Charlotte Smith
Term Project
Fall 2007

Total Coliform Results in Very Large Drinking Water Systems

Background:

In the United States, water utilities must collect water samples each month and test for Coliform bacteria as a surrogate for microbes in the drinking water system. The number of tests is a function of population size, with larger utilities collecting more tests than smaller utilities. If a test is positive, the utility collects 3 additional samples for testing (one at the same site, one upstream and one downstream). Drinking water regulations stipulate that no more than 5% of the tests performed each month can be positive for Coliform bacteria.

The Association of Metropolitan Water Agencies (AMWA), an industry association representing large public water systems recently requested data from its members related to Coliform test results. The utilities entered the number of Coliform tests performed for the year 2005, as well as the number of routine and repeat tests that were positive for Coliform bacteria.

Purpose:

The object of this exercise was to determine the odds of a positive test in a routine and repeat sample. An additional objective was to plot the distribution of positive routine tests for the 52 drinking water utilities that responded to AMWA's request.

Methods:

AMWA requested the data via email to its members with an attached Excel spreadsheet. Fifty-two Excel spreadsheets were provided by AMWA for analysis. The spreadsheets were converted to .csv format and analyzed using a program written in R. Separate loops were created to read the number of samples, number of positive routine samples and number of positive repeat samples to obtain probabilities and odds ratios.

In addition, a loop was created to obtain the proportion of routine samples by utility (for the plot).

The R code is found in Appendix 1.

Results:

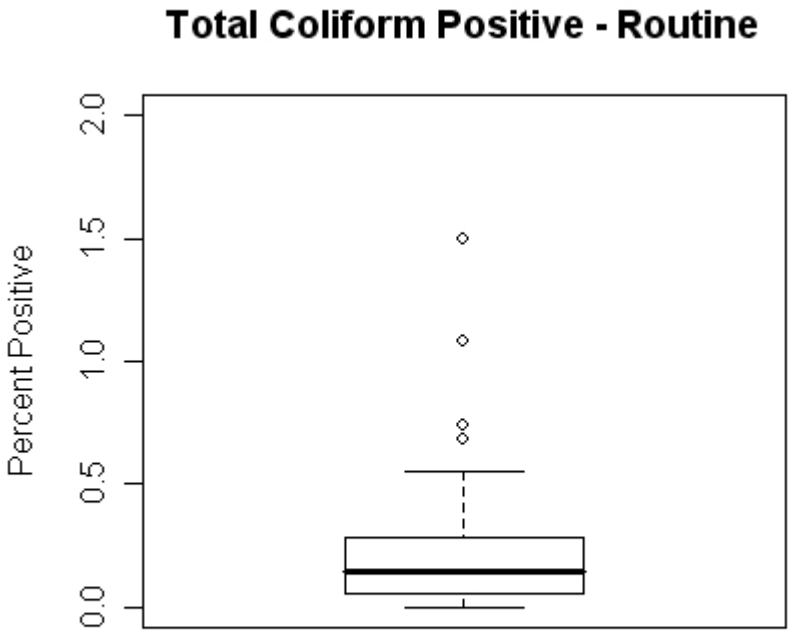
For the 52 systems in the AMWA database, the odds of a TC positive in a repeat test (N=977) were 8.2 times the odds of a TC positive in a routine test (N=139,666).

Alternatively, the probability of a TC positive result for a repeat test was 1.7%, and the probability of a TC positive result for a routine test was 0.2%.

```
       Repeat Routine
TC Pos     17     300
TC Neg    960  139366
```

```
                Repeat     Routine
risk       0.01740020 0.002147982
risk.ratio 8.10072330 1.000000000
odds       0.01770833 0.002152605
odds.ratio 8.22646528 1.000000000
```

The plot below shows the distribution of Total Coliform bacteria tests that were positive for all 52 water utilities that participated in the AMWA survey. For nearly all of the utilities, less than one percent of the tests were positive. In about half of the utilities, less than 0.1% of the tests performed in 2005 were positive for Coliform bacteria.



Conclusions:

The 52 AMWA members that responded to the request for data were well below the 5% allowable level for positive Coliform tests in their systems. There may be some selection bias in that drinking water systems that were above the level may have preferred not to participate in this voluntary survey.

