What are the proportion of people who are infected with HIV?

How can blood donations be efficiently screened for diseases?

What is the probability of transmission of a pathogen from an insect vector to a plant?

What chemical compounds could be potentially useful in a new drug to cure a disease?

Other terms:
- Pooled testing
- Pooled experiments
- Special case of composite sampling

What is group testing?

- Introduced to area in 2002
- Work with Josh Tebbs
  - Worked with at OSU 2001-3
  - One of 9 North Carolina State University PhD graduates in this area
  - Assistant Professor in the Department of Statistics at the University of South Carolina
- Just Group It! World Tour
  - Introduction to group testing
  - Stops
    - Lincoln and Omaha
    - Oslo, Norway

Testing an item for a binary trait
- Suppose people are being tested for a disease
- What is the prevalence of the disease?
- Who has the disease?

Individual testing
- Problem: Cost
- Problem: Time
What is group testing?

- Group testing
  - If the GROUP sample is negative, then all 4 people do not have the disease
  - If the GROUP sample is positive, then at least ONE of the 4 people have the disease
  - “Retesting” can be done to determine which people are positive
  - Cost and time savings!
  - Strategy works well when prevalence of the trait is small
  - If prevalence is large, all groups may test positive

Basic statistics

- Group testing research is split into two areas
  - Statistical
  - Combinatorial
- Combinatorial group testing research – see Du and Hwang (2000)
  - Deterministic model for the identification of positive items
  - Try to minimize the number of retests to find the positive items in a group
  - Upper bound for the number of positive items often needs to be assumed.
- Statistical group testing research
  - Each item’s binary response is treated as a random variable
  - Probability distributions used then to help determine:
    - Prevalence of a trait in a population (Estimation problem)
    - Which items are positive (Identification problem)

What is group testing?

- Purpose
  - Basic statistics ideas
  - Show examples of where group testing is used

Basic statistics

- Notation
  - Individual responses
    - \( Y_{ik} = 1 \) if the \( i^{th} \) item in the \( k^{th} \) group has the trait (positive)
    - \( Y_{ik} = 0 \) otherwise (negative) for \( i = 1, \ldots, I_k \) and \( k = 1, \ldots, K \)
  - \( Y_{ik} \) are i.i.d. Bernoulli(\( p \)) random variables
    - \( p = P(Y_{ik} = 1) \)
    - \( p \) can be thought of as the “individual probability” or “prevalence in a population”
  - Assume equal group sizes, \( I_1 = I_2 = \ldots = I_K = I \)
Basic statistics

- Notation (continued)
  - Group responses
    - $Z_k = 1$ denotes a positive response
    - $Z_k = 0$ denotes a negative response for the $k^{th}$ group
    - $Z_k$ are i.i.d. Bernoulli($\theta$) random variables
      - $\theta = P(Z_k = 1)$
  - Individual and group relationship
    - $Z_k = 1$ if and only if $\sum_{i=1}^{I} Y_{ik} > 0$
    - $Z_k = 0$ if and only if $\sum_{i=1}^{I} Y_{ik} = 0$
    - $Y_{ik}$’s are observable when $Z_k = 0$ and there are no testing errors
    - $Y_{ik}$’s are unobservable when $Z_k = 1$

Basic statistics

- Example random variables

Basic statistics

- Example observed values

Basic statistics

- Example observed values
What is the relationship between $p = P(Y_{ik} = 1)$ and $\theta = P(Z_k = 1)$?

- Want to make inferences about $p$!
- $\theta = P(Z_k = 1) = P(\sum_{i=1}^{I} Y_{ik} > 0) = P(\text{at least one item is positive})$
- $1 - P(\sum_{i=1}^{I} Y_{ik} = 0) = 1 - P(\text{no items are positive})$
- $1 - P(Y_{ik} = 0, \forall i) = 1 - P(\text{all items are negative})$
- $1 - P(Y_{1k} = 0) \ast P(Y_{2k} = 0) \ast \cdots \ast P(Y_{Ik} = 0) = 1 - (1 - p)^I$
- Then $p = 1 - (1 - \theta)^{1/I}$

Choice of the group size, $I$, is critical!

- $p = 1 - (1 - \theta)^{1/I}$ and $\theta = 1 - (1 - p)^I$
- If $\theta$ is close to 1, all groups are likely to test positive
- If $\theta$ is close to 0, all groups are likely to test negative
- Choose group size, $I$, so that this does not happen
- “Rule of thumb” is to choose $I$ so that $\theta = 0.5$
  - Other values of $\theta$ between 0.2 and 0.8 may be optimal
    - Optimal means smallest MSE
    - Table in Swallow (*Phytopathology*, 1985)
  - Problem: Need to know $p$!

What is an estimate of $p$?

