = 10$). Den antagne α-værdi samt metoden bag tabellens udledning er ukendte.

Statistical uncertainty (in sampling) and number of tuber tested

Subsamples (n)	Cores in a subsamples (n)	Infected subsamples (n)	Sample infection (%)	Lower detection (%)	Higher detection (%)
10	10	0	0	0,00000	3,62167
10	10	1	1	0,02531	5,71816
10	10	2	2	0,25501	7,80042
10	10	3	4	0,68833	10,02924
10	10	4	5	1,28763	12,52323
10	10	5	7	2,05000	15,43237
10	10	6	9	2,99740	19,00163
10	10	7	12	4,18028	23,71515
10	10	8	15	5,69930	30,79164
10	10	9	21	7,77735	45,00950
10	10	10	100	11,09522	100,00000

Calculated by a statiscian, for fields were the virus expect to be even distributed in the field.

If you want to lower the detection limit, you need more potatoes.

Hvis vi bruger metoden fra foregående afsnit og vælger at sætte $\alpha = 2\%$, får vi en meget lignende tabel:

Antal positive delprøver	Minimum virusforekomst (%)	Maksimum virusforekomst (%)
0	0	3.83
1	0.03	5.84
2	0.24	7.86
3	0.67	10.06
4	1.27	12.55
5	2.04	15.48
6	2.99	19.13
7	4.17	24.00
8	5.65	31.37
9	7.63	46.18
10	10.67	100

Forskellene kan tilskrives større numerisk usikkerhed i beregningsprocessen i den tidligere tabel.

Beregning af nye sikkerhedsintervaller

Til besvarelsen er vedlagt et R-script, der kan benyttes til at producere tabeller med vilkårlige værdier for n, N og α. Som eksempel følger seks tabeller **alle med $\alpha = 2\%$**. I tabel 1 til 3 er det samlede antal knolde 100. I tabel 4 til 6 er det samlede antal knolde 200. I begge tilfælde er knoldene blevet opdelt i $N = 20, 10$ og 5 delprøver (svarende til 20, 10 og 5 prøvebrønde).

Som bemærket i fodnoten til Landbrugsstyrelsens nuværende tabel (se *Sammenligning med tidligere sikkershedsintervaller*), så skal der indsamles flere kartofler for at opnå en lavere detektionsgrænse. Hvis der indsamles 100 kartofler, og ingen af dem er positive, så er det ligegyldigt, hvor mange delprøver, man opdeler dem i; man vil ikke kunne uddrive yderligere information angående den sande virusforeskomst (p), om man så testede hver eneste knold separat. Dette fremgår af tabellerne i det følgende, hvis man studerer sikkerhedsgrænsen for nul positive delprøver, som er ens i tabel 1-3 henholdsvis tabel 4-6. Det fremgår også af tabellerne at jo flere kartofler der indgår i prøven jo mindre bliver den øvre grænseværdi, herunder også hvis der ingen positive prøver er.

Tabel 1. Sikkerhedsintervaller: 100 knolde opdelt i 20 prøvebrønde

Antal positive delprøver	Minimum virusforekomst (%)	Maksimum virusforekomst (%)
0	0	3.83
1	0.03	5.65
2	0.23	7.33
3	0.63	8.99
4	1.17	10.68
5	1.82	12.40
6	2.55	14.20
7	3.37	16.09
8	4.27	18.09
9	5.26	20.23
10	6.34	22.53
11	7.51	25.04
12	8.79	27.81
13	10.20	30.91
14	11.76	34.45
15	13.50	38.59
16	15.48	43.61
17	17.78	50.05
18	20.54	59.15
19	24.06	74.78
20	29.22	100

Tabel 2. Sikkerhedsintervaller: 100 knolde opdelt i 10 prøvebrønde

Antal positive delprøver	Minimum virusforekomst (%)	Maksimum virusforekomst (%)
0	0	3.83
1	0.03	5.84
2	0.24	7.86
3	0.67	10.06
4	1.27	12.55
5	2.04	15.48
6	2.99	19.13
7	4.17	24.00
8	5.65	31.37
9	7.63	46.18
10	10.67	100

Tabel 3. Sikkerhedsintervaller: 100 knolde opdelt i 5 prøvebrønde

Antal positive delprøver	Minimum virusforekomst (%)	Maksimum virusforekomst (%)
0	0	3.83
1	0.03	6.29
2	0.25	9.38
3	0.75	14.07
4	1.58	24.06
5	3.01	100

Tabel 4. Sikkerhedsintervaller: 200 knolde opdelt i 20 prøvebrønde

Antal positive delprøver	Minimum virusforekomst (%)	Maksimum virusforekomst (%)
0	0	1.93
1	0.02	2.86
2	0.12	3.73
3	0.32	4.60
4	0.59	5.49
5	0.91	6.40
6	1.29	7.37
7	1.70	8.40
8	2.16	9.49
9	2.67	10.68
10	3.22	11.98
11	3.83	13.42
12	4.50	15.03
13	5.24	16.88
14	6.06	19.03
15	7.00	21.63
16	8.07	24.90
17	9.32	29.32
18	10.86	36.08
19	12.86	49.78
20	15.87	100

Tabel 5. Sikkerhedsintervaller: 200 knolde opdelt i 10 prøvebrønde

Antal positive delprøver	Minimum virusforekomst (%)	Maksimum virusforekomst (%)
0	0	1.93
1	0.02	2.96
2	0.12	4.01
3	0.34	5.16
4	0.64	6.48
5	1.03	8.06
6	1.51	10.07
7	2.11	12.82
8	2.87	17.16
9	3.89	26.64
10	5.49	100

Tabel 6. Sikkerhedsintervaller: 200 knolde opdelt i 5 prøvebrønde

Antal positive delprøver	Minimum virusforekomst (%)	Maksimum virusforekomst (%)
0	0	1.93
1	0.02	3.19
2	0.13	4.80
3	0.38	7.30
4	0.80	12.85
5	1.52	100