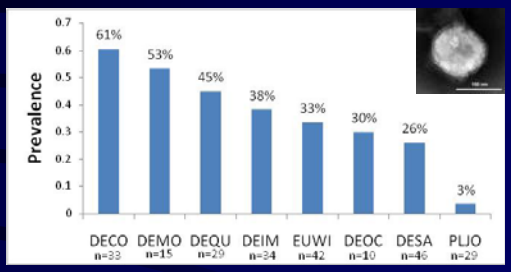
PATHOGEN SURVEILLANCE:
Designs and Analyses



Matthew J. Gray

University of Tennessee
Center for Wildlife Health
Department of Forestry, Wildlife and Fisheries

Lecture Outline

I. Uses of Surveillance Data

II. Statistical Inference

III. Sample Size and Sample Design

IV. Confidence Intervals and Tests

Goal of Pathogen Surveillance

To obtain an unbiased estimate of pathogen or disease prevalence in a population

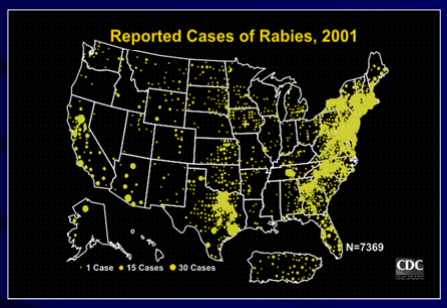
Pathogen Prevalence

An estimate of the proportion of individuals in a population that are infected with a pathogen

Infection  Disease

Uses of Surveillance Data

Occurrence and Distribution

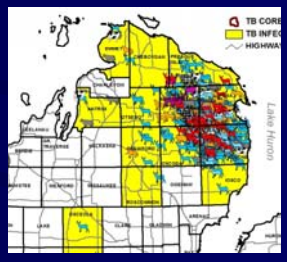


Evidence of Emergence

Pathogen or disease that is increasing in distribution, prevalence, or host range

Uses of Surveillance Data

Evidence of Hotspots



Identification of Mechanisms of Emergence

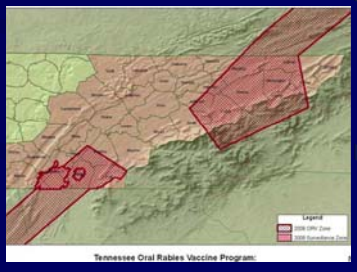
Uses of Surveillance Data

Disease Intervention Strategies

Interrupt Host-Pathogen Cycle

Reduce Stressors

1. High transmission
2. Distribution expansion
3. Stressors

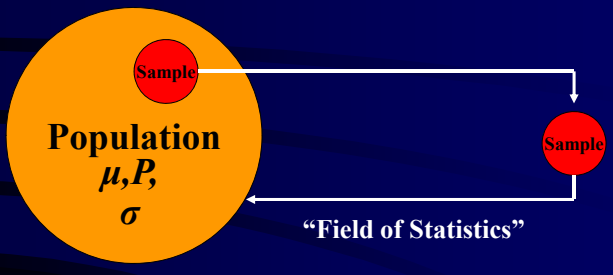


Individual vs. Population

What conclusions can be made?

Uses and Benefits of Individual vs. Population Data?

Statistical Inference on Populations



Subset of all individuals

$\bar{x}$ $\hat{p}$ S

"Point Estimates"

"Parameters"

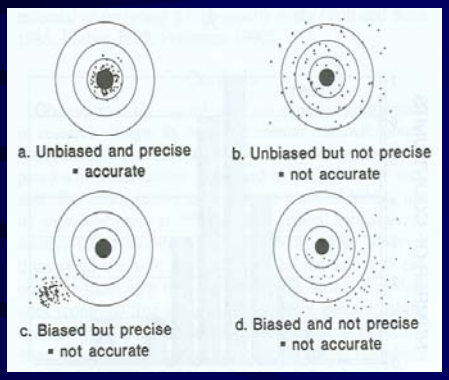
Is statistics necessary for reports on individual cases?

Measures of Reliability

How variable is your estimate?

