Research Repository UCD

Title	Frequentist and Bayesian approaches to prevalence estimation using examples from Johne's disease
Authors(s)	Messam, Locksley L. McV., Branscum, Adam J., Collins, Michael T., Gardner, Ian A.
Publication date	2008-01-01
Publication information	Messam, Locksley L. McV., Adam J. Branscum, Michael T. Collins, and Ian A. Gardner. "Frequentist and Bayesian Approaches to Prevalence Estimation Using Examples from Johne's Disease" 9, no. 1 (January 1, 2008).
Publisher	Cambridge University Press
Item record/more information	http://hdl.handle.net/10197/10175
Publisher's statement	This article has been published in a revised form in Animal Health Research Reviews [http://doi.org/10.1017/S1466252307001314]. This version is free to view and download for private research and study only. Not for re-distribution, re-sale or use in derivative works. © Cambridge University Press.
Publisher's version (DOI)	10.1017/S1466252307001314

Downloaded 2023-04-07T10:52:29Z

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa)

© Some rights reserved. For more information

Appendix A

1. For $\hat{P}_A = x/n$ ($x = 1, 2, \dots, n-1$), an exact $100(1-\alpha)\%$ confidence interval may be calculated by solving the following equations:

$$\sum_{k=x}^n \binom{n}{k} P_A^k (1 - P_A)^{n-k} = \alpha/2 \quad (1)$$

and

$$\sum_{k=0}^x \binom{n}{k} P_A^k (1 - P_A)^{n-k} = \alpha/2 \quad (2)$$

to get upper and lower confidence bounds respectively. For $x = 0$ the lower bound is 0 and for $x = n$, the upper bound is 1. An Excel spreadsheet for calculation of exact $100(1-\alpha)\%$ confidence intervals is available from the authors on request.

2. The Pooled Prevalence Calculator (PPC) can be used to calculate prevalence estimates and CIs (asymptotic and exact) for the following scenarios:

- Individual-level prevalence using a fixed pool size and assuming a perfect test.
- Individual-level prevalence using a fixed pool size and known sensitivity (Se) and specificity (Sp).
- Individual-level prevalence using a fixed pool size and uncertain sensitivity (Se) and specificity (Sp)¹.

The following inputs are required:

- Pool size (s_p)
- Number of pools (n_p)
- Number of positive pools (x_p)
- Test sensitivity (Se)*
- Test specificity (Sp)*
- Sample size used (in original study) to estimate sensitivity (n_1)**
- Sample size used (in original study) to estimate specificity (n_2)**
- Lower and upper confidence levels

The output provides:

- Individual-level prevalence estimate
- Confidence limits (exact or asymptotic as specified by user)

*When known test sensitivity and specificity are used

**When test sensitivity and specificity are uncertain.

¹Only asymptotic CIs available

3. McNemar's exact test

$$\text{p-value} = 2 \times \sum_{k=0}^{n_{12}} \binom{n_{12} + n_{21}}{k} \left(\frac{1}{2}\right)^{n_{12} + n_{21}} \quad \text{if } n_{12} < \frac{1}{2}(n_{12} + n_{21})$$

$$\text{p-value} = 2 \times \sum_{k=n_{12}}^{n_{12} + n_{21}} \binom{n_{12} + n_{21}}{k} \left(\frac{1}{2}\right)^{n_{12} + n_{21}} \quad \text{if } n_{12} > \frac{1}{2}(n_{12} + n_{21})$$

$$\text{p-value} = 1 \text{ if } n_{12} = \frac{1}{2}(n_{12} + n_{21})$$

4.

Table 1. 2×2 Contingency table for prevalence data (matched pairs)

	2 nd Survey		
Initial survey	T+	T-	Total
T+	$\dot{p}_{11}$	$\dot{p}_{12}$	$\dot{p}_{1+} = 0.05$
T-	$\dot{p}_{21}$	$\dot{p}_{22}$	$\dot{p}_{2+} = 0.95$
Total	$\dot{p}_{+1} = 0.10$	$\dot{p}_{+2} = 0.90$	1

Given $\dot{p}_{11} + \dot{p}_{22} = 0.87$, $\dot{p}_{11} + \dot{p}_{12} = 0.05$, $\dot{p}_{11} + \dot{p}_{21} = 0.10$, $\dot{p}_{21} + \dot{p}_{22} = 0.95$,

$\dot{p}_{12} + \dot{p}_{22} = 0.90$, then $\dot{p}_{21} - \dot{p}_{12} = 0.05$ and $\dot{p}_{21} + \dot{p}_{12} = 0.13$.

