

CALIFORNIA STATE UNIVERSITY, HAYWARD  
STATISTICS DEPARTMENT

Statistics 6401 Graduate Probability I  
Fall 2004

Take Home Midterm 1

1. Suppose that a large number,  $n$ , of blood samples are to be screened for a relatively rare disease. If each sample is assayed individually,  $n$ , tests will be required. On the other hand, if each sample is divided in half and one of the halves is put into a pool with all the others, the pooled lot can be tested. Then, provided that the test method is sensitive enough, if this test is negative, no further assays are necessary and only one test has to be preformed. If the test on the pooled blood is positive, each reserved half-sample can be tested individually. In this case, a total of  $n + 1$  tests will be required. It is therefore plausible, assuming that the disease is rare, that some savings can be achieved through this pooling procedure.

To analyze this situation, suppose that the  $n$  samples are first grouped into  $m$  groups of  $k$  sample each, or  $n = mk$ . Each group is then tested, if a group tests positive then each individual in the group is tested. Let  $X_i$  be the number of tests run on the  $i^{th}$  group, the total number of tests run is  $N = \sum_{i=1}^m X_i$ .

- (a) Compute the expected total number of tests  $E[N]$ . Hint: Find the  $E[X_i]$  first.  
(b) Compute the probability that a sample tests positive with prevalence  $\pi = P(D)$ , sensitivity  $\eta = P(+|D)$ , and specificity  $\theta = P(-|D^c)$ .  
(c) Run the following S-Plus program many times and compare the simulated value of  $N$  with the expected value  $E[N]$ .  
(d) Change the sensitivity and specificity to 0.95, how does this change the results?

### Splus program that simulates groups testing.

```
n <- 100000 # samples
m <- 10000 # groups
k <- n/m # samples in each group Note: n = m*k

Pi <- 0.05 # prevalence P(D)
Eta <- 1 # sensitivity P(+|D)
Theta <- 1 # specificity P(-|D^c)

p <- 1 - ( Eta*Pi + (1-Theta)*(1-Pi) ) # probability a sample tests negative

N <- 0 # count of number of positive tests

for (i in 1:m){
  # test m groups
  if(rbinom(1,1,p^k) == 1){
    # group is negative, with probability p^k, and only one test needs to be performed
    N <- N + 1
  }
  else{
    # group is positive, so one test is done and k more need to be performed
    N <- N + k + 1
  }
}

# the simulates value for N
N
```

2. A smooth probability density estimate can be constructed in the following way. Let  $w(x)$  be a nonnegative, symmetric weight function, centered at zero and integrating to 1. For example,  $w(x)$  can be a standard normal density. The function

$$w_h(x) = \frac{1}{h} w\left(\frac{x}{h}\right) \quad (1)$$

is a rescaled version of  $w$ . As  $h$  approaches infinity,  $w_h$  becomes more spread out and flatter. If  $w(x)$  is the standard normal density, then  $w_h(x)$  is the normal density with standard deviation  $h$ . If  $X_1, \dots, X_n$  is a sample from a probability density function,  $f$ , and estimate of  $f$  is

$$f_h(x) = \frac{1}{n} \sum_{i=1}^n w_h(x - X_i). \quad (2)$$

This estimate, called the *kernel probability density estimate*, consists of the superposition of “hills” centered over the observations. In the case where  $w(x)$  is the standard normal density  $w_h(x - X_i)$  is the normal density with mean  $X_i$  and standard deviation  $h$ .

The parameter  $h$ , the bandwidth of the estimating function, controls its smoothness and corresponds to the bin width of the histogram.

- (a) In S-Plus type at the command prompt

```
> help(density)
```

What parameter of the function controls the bandwidth? What are the possible weight functions?

- (b) Download the data file `beeswax.dat` from the class website and run the following S-Plus program. Changing the bandwidth parameter, choose what you think is the best value and print the picture.
- (c) Try the other window functions allowed. How does the density estimate change? Hint: It may help to lower the bandwidth to see the differences.

```
### density plots for meltpt of beeswax

bees <- matrix(scan("H:\\hw3\\beeswax.dat"), ncol=2, byrow=T)
bees

meltpt <- bees[,1]
phydcarb <- bees[,2]

par(mfrow=c(2,1))

hist(meltpt)

# density plot for meltpt

plot(density(meltpt, n=10, window="g"), type="b", xlab="Melting Point C", ylab="density")
```