S - - -

$\hat{p}$ - - -



How close is your estimate a the true prevalence?

- - - → σ

- - - → P

- **Precision:** numerical closeness of measurements to each other
- **Bias:** numerical closeness of measurements to a true population parameter (P)
- **Accuracy:** unbiased + precision

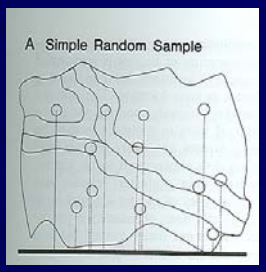
Surveillance Designs

Collecting Unbiased, Representative Sample

Random Sampling

All individuals or surveillance locations have an equal probability of being sampled

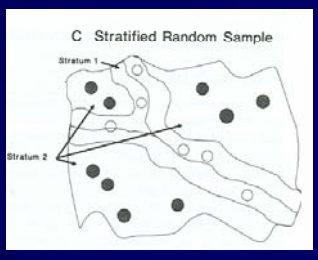
Random Numbers Table or Programs



Stratified Random Sampling

All individuals within a specified location or category have an equal probability of being sampled

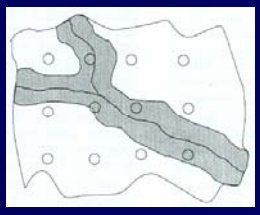
Habitat Type/Condition; Gender; Age Class



Surveillance Designs

Systematic Sampling

Individuals or locations in specified intervals have an equal probability of being sampled



Biased if Not a Uniform Distribution

Other Designs: Cluster sampling, Multi-stage sampling, Adaptive Sampling

Haphazard Sampling

Individuals are selected based on ease of access or in a way that does not follow an unbiased random process.

Case Studies: Inferences Limited to the Sample

Estimating Required Sample Size

Detect a Pathogen

Information Needed

- Assumed Pathogen Prevalence Level (APPL)
- Estimated Host Population Size
- Confidence in detection (95%)

Population Size	10% APPL	5% APPL	2% APPL
50	20	35	50
100	23	45	75
250	25	50	110
500	26	55	130
2000	27	60	145
>100,000	30	60	150

(Amos 1985, Thoesen 1994)

Estimating Required Sample Size

Precise Estimate of Prevalence

$Z_{\alpha/2}$ =1.96
(95% confidence)

$$n = p(1-p)\left[\frac{1.96}{d}\right]^2$$

p = Prevalence from a previous study

d = error in estimation

"Error in Estimation" is the amount of error you are willing to tolerate in your estimate of prevalence

Error = 5%	Error = 10%	Error = 10%
p = 85%	p = 85%	p = unknown
$n = (0.85)(0.15)\left[\frac{(1.96)}{0.05}\right]^2 \approx 196$	$n = (0.85)(0.15)\left[\frac{(1.96)}{0.10}\right]^2 \approx 49$	$n = (0.25)\left[\frac{(1.96)}{0.10}\right]^2 \approx 96$

What happens if estimation error increases?
What happens if prevalence is near 0.5?

$0.01 < P(1-p) < 0.25$

Estimating Prevalence

$$\hat{p}_i = \frac{n_i}{N_i}$$

$$\hat{p}_1 = \frac{4}{40} = 10\%$$

Estimate of Precision

$$S = \sqrt{\frac{\hat{p}_i \hat{q}_i}{n}} \quad where, \quad \hat{q}_i = 1 - \hat{p}_i$$

Standard Deviation, S: Expected Average Deviation in p-hat around P

For Large n, 95% Confidence Interval:

$$CI(95\%) = \hat{p}_i \pm 1.96(S)$$

Estimating Prevalence and CI

$$\hat{p}_i = \frac{n_i}{N_i} \qquad S = \sqrt{\frac{\hat{p}_i \hat{q}_i}{n}} \qquad CI(95\%) = \hat{p}_i \pm 1.96(S)$$

Infection Data:

Age Class	Infected	Sampled	P_Hat	Q_Hat	S	EM	Lower	Upper
Juv	9	35	0.2571	0.7429	0.0739	0.145	0.112	0.402
Subadult	10	40	0.25	0.75	0.0685	0.134	0.116	0.384
Adult_F	5	15	0.3333	0.6667	0.1217	0.239	0.095	0.572
Adult_M	3	30	0.1	0.9	0.0548	0.107	-0.007	0.207

$CI(Juv) = 0.112 < P < 0.402$ $CI(F) = 0.095 < P < 0.572$

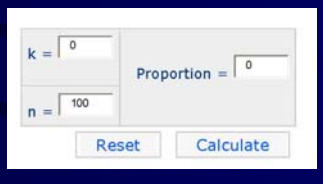
$CI(SA) = 0.116 < P < 0.384$ $CI(M) = 0 < P < 0.207$

Is Prevalence Different Among Age Classes?

Estimating Confidence Intervals

Small Sample Size or Prevalence = 0

Wilson Score Method



http://faculty.vassar.edu/lowry/prop1.html

Journal of the American Statistical Association 22:209-212

Hypothesis Testing

Two Proportions

Different if Z > 1.96

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\hat{p} = \frac{X + Y}{n_1 + n_2}$$

Age Class	Infected	Sampled	P_i	P_Hat	sq(P*Q)	SqRt	Den	Num	Z
Juv	9	35	0.2571	0.1846	0.388	0.249	0.097	0.157	1.6279
Adult_M	3	30	0.1						

$$\hat{p} = \left(\frac{9 + 3}{35 + 30}\right) = 0.185$$

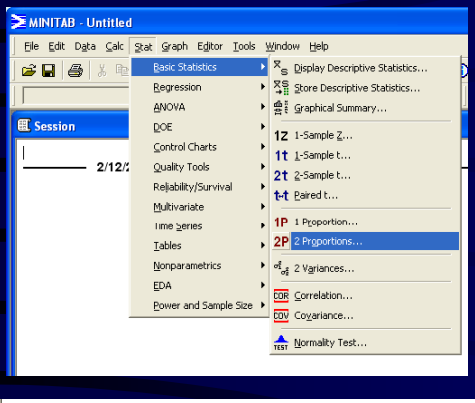
$$Z = \frac{(0.257 - 0.10)}{\left[\sqrt{0.185 \times 0.815}\sqrt{\frac{1}{35} + \frac{1}{30}}\right]} = 1.63$$

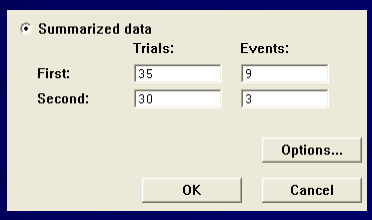
P = 0.104

Hypothesis Testing

Two Proportions

Minitab





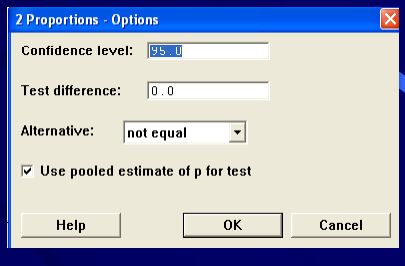

```
Test and CI for Two Proportions

Sample  X   N  Sample p
1       9  35  0.257143
2       3  30  0.100000


Difference = p (1) - p (2)
Estimate for difference:  0.157143
95% CI for difference:  (-0.0231070, 0.337393)
Test for difference = 0 (vs not = 0):  Z = 1.63  P-Value = 0.104


* NOTE * The normal approximation may be inaccurate for small samples.

Fisher's exact test: P-Value = 0.122
```



Hypothesis Testing

Multiple Proportions: One Hypothesis

Prevalence Different among 4 Age Classes?