5. Derivation of (39)

Let

C_I = the total cost of obtaining and testing n individual samples for a prevalence survey.

C_p = the total cost of obtaining and testing n_p pooled samples for a prevalence survey.

c_T = cost of performing a single diagnostic test T .

c_s = cost of obtaining a single sample s .

s_p = pool size

c_p = cost of pooling s_p samples into a pool (where s_p = pool size).

For pooled estimation to be more economic:

$$C_p < C_I \text{ i.e.,}$$

$$n_p (c_T + c_p + s_p c_s) < n (c_T + c_s) \text{ and hence}$$

$$c_p < \left(\left(\frac{n}{n_p} \right) - 1 \right) c_T + \left(\left(\frac{n}{n_p} \right) - s_p \right) c_s$$

6. Sample sizes necessary to substantiate freedom from disease (i.e., prevalence below a prescribed value) may be calculated using the “Freedom from Disease” drop-down menu within the “Survey Toolbox” or “FreeCalc”. Both are available at www.epiweb.massey.ac.nz. The following steps are necessary:
- a Define the minimum expected prevalence at the herd level.
 - b Define the acceptable level for a Type I error (Concluding that the prevalence is not above the minimum expected herd prevalence when it is so).
 - c Define the acceptable level for a Type II error (Concluding that the prevalence is above the minimum expected herd prevalence when it is not).
 - d Calculate the no. of herds to be sampled using FreeCalc (Cameron and Baldock, 1998).
To do the calculation, one has to first choose a HSe and HSp which will allow you to determine the number of herds necessary to substantiate freedom from disease. This is done by trial and error. (The program also calculates the maximum number of positive herds to be expected if the population is to be considered disease free.)
 - e Define the minimum expected prevalence within herds.
 - f Define the Type I error = 1-HSe (Concluding that the individual-level prevalence is not above the minimum expected prevalence when it is in fact so).
 - g Define the Type II error rate = 1-HSp (Concluding that the individual-level prevalence is above the minimum expected herd prevalence when it is not so).

- h Calculate the no. of animals to be sampled from each herd inputting, each time, the respective herd's size. The program provides the maximum allowable number of positive animals if prevalence below the given level is to be substantiated.

Accepted for publication (AHRR) - uncorrected version

Appendix B - WinBUGS code

1. Apparent prevalence

```

a)
Model
{
  y ~ dbin(ap,n)           # Likelihood function for apparent prevalence
  ap ~ dbeta (1, 1)        # Prior distribution for apparent prevalence reflecting a lack of prior knowledge of prevalence
}
list(y=20, n =120)        # Data

b)
Model
{
  y ~ dbin(ap,n)           # Likelihood function for apparent prevalence
  ap ~ dbeta (5.62, 42.57) # Prior distribution for apparent prevalence prevalence (Mode =0.10, 95% certain apparent prevalence < 0.2)
}
list(y=20, n =120)        # Data

```

2. Low prevalence (Rogan-Gladen-estimator < 0)

```

Model
{
  y ~ dbin(ap,n)           # Likelihood function for apparent prevalence
  ap <- SeEL*tp+(1-SpEL)*(1-tp) # Model expressing relationship between apparent and true prevalence
  z ~ dbern(0.9)           # Likelihood function expressing herd infection probability
  SeEL ~ dbeta(6.28, 13.32) # Prior distribution for ELISA sensitivity. Mode =0.3, 95% certain SeEL <0.5
  SpEL ~ dbeta(384.13, 16.96) # Prior distribution for ELISA specificity. Mode =0.96, 95% certain SpEL >0.94
  tpa ~ dbeta (1, 1)       # Prior distribution (reflecting lack of knowledge) of true prevalence
  tp <- z*tpa              # Relationship between true prevalence and herd infection probability
  a <- equals(tp,0)        # Function calculating probability of herd infection=0
}
list(y=2, n =60)          # Data

