Incidentalome_Figure.R

jamesdiao

Mon Jun 13 16:11:01 2016

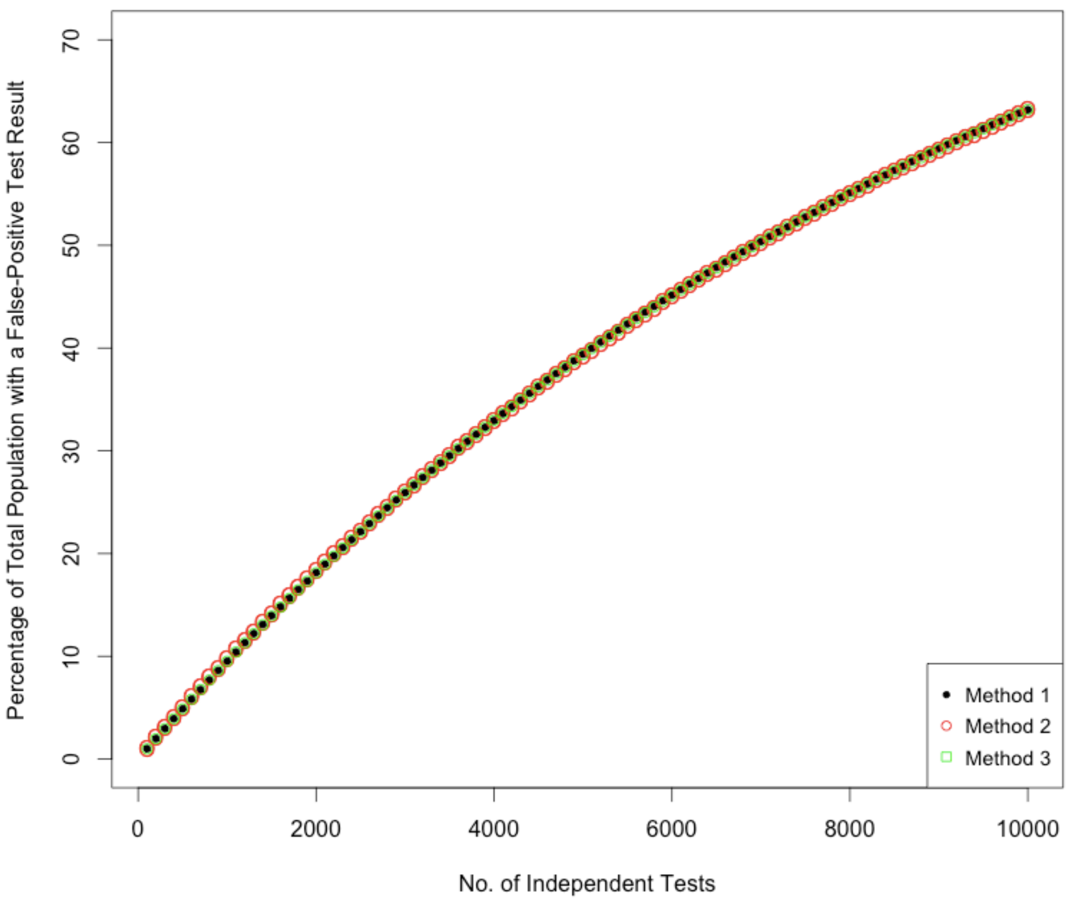
```r
#Method 1: simulation | sample
method_1 <- function(false_positive_rate, population, test_num) {
  results <- rep(0,population)
  result_list <- rep(0,test_num)
  for (i in 1:test_num)
  {
    results[sample(population,10)] <- 1
    result_list[i] <- mean(results)
  }
  return(result_list)
}
#Method 2: simulation | rbinomial
method_2 <- function(false_positive_rate, population, test_num) {
  results <- rep(0,population)
  result_list <- rep(0,test_num)
  for (i in 1:test_num)
  {
    results <- results | rbinom(population,1,prob = false_positive_rate)
    result_list[i] <- mean(results)
  }
  return (result_list)
}
#Method 3: calculation | expected values
method_3 <- function(false_positive_rate, test_num) {
  return(1-(1-false_positive_rate)^c(1:test_num))
}

result_list_1 <- method_1(0.0001, 100000, 10000)
result_list_2 <- method_2(0.0001, 100000, 10000)
result_list_3 <- method_3(0.0001, 10000)
subpoints <- seq(100,10000,100)

plot(subpoints,100*result_list_1[subpoints], main = "Figure. Percentage of
Total Population with a False-Positive Test Result", pch = 20, ylim =
c(0,70), xlab = "No. of Independent Tests", ylab = "Percentage of Total
Population with a False-Positive Test Result")
points(subpoints,100*result_list_2[subpoints], pch = "O", col = "red")
points(subpoints,100*result_list_3[subpoints], col = "green", ylim = c(0,70))

legend("bottomright",c("Method 1","Method 2", "Method 3"), col =
c("black","red","green"), pch = c(20,1,0), cex = 0.9)
```

Figure. Percentage of Total Population with a False-Positive Test Result



Assumptions:

1. Kohane et al. samples 10 false-positives from the population for each test, assuming that there will be **exactly 10 with each test**. This is the expected value, but the true number of false-positives could be almost anything. A more realistic (and noisier) method would be to simulate each trial as *n* Bernoulli trials with *p* as the false-positive-rate (Method 2, in red).
2. Kohane et al. assumes that **all tests are mutually independent**. In fact, a patient who receives 1 false-positive may be more or less likely to test positive on subsequent tests.
3. Kohane et al. assumes that the **false-positive-rate is constant** between all tests.
4. Kohane et al. assumes that **false-positives can come from the whole population** (100,000), when they can actually only come from population-1, or 99,999. This is because one person is a true positive and excluded from the pool for selection. This means that Kohane et al.'s estimate is very very slightly biased upward (though the number 100,000 was arbitrary anyway, so it really doesn't matter).

The slope/concavity of the curve comes from some patients receiving multiple false-positives. A simple probability calculation gives the equation:

$$P(at\ least\ 1\ false\ positive) = 1 - (1 - false.positive.rate)^n$$

This is illustrated by method 3, plotted in green, which follows our simulation closely (r = 0.999999). **Thus, the curve is dictated entirely by the false-positive-rate.**