Chi-square Test of Homogeneity SAS®

```
data one;
input age $ infect $ count;
cards;
juv yes 9
juv no 26
sub yes 10
sub no 30
AF yes 5
AF no 10
AM yes 3
AM no 27
;
proc freq;
weight count;
tables age*infect/nocol nopct chisq;
run;
```

age infect

Frequency Row Pct	no	yes	Total
AF	10 66.67	5 33.33	15
AM	27 90.00	3 10.00	30
juv	26 74.29	9 25.71	35
sub	30 75.00	10 25.00	40
Total	93	27	120

Statistics for Table of age by infect

Statistic	DF	Value	Prob
Chi-Square	3	4.0485	0.2563
Likelihood Ratio Chi-Square	3	4.4688	0.2151
Mantel-Haenszel Chi-Square	1	0.1010	0.7506

Hypothesis Testing

Multiple Proportions: Two Hypotheses

Prevalence Different among 2 Land Uses and 3 Seasons?

Logistic Regression SAS®

species=RACA

The LOGISTIC Procedure

Contrast Test Results

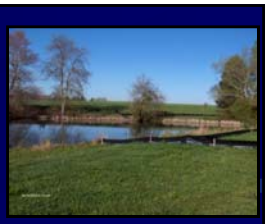
Contrast	DF	Wald Chi-Square	Pr > ChiSq
Overall--Are treatments different	1	0.0752	0.7839
June versus Oct	1	0.7997	0.3712
June versus Feb	1	14.2209	0.0002
Oct versus Feb	1	5.7014	0.0170
Overall--Are seasons different	2	15.9347	0.0003

Odds Ratio Estimates

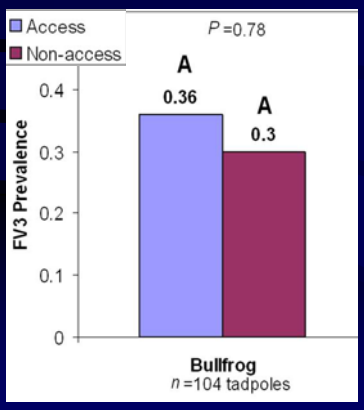
Effect			Point Estimate	95% Wald Confidence Limits	
trt	access	vs noaccess	1.137	0.454	2.847
season B vs A			1.832	0.486	6.907
season C vs A			7.709	2.667	22.283

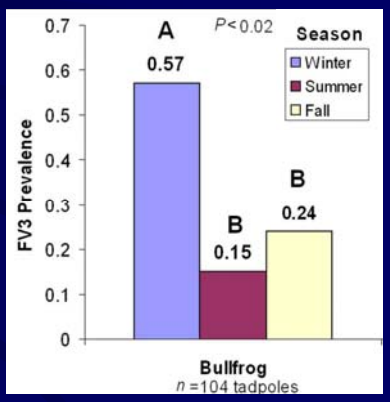
Results

Cattle Land Use and Season



7.7X More Likely!!





Bd Surveillance

Non-lethal Techniques: Brem et al. (2007)

Swabbing Preferred

Adults:

Swab 5 times in 5 locations

- Rear feet (webbing)
- Inner thighs
- Ventral Abdomen

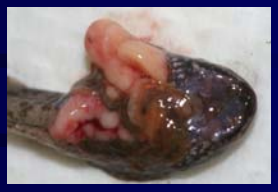
Larvae:

Swab Oral Cavity 5 times

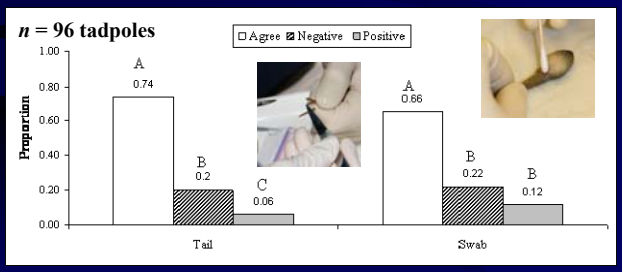
Store in 70% EtOH

Ranavirus Surveillance

Lethal Collection:
Liver Preferred
St-Armour & Lesbarrères (2007)



Non-lethal Techniques: Gray et al. (2012)



Lethal followed by Tail

Misclassification Decreases as Disease Progresses
Greer and Collins (2007)

Toe Clips
False negative = 7%
False positive = 3%
St-Armour & Lesbarrères (2007)