```

3. Apparent prevalence = 0.

```

Model
{
  y ~ dbin(ap,n)           # Likelihood function for apparent prevalence
  ap <- SeEL*tp+(1-SpEL)*(1-tp) # Model expressing relationship between apparent and true prevalence
  z ~ dbern(0.9)           # Likelihood function expressing herd infection probability
  SeEL ~ dbeta(6.28, 13.32) # Prior distribution for ELISA sensitivity. Mode = 0.3, 95% certain SeEL < 0.5
  SpEL ~ dbeta(384.13, 16.96) # Prior distribution for ELISA specificity. Mode =0.96, 95% certain SpEL >0.94
  tpa ~ dbeta (1, 1)       # Prior distribution of true prevalence
  tp <- z*tpa              # Relationship between true prevalence and herd infection probability
  a <- equals(tp,0)        # Function calculating posterior probability of herd infection
}
list(y=0, n =60)          #Data

```

4. Hypergeometric sampling using BDFree

The following inputs are necessary:

1. Prior for Se: $Se \sim \text{beta}(6.28, 13.32)$
2. Prior for Sp: $Sp \sim \text{beta}(384.13, 16.96)$
3. Prior for P_T : $P_T \sim \text{beta}(1,1)$
4. Herd size: $N = 675$
5. Number of animals sampled: $n = 675$
6. Number of animals testing positive (number of reactors): $P_A = 23$
7. User specified value x , where $0 < x < 1$ used to calculate $\Pr(P_T < x)$. For this example $x = 0.00000001 \approx 0$. Used to estimate the approximate posterior probability of infection.

5. Pooled prevalence

```

Model
{
  y ~ dbin(ap, n)                                # Likelihood function
  ap <- Sep-(Spp+Sep-1)*(pow((1-tp),k))          # Function expressing relationship between prevalence of positive pools and true
                                                    prevalence
  Sep ~ dbeta(7.04, 5.03)                        # Prior distribution for pooled fecal culture sensitivity(Sep). Mode=0.60, 95% sure Sep > 0.35
  Spp ~ dbeta(919.87, 1.92)                      # Prior distribution for pooled fecal culture specificity(Spp). Mode=0.999, 95% sure Spp >0.995
  tpstar ~ dbeta(5.87,33.61)                    # Prior distribution for true prevalence given infection ( i.e. true prevalence >0). Mode=0.13, 95%
                                                    sure tpstar <0.25

  Z ~ dbern(0.95)                                # Prior distribution for probability of infection
  tp <- Z*tpstar                                # Function expressing relationship between probability of infection and true prevalence

  a <- equals(tp,0)                             # Probability that true prevalence = 0
}

list(y=2, n=6, k=0.1)                          # Data

```


6. Herd prevalence and prevalence distribution estimation

```

Model
{
  for(i in 1:k)                                # Command designating number of loops to be performed
  {
    y[i] ~ dbin(ap[i], n)                        # Likelihood function for the ith herd
    ap[i] <- tp[i]*SeEL + (1-tp[i])*(1-SpEL)      # Function showing relationship between apparent and true prevalences for ith herd
    tp[i] <- tpstar[i]*pinf[i]                  # Function showing relationship between true prevalence and infection probability for ith herd

    pinf[i] ~ dbern(tau)                        # Prior distribution of probability of infection for ith. herd
    tpstar[i] ~ dbeta(alpha,beta)               # Prior distribution of true prevalence for ith herd given true prevalence >0
  }

  SeEL ~ dbeta(6.28, 13.32)                    # Prior distribution of ELISA sensitivity. Mode=0.3, 95% sure SeEL < 0.5
  SpEL ~ dbeta(384.13, 16.96)                  # Prior distribution of ELISA specificity. Mode=0.96, 95% sure SpEL > 0.94

  tau ~ dbeta(4.8, 3.6)                        # Prior distribution for herd probability of infection. Mode=0.60, 95% sure > 0.3

  alpha <- mu*psi
  beta <- psi*(1-mu)                           #Relationship of distribution parameters to mean (mu) and psi (variance related parameter)

  mu ~ dbeta(3.283, 17.744)                    # Prior distribution for mean (mu) of prevalence distribution Mode=0.12, 95% sure < 0.30.
  psi ~ dgamma(4.524, 0.387)                  # Prior distribution for psi (variability related parameter) of prevalence distribution. Uses median
                                              # of 95th percentile of prevalence. Distribution=0.30 and 99% sure this number is < 0.50

  H ~ dbern(tau)                               # Prior distribution of probability of infection for randomly chosen herd
  tpstarH ~ dbeta(alpha,beta)                  # Prior distribution of true prevalence for randomly chosen herd given true prevalence >0
  tpH <- H*tpstarH                             # Function showing relationship between true prevalence and infection probability for ith herd

  a1 <- 1-step(tpH - 0.05)                     # Probability that a randomly chosen herd has true prevalence < 0.05
  a2 <- 1-step(tpH-0.2)                       # Probability that a randomly chosen herd has true prevalence < 0.20

}
list(k=29, n=60, y=c(2,0,3,2,3,6,0,9,3,13,3,4,1,7,2,2,0,6,2,2,8,1,5,0,5,4,2,0,15)) #Data