Appendix 1

R code:

```
#Samples in all systems
files<-list.files("C:\\CSV_F_1\\",full=TRUE, pattern=".csv")
samps <- c()
for(f in files)   {
data <- read.csv(f, header=F, stringsAsFactors=F)
y <- as.numeric(data[5,6])
cat("y",y, "\n")
samps <- c(samps, y)
}
samps
samples<-sum(samps, na.rm = T)
samples

#Positive TC samples for all systems
files<-list.files("C:\\CSV_F_1\\",full=TRUE, pattern=".csv")
tc.pos.a <- c()
for(f in files)   {
data <- read.csv(f, header=F, stringsAsFactors=F)
x <- as.numeric(data[5,7])
cat("x",x,"\n")
tc.pos.a <- c(tc.pos.a, x)
}
tc.pos.a
tc.pos.all<-sum(tc.pos.a, na.rm = T)
tc.pos.all

#TC negative
tc.neg<-samples-tc.pos.all
tc.neg


#proportion of samples TC positive by system
files<-list.files("C:\\CSV_F_1\\",full=TRUE, pattern=".csv")
tc.pos <- c()
for(f in files)   {
data <- read.csv(f, header=F, stringsAsFactors=F)
n <- as.numeric(data[5,7])
d <- as.numeric(data[5,6])
cat("n",n,"d",d,"\n")
tc.pos <- c(tc.pos, n/d)
}
percent<-tc.pos*100
boxplot(percent,main='Total Coliform Positive - Routine',ylim=c(0,2.0),
ylab='Percent Positive')

#proportion repeat samples TC positive:
files<-list.files("C:\\CSV_F_1\\",full=TRUE, pattern=".csv")
rpt.tc.pos <- c()
for(f in files)   {
data <- read.csv(f, header=F, stringsAsFactors=F)
nn <- as.numeric(data[5,12])
dd <- as.numeric(data[5,11])
```

```
cat("nn",nn,"dd",dd,"\n")
rpt.tc.pos <- c(rpt.tc.pos, nn/dd)
}
rpt.tc.pos
rpt.percent<-rpt.tc.pos*100
boxplot(rpt.percent,main='Total Coliform Positive - Repeat',ylim=c(0,1.0),
ylab='Percent Positive')

# Repeat samples in all systems
files<-list.files("C:\\CSV_F_1\\",full=TRUE, pattern=".csv")
rpt.samps <- c()
for(f in files)   {
data <- read.csv(f, header=F, stringsAsFactors=F)
y <- as.numeric(data[5,11])
cat("y",y, "\n")
rpt.samps <- c(rpt.samps, y)
}
rpt.samps
rpt.samples<-sum(rpt.samps, na.rm = T)
rpt.samples

#Positive TC samples for all systems
files<-list.files("C:\\CSV_F_1\\",full=TRUE, pattern=".csv")
rpt.tc.pos.a <- c()
for(f in files)   {
data <- read.csv(f, header=F, stringsAsFactors=F)
x <- as.numeric(data[5,12])
cat("x",x,"\n")
rpt.tc.pos.a <- c(rpt.tc.pos.a, x)
}
rpt.tc.pos.a
rpt.tc.pos.all<-sum(rpt.tc.pos.a, na.rm = T)
rpt.tc.pos.all

#TC negative
rpt.tc.neg<-rpt.samples-rpt.tc.pos.all
rpt.tc.neg


#Make table of Routine Repeat TC positive data and Samples for each system:
rpt.tbl<-cbind(Code, rpt.tc.pos.a, rpt.samps)
rpt.tbl
#To find original names of strange spreadsheets:
files<-list.files("C:\\CSV_F_1\\",full=TRUE, pattern=".csv")
files

#Note: fixed Lakehaven, Orange, San Juan, Augusta, Desmoines,
# Beaver, Santa Fe - all ok now
```

```
#####################################################################
#########ODDS RATIOs -- See page 46 chapter 2 Aragon_2007#####
#####################################################################

tc.pos.all
tc.neg
samples
rpt.tc.pos.all
rpt.tc.neg
rpt.samples

dat<-matrix(c(300,139366,17,960), 2,2)
rownames(dat)<- c("TC Pos", "TC Neg")
colnames(dat)<- c("Routine","Repeat")
coltot<-apply(dat,2,sum); #should be same as samples
risk<-dat["TC Pos",]/coltot
risk.ratio<-risk/risk[2]
odds<-risk/(1-risk)
odds.ratio<-odds/odds[2]
#Display results
dat
rbind(risk, risk.ratio, odds, odds.ratio)


#Repeat as the "exposure"
dat<-matrix(c(17,960, 300,139366), 2,2)
rownames(dat)<- c("TC Pos", "TC Neg")
