

Epistatistics: From epidemiologic measures to inference

Tomás Aragón, MD, DrPH

Center for Infectious Disease Preparedness
UC Berkeley School of Public Health
URL: <http://www.idready.org>
Email: aragon@berkeley.edu

November 11, 2006

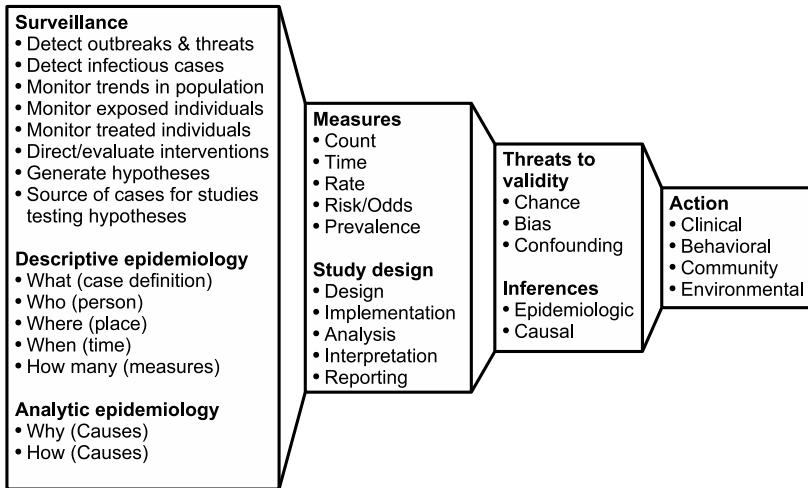


Outline

- 1 The epidemiologic approach and inference
- 2 From epidemiologic measures to inference
- 3 Practical examples from public health



The epidemiologic approach: Steps to public health action



Inference in epidemiology

- Types of inference
 - Statistical inference
 - Epidemiologic inference
 - Causal inference
- Threats to valid inferences
 - Chance (random error)
 - Bias (systematic error)
 - Confounding (alternative causal explanations)



From epidemiologic measures to inference

- 1 What is the point estimate? (estimation)
- 2 What is the variability of this estimate? (confidence interval)
- 3 Is this estimate consistent with a reference value? (p value and Type I error)
- 4 What is the chance of detecting a meaningful difference (effect size) from the reference value, if one exists? (power and Type II error)
- 5 How many subjects are required to be able to detect a meaningful difference, if one exists? (sample size)



1 Estimation

What is the point estimate?

- Measures of occurrence
 - Count of new or existing cases
 - Time until an event occurs
 - Rate
 - Risk
 - Prevalence
- Measures of association
 - Rate ratio
 - Risk ratio
 - Odds ratio



2 Confidence interval

- Question to ask:
 - What is the variability of this estimate?
 - What are the components of variability?
 - What is a confidence interval?
- Methods for constructing confidence intervals
 - Normal distribution approximation methods
($R_L, R_U = R \pm Z \times SE(R)$)
 - Exact methods using known distributions
 - Exact method approximations using derived formulas
 - Resampling (simulation) methods for unknown distributions



3. p values and Type I error (α)

Is this estimate consistent with a reference value?

- p value: Under the null hypothesis (estimate = reference), what is the probability of observing estimate (or more extreme value)?

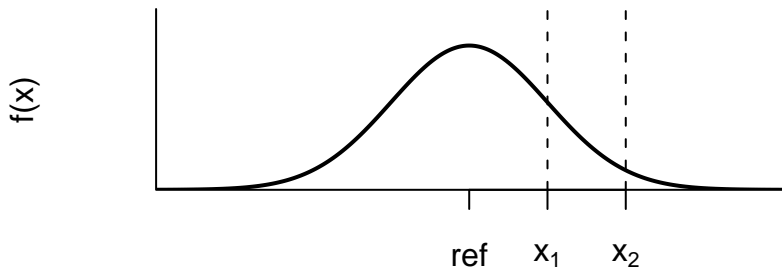
What is driving the p value?

- Type I error: At what p value are you willing to incorrectly infer the values are not consistent when they actually are (null hypothesis true)?

$$\begin{aligned}\alpha &= \text{Type I error} \\ &= P(\text{Rejecting null hypothesis} \mid \text{Null hypothesis true})\end{aligned}$$



Figure 1. Assessing whether some measure (x_1, x_2, \dots) is consistent with a reference value



4. Power and Type II error (β)

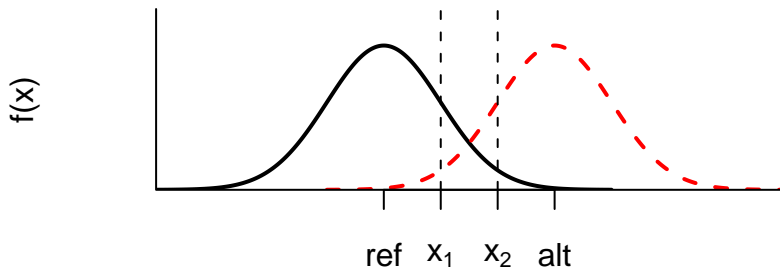
What is the chance of detecting a meaningful difference from the reference value, if one exists?

- What is a meaningful difference (effect size) worth detecting?
- What is the probability of detecting this effect size if it exists? (power)
- What is the probability of failing to reject the null hypothesis when the effect size actually exists? (β)

$$\begin{aligned}\text{power} &= 1 - \beta \\ &= P(\text{Rejecting null hypothesis} \mid \text{Effect size exists})\end{aligned}$$



Figure 2. Assessing statistical power to detect a meaningful difference



5 Sample size calculations

How many subjects are required to be able to detect a meaningful difference, if one exists?

Sample size calculation integrates all the previous concepts:

- Determine variability of estimate?
- Select an effect size?
- Select α (Type I error)
- Select β (Type II error); power = $1 - \beta$
- Calculate sample size



Review: From epidemiologic measures to inference

- 1 Measure estimate
- 2 Assess variability (confidence intervals)
- 3 Test consistency (p value and Type I error [α])
- 4 Detect differences (effect size, power, and Type II error [β])
- 5 Calculate sample sizes



Example 1: Estimation

Meningococcal disease:

Incidence rate:

$$r = \frac{x}{PT},$$

where x is count of incident cases, and PT is person-time at risk.

Consider meningococcal disease in the general population where the incidence rate is about 1 per 100,000 per year.

The population of San Francisco is about 800,000, and the expected number of cases annually is about 8.



Example 2: p values

For convenient, we assume the occurrence of new cases follows a Poisson distribution.

The Poisson distribution is a discrete probability distribution with the following *density function*:

$$P(X = x) = \frac{x^{-\lambda} \lambda^x}{x!},$$

where X is the random variable, x is the observed count, and λ is the expected count.

And, here is the Poisson *distribution function*:

$$P(X \leq x) = \sum_{k=0}^x \frac{k^{-\lambda} \lambda^k}{k!}.$$



Example 2: p values (continued)

Meningococcal disease in the general population about 1 per 100,000 per year. In 2005, 7 cases were reported in San Francisco (population = 795,292). Assume the occurrence of new cases follow a Poisson distribution

1. What's the chance of seeing 7 cases? [i.e., $P(X = 7)$]
2. What's the chance of seeing 7 or fewer cases? [i.e., $P(X \leq 7)$]
3. What's the chance of seeing 12 or more cases? [i.e., $P(X \geq 12)$]

```
> lam = 795292*(1/100000)           #expected count
> dpois(x = 7, lambda = lam)         #P(X=7))
[1] 0.1403933
> ppois(q = 7, lambda = lam)        #P(X<=7)
[1] 0.4595517
> 1 - ppois(q = 11, lambda = lam)   #P(X>=12)
[1] 0.1085553
```



Example 3: Confidence interval

Method 1. Normal approximation:

$$r_L, r_U = r \pm Z \times SE(r),$$

where $SE(r) = \sqrt{x/PT^2}$, and Z is the quantile value for the standard normal density function. For a 95% CI, $Z = 1.96$.

What are the limitations of normal approximation methods?



Example 3: Confidence interval (continued)

Method 1 (cont). Normal approximation:

In 2005, 7 cases of meningococcal disease were reported in San Francisco (population = 795,292). Using the normal approximation method, calculate the 95% CI.

```
> x = 7; PT = 795292; r = x/PT; Z = 1.96  
> r.LL = r-Z*sqrt(x/PT^2); r.UL = r+Z*sqrt(x/PT^2)  
> r*100000  
[1] 0.8801799  
> r.LL*100000  
[1] 0.2281335  
> r.UL*100000  
[1] 1.532226
```



Example 3: Confidence interval (continued)

Method 2. Exact method approximation using Byar's confidence limits:

$$r_L, r_U = (x + 0.5) \left(1 - \frac{1}{9(x + 0.5)} \pm \frac{Z}{3} \sqrt{\frac{1}{(x + 0.5)}} \right)^3 / PT$$

In 2005, 7 cases of meningococcal disease were reported in San Francisco (population = 795,292). Calculate the 95% CI using Byar's formula.

```
> library(epitools)
> pois.byar(x=7, pt=795292, conf.level=0.95)
  x      pt      rate      lower      upper conf.level
1 7 795292 8.802e-06 3.566e-06 1.651e-05          0.95
```



Example 3: Confidence interval (continued)

Method 3. Exact method using Poisson distribution:

In 2005, 7 cases of meningococcal disease were reported in San Francisco (population = 795,292). Calculate the 95% CI using an exact method.

```
> library(epitools)
> pois.exact(x=7, pt=795292, conf.level=0.95)
  x      pt      rate      lower      upper conf.level
1 7 795292 8.8018e-06 3.5388e-06 1.8135e-05      0.95
```



Summary: From epidemiologic measures to inference

- 1 What is the point estimate? (estimation)
- 2 What is the variability of this estimate? (confidence interval)
- 3 Is this estimate consistent with a reference value? (p value and Type I error)
- 4 What is the chance of detecting a meaningful difference (effect size) from the reference value, if one exists? (power and Type II error)
- 5 How many subjects are required to be able to detect a meaningful difference, if one exists? (sample size)