```

7. Prevalence differences based on the same test

```

Model
{
  y1 ~ dbin(ap1, n1)                # Likelihood function for apparent prevalence
  y2 ~ dbin(ap2, n2)

  ap1 <- SeEL*tp1+(1-SpEL)*(1-tp1)  # Model expressing relationship between apparent and true prevalence (both herds)
  ap2 <- SeEL*tp2+(1-SpEL)*(1-tp2)
  SeEL ~ dbeta(6.28, 13.32)         # Prior distribution for ELISA sensitivity. Mode =0.3, 95% certain SeEL <0.5
  SpEL ~ dbeta(384.13, 16.96)       # Prior distribution for ELISA specificity. Mode =0.96, 95% certain SpEL >0.94

  tp1 ~ dbeta (1,1)                 # Prior distribution (reflecting lack of knowledge) of true prevalence
  tp2 ~ dbeta (1,1)

  diff<- tp2-tp1                    # Function calculating difference between prevalences
  a<- step(diff)                    # Function determining if prevalence difference >0
}
list(y1=26, y2 = 30, n1=200, n2=200) # Data

```

8. Prevalence differences based on different tests

```

Model
{
  y1 ~ dbin(ap1, n1)                # Likelihood function for apparent prevalence
  y2 ~ dbin(ap2, n2)

  ap1 <- SeEL*tp1+(1-SpEL)*(1-tp1)  # Model expressing relationship between apparent and true prevalence(both herds)
  ap2 <- SeFC*tp2+(1-SpFC)*(1-tp2)

  SeEL ~ dbeta(6.28, 13.32)         # Prior distribution for ELISA sensitivity. Mode =0.3, 95% certain SeEL <0.5
  SpEL ~ dbeta(384.13, 16.96)       # Prior distribution for ELISA specificity. Mode =0.96, 95% certain SeEL >0.94
  SeFC ~ dbeta(7.04, 5.03)          # Prior distribution for fecal culture sensitivity. Mode =0.60, 95% certain SeFC >0.35
  SpFC ~ dbeta(919.87, 1.92)        # Prior distribution for fecal culture specificity. Mode =0.999, 95% certain SeFC >0.995

  tp1 ~ dbeta (1,1)                 # Prior distribution (reflecting ignorance) of true prevalence
  tp2 ~ dbeta (1,1)

  diff<- tp2-tp1                    #Function calculating difference between prevalences
  a<- step(diff)                    # Function determining if prevalence difference >0
}
list(y1=26, y2 = 30, n1=200, n2=200) # Data

```

Tables

Table 1. Priors for sensitivity (Se) and specificity (Sp) of IDEXX serum ELISA and HEY fecal culture based on validation studies and updated with expert opinion

Test Parameter	Expert opinion ¹	Beta prior
Se (ELISA)	Most likely $Se = 0.3$, 95 % certain $Se < 0.5$	(6.28, 13.32)
Sp (ELISA)	Most likely $Sp = 0.96$, 95 % certain $Sp > 0.94$	(384.13, 16.96)
Se (fecal culture)	Most likely $Se = 0.60$, 95 % certain $Se > 0.35$	(7.04, 5.03)
Sp (fecal culture)	Most likely $Sp = 0.999$, 95 % certain $Sp > 0.995$	(919.87, 1.92)

¹Most likely sensitivity and specificity are used as point estimates for frequentist calculations

Table 2. Results of herd prevalence survey for Johne's disease using IDEXX ELISA for 29 central California dairies (60 cows per herd).