- Let $T$ be a random variable denoting the number of positive groups
  - $T = \sum_{k=1}^{K} Z_k$
  - $T \sim \text{Binomial}(K, \theta)$
  - MLE for $\theta$ is $\hat{\theta} = T/K$
- Use invariance property of MLEs, to get the MLE for $p$ to be $\hat{p} = 1 - (1 - \hat{\theta})^{1/I}$
- Positively bias for finite samples
- Ways to correct bias are discussed in
  - Tebbs, Bilder, and Moser (*Communications*, 2003)
  - Bilder and Tebbs (*Biometrical Journal*, 2005)

Using delta-method, one can show that $\sqrt{n} \left( \hat{p} - p \right) \xrightarrow{d} N(0, V(p))$ where $V(p) = I^{-2}[1 - (1 - p)^I](1 - p)^2 - I$.

- With individual testing $I = 1$, this simplifies to $V(p) = p(1 - p)$
- $(1 - \alpha)100\%$ Wald confidence interval (Bhattacharyya et al., *American Journal of Epidemiology*, 1979): $\hat{p} \pm z_{1-\alpha/2} \sqrt{V(\hat{p})/K}$
- Poor coverage!
- Tebbs and Bilder (*JABES*, 2004)
  - Adaptation of Blaker’s (2001) interval for a proportion under individual testing is the best
  - 95% C.I., $K = 40$, and $I = 10$
Basic statistics

- Testing or measurement errors
  - False positive – group tests positive when all items are really negative
  - False negative – group tests negative when at least one item is really positive
  - What happens to \( p \)?
    - Let \( \tilde{Z}_k = 1 \) if the group is truly positive
    - Let \( \tilde{Z}_k = 0 \) if the group is truly negative
  - Sensitivity = \( \eta = P(Z_k = 1 | \tilde{Z}_k = 1) \) for all \( k \)
  - Specificity = \( \delta = P(Z_k = 0 | \tilde{Z}_k = 0) \) for all \( k \)
    - Want to be as close to 1 as possible (often are close)
    - Usually treated as fixed constants
  - \( P(Z_k = 1) = \theta = \eta + (1 - \delta - \eta)[1 - P(\tilde{Z}_k = 1)] \)
  - \( p = 1 - [1 - P(\tilde{Z}_k = 1)]^{1/3} \)

Basic statistics

- Identification problem
  - Dorfman \((Annals of Mathematical Statistics, 1943)\)
    - Retest all items in a positive group
    - Often credited for the very first use of group testing
  - Sterrett \((Annals of Mathematical Statistics, 1957)\)
    - Retest randomly selected individual items until first positive is found
    - Remaining items are tested in a smaller group
      - If this smaller group is negative, retesting is completed
      - If this smaller group is positive, the same retesting procedure as initially performed continues
    - Procedure ends when all individuals are exhausted or a group tests negative
  - Smaller number of expected retests than Dorfman

Basic statistics

- Testing or measurement errors (continued)
  - Does group testing result in a loss of accuracy (i.e. lower \( \eta \) and \( \delta \)) when compared to individual testing?
    - ELISA tests for HIV screening – Group size \( \leq 15 \) have negligible loss (Kline et al., *Journal of Clinical Microbiology*, 1989)
    - Rapid HIV antibody assays – Group size \( \leq 20 \) no loss (Soroka et al., *Journal of Clinical Virology*, 2003)
    - NATs – Group size \( \leq 50 \) no loss (Bush et al., *New England Journal of Medicine*, 1991)
    - Behets et al. \((AIDS, 1990)\) show that the specificity is actually higher with group testing
    - Less number of errors overall with group testing since there are less tests!

Basic statistics

- Identification problem (continued)
  - Sobel and Elashoff \((Biometrika, 1975)\) use halving
    - Positive groups are divided into halves for retesting
    - Subsets that test positive are again halved and retested until all positive items have been identified
    - Litvak et al. \((JASA, 1994)\) presents a variation
      - Positive groups are split into several subgroups
    - See Gupta and Malina \((Statistics in Medicine, 1998)\) for a summary
Hepatitis C prevalence

- Worldwide prevalence is around 3%
- Liu et al. (*Transfusion, 1997*)
  - First paper on Hepatitis C virus (HCV) and group testing
  - HCV prevalence in Xuzhou City, China
  - Show how well group testing does compared to individual testing
    - BOTH individual and group testing data collected!
- ELISA test
  - Blood samples
  - Detect antibodies produced by the body when infected with HCV
  - Testing errors were not accounted for in their final estimates
- Individual testing
  - 1,875 blood samples screened
  - There were 42 positives

Blood donation screening

- Screening for infectious diseases is needed to ensure safety of blood supply
- Group testing is used!
- Dodd et al. (*Transfusion, 2002*)
  - American Red Cross blood donors
  - HIV, Hepatitis B, Hepatitis C, and human T cell lymphotropic virus
    - Estimation problem
    - Identification problem
  - How many donations need to be screened?
    - For this study (1998 – 2001), there were 19,811,809
  - Prevalence very small
    - Initial screening of people through a questionnaire also lowers prevalence

Hepatitis C prevalence

- Group testing
  - $K = 375$ groups
  - $I = 5$ individuals per group (samples pooled consecutively)
  - $t = \sum_{k=1}^{K} z_k = 37$ positive groups
- Estimates of $p$, probability individual is positive
  - Using individual data: $\hat{p} = 42/1875 = 0.0224$
  - Using group data: $\hat{p} = 1 - (1 - \hat{\theta})^{1/5} = 1 - (1 - 37/375)^{1/5} = 0.0206$
- Which is easier and more cost effective?
  - 1875 tests using individual testing
  - 375 tests using group testing
- Only the estimation problem of interest here

Blood donation screening

- Dodd et al. (*Transfusion, 2002*)
  - Specifically for HIV and Hepatitis C
    - Starting in 1999, NATs for groups
      - Actually look for HCV RNA and HIV RNA
      - Groups of 128 samples from March to September 1999
      - Groups of 16 after September 1999
    - Each positive group has all of its items retested (Dorfman method)
  - Stramer et al. (*Transfusion, 2000*) discusses the exact process of declaring negative or positive
Multiple vector transfer designs

- Plant pathologists often want to estimate the probability, $p$, an insect vector transfers a pathogen (virus, bacteria, etc.) to a plant
  - Swallow (Phytopathology, 1985, 1987)

- British plant pathologists were first to use group testing (Watson, 1936) despite Dorfman (1943) usually receiving the credit

Multiple vector transfer designs

- Group testing (multiple vector transfer)

\[
\begin{align*}
    z_i = 0 & \text{ if plant is negative, } 1 \text{ if plant is positive} \\
    z_i = 0 & \text{ if plant is negative, } 1 \text{ if plant is positive} \\
    z_{K-1} = 0 & \text{ if plant is negative, } 1 \text{ if plant is positive}
\end{align*}
\]

- Otherwise infeasible experiments are made feasible by using group testing!

- Ornaghi et al. (Maydica, 1999)
  - Location: Argentina
  - Plant: Corn
  - Planthopper: *Delphacodes kuscheli*
  - Virus: Mal Rio Cuarto
    - $120$ million in damages during the 1996–1997 agricultural season in Argentina
    - Most important corn virus (Lenardon et al., Plant Disease, 1998)
  - Goal: Estimate the probability of virus transmission by planthoppers that are known sources of the virus
  - Study done in stages – examine just the fourth stage
Multiple vector transfer designs

- Ornaghi et al. (*Maydica*, 1999)
  - $K = 24$ test plants were individually isolated in cages with the planthopper vectors for 48 hours at a common temperature
  - 7 insect vectors per plant
  - ELISA tests used to judge each test plant as infected or not
  - $t = \sum_{i=1}^{r} z_t = 3$ test plants were observed as infected
    - $\hat{p} = 1 - (1 - \hat{\theta})^{1/7} = 1 - (1 - 3/24)^{1/7} = 0.019$
    - 95% Wald C.I. for $p$: $(-0.0023, 0.0401)$
      - Lower bound is negative!
      - Blaker C.I. for $p$: $(0.0051, 0.0513)$
  - Estimation problem only
  - Does not matter which vector transmits the virus

Drug discovery experiments

- Screen hundreds of thousands of chemical compounds to look for potentially good ones
  - These compounds may eventually lead to new drugs
  - Only a very small amount are “active” or “potent”
  - 1 out of 10,000 according to Delvin (1997)
- Use group testing!
  - The process
    - Chemical compounds are placed in the wells of a plate
    - All compounds in a row (or column) are combined into a group
    - Each group is tested to determine active or inactive pools
    - In active pools, “decoding” is used to further identify which compounds are active – identification problem

Other examples

- Veterinary
  - Bovine viral diarrhea virus infection in cattle (Munoz-Zanzi et al., *J. of Veterinary Diagnostic Investigation*, 2000)
  - Avian pneumovirus (APV) in turkeys (Maherchandani et al., *J. of Veterinary Diagnostic Investigation*, 2004)
- Quality control – identify defective items
  - Johnson, Kotz, and Wu (1991) book – see Section 2
- DNA or RNA pooling
  - Pfeiffer et al. (*Genetic Epidemiology*, 2002)
    - “Efficiency of DNA pooling to estimate joint allele frequencies and measure linkage disequilibrium”
  - Kendiziorski et al. (*Biostatistics*, 2003)
    - “The efficiency of pooling mRNA in microarray experiments”

References

- Zhu, Hughes-Oliver, and Young (*Biometrics*, 2001)
- Xie, Tatsoka, Sacks, and Young (*JASA*, 2001)
- Works at GlaxoSmithKline
References