No. herds	No. positive (%)
5	0 (0.0)
2	1 (1.7)
7	2 (3.3)
4	3 (5.0)
2	4 (6.7)
2	5 (8.3)
2	6 (10.0)
2	7 (11.6)
1	8 (13.3)
1	13 (21.7)
1	15 (25.0)

Table 3. 2×2 Contingency table for prevalence data

	Sample 1	Sample 2	Total
T+	$n_{11} = O_{11} = 26$	$n_{12} = O_{12} = 30$	$n_{1+} = r_1 = 56$
T-	$n_{21} = O_{21} = 174$	$n_{22} = O_{22} = 170$	$n_{2+} = r_2 = 344$
Total	$n_{+1} = c_1 = 200$	$n_{+2} = c_2 = 200$	$n = 400$

Accepted for publication (AHRR) - uncorrected version

Table 4. 2×2 Contingency table for prevalence data (matched pairs)

	2 nd survey		
Initial survey	T+	T-	Total
T+	$n_{11} = 4$	$n_{12} = 22$	$n_{1+} = 26$
T-	$n_{21} = 26$	$n_{22} = 148$	$n_{2+} = 174$
Total	$n_{+1} = 30$	$n_{+2} = 170$	$n_p = 200$

n_p = the total number of pairs

Accepted for publication (AHRR) - uncorrected version

Table 5. Sample sizes to estimate hypothesized true prevalences ($\dot{P}_T$) with a given precision (Δ) based on *a priori* apparent prevalence estimates ($\dot{P}_A$) for Serum ELISA and fecal culture assuming known sensitivity (Se) and specificity (Sp).

	Serum ELISA (Se = 0.3, Sp = 0.96)						Fecal culture (Se = 0.6, Sp = 0.999)					
$\dot{P}_A$	0.100	0.120	0.140	0.160	0.180	0.200	0.100	0.120	0.140	0.160	0.180	0.200
$\dot{P}_T$	0.231	0.308	0.385	0.462	0.538	0.615	0.165	0.199	0.232	0.265	0.299	0.332
Δ												
0.10	512	601	685	764	839	910	96	114	129	144	159	172
0.15	228	267	305	340	373	405	43	51	58	64	71	77
0.20	128	151	172	191	210	228	NA	NA	33	36	40	43

NA = Not applicable

Table 6. Sample sizes (numbers in each group) required to detect differences in true prevalence (P_T) for independent samples ($\alpha = 0.05$, $1-\beta = 0.80$ and $k=1$ ($n_2 = n_1$)).

Serum ELISA (Se = 0.3, Sp = 0.96)							Fecal culture (Se = 0.6, Sp = 0.999)					
P_T		0.001	0.1	0.2	0.3	0.4		0.001	0.1	0.2	0.3	0.4
	P_A	0.040	0.066	0.092	0.118	0.144	P_A	0.0016	0.061	0.121	0.181	0.241
0.001	0.040	NA	1190	361	188	121	0.0016	NA	134	63	40	29
0.1	0.066		NA	1687	484	242	0.061		NA	360	115	61
0.2	0.092			NA	2179	603	0.121			NA	559	161
0.3	0.118				NA	2640	0.181				NA	726
0.4	0.144					NA	0.241					NA

NA = Not applicable

Table 7. Samples sizes¹ required to estimate hypothesized true prevalences ($\dot{P}_T$) with a given precision (Δ) based on an *a priori* apparent prevalence estimate ($\dot{P}_A$) using pooled fecal culture with known sensitivity ($PSe = 0.6$) and specificity ($PSp = 0.999$).

Pool size	5				10			
$\dot{P}_A$	0.02	0.04	0.06	0.08	0.02	0.04	0.06	0.08
$\dot{P}_T$	0.034	0.070	0.110	0.154	0.036	0.078	0.136	0.249
Δ								
0.02	492 (99)	900 (180)	1258 (252)	1571 (315)	431 (44)	690 (69)	866 (87)	1102 (111)
0.03	219 (44)	400 (80)	560 (112)	699 (140)	192 (20)	307 (31)	385 (39)	490 (49)
0.04	NA	225 (45)	315 (63)	393 (79)	NA	173 (18)	217 (22)	276 (28)

¹Numbers of pools in parentheses

NA = Not applicable

Table 8. Sample sizes required to estimate a hypothesized true prevalence ($\hat{P}_r$) with precision (Δ) for individual (I) and pooled testing (P) using fecal culture with known sensitivity ($Se = 0.60$) and specificity ($Sp = 0.999$) (Pool size = 5).

True prevalence												
	0.10		0.12		0.14		0.16		0.18		0.20	
Δ	I	P	I	P	I	P	I	P	I	P	I	P
0.05	245	189	290	214	333	237	375	258	416	277	455	295
0.10	62	48	73	54	84	60	94	65	104	70	114	74
0.15	NA	NA	NA	NA	NA	NA	42	29	47	31	51	33

NA = Not applicable

Table 9. Sample sizes required to estimate a hypothesized true prevalence ($\dot{P}_T$) with precision (Δ) for individual (I) and pooled testing (P) using fecal culture with known sensitivity ($Se = 0.60$) and specificity ($Sp = 0.999$) (Pool size = 10).

True prevalence												
	0.10		0.12		0.14		0.16		0.18		0.20	
Δ	I	P	I	P	I	P	I	P	I	P	I	P
0.05	245	124	290	134	333	140	375	148	416	154	455	160
0.10	62	32	73	34	84	36	94	38	104	40	114	40
0.15	NA	NA	NA	NA	NA	NA	42	17	47	17	51	18

NA = Not applicable

Table 10. Number of herds required to estimate a hypothesized true herd prevalence ($HP_T = 0.7$) with precision ($\Delta = 0.1$) and $z_{\alpha/2} = 1.96$ for different combinations of herd sensitivity (HSe) and herd specificity (HSp).

<i>HSe</i>	<i>HSp</i>						
	0.5	0.6	0.7	0.8	0.9	0.95	0.99
0.5	NA	9570	2367	1033	566	440	366
0.6	9416	2386	1067	600	381	313	270
0.7	2213	1016	585	381	267	228	202
0.8	879	523	350	252	190	167	152
0.9	413	289	216	168	135	122	113
0.95	287	215	169	137	113	104	97
0.99	212	168	138	115	98	90	85

NA = Not applicable

Table 11. Number of herds required to estimate a hypothesized true herd prevalence ($HP_T = 0.9$) with precision ($\Delta = 0.1$) and $z_{\alpha/2} = 1.96$ for different combinations of herd sensitivity (HSe) and herd specificity (HSp).

<i>HSe</i>	<i>HSp</i>						
	0.5	0.6	0.7	0.8	0.9	0.95	0.99
0.5	NA	9601	2398	1064	597	471	397
0.6	9293	2340	1047	592	381	315	275
0.7	2090	944	539	350	246	211	188
0.8	756	438	289	206	155	137	124
0.9	290	196	144	111	89	81	75
0.95	164	120	93	75	63	58	54
0.99	89	71	59	50	44	41	39

NA = Not applicable

Table 12. Sample sizes and upper limits of number of reactors necessary to achieve herd sensitivity and specificity of 0.95 and 0.98, respectively, when using a test with $Se = 0.60$ and $Sp = 0.999$ for different herd sizes.

Herd size	Required sample size	Limiting number of reactors
300	296	2
400	351	2
500	382	2
1000	436	2
1500	453	2

Table 13. Number of cows necessary to sample in a herd of size 1000 to establish with 95% confidence that *Mycobacterium avium* subspecies *paratuberculosis* (MAP) infection, if present is less than a given true prevalence (P_T).

$P_T (MAP) <$	Fecal culture (Se = 0.60, Sp = 0.999)		Serum ELISA (Se = 0.30, Sp = 0.96)	
	Sample size	Maximum number testing positive	Sample size	Maximum number testing positive
0.02	336	1	>1000	NA
0.04	181	1	>1000	NA
0.06	123	1	>1000	NA
0.08	94	1	>1000	NA
0.10	48	0	759	39
0.12	40	0	562	30
0.14	34	0	434	24
0.16	30	0	350	20
0.18	26	0	288	17
0.20	24	0	233	14

NA= not applicable (herd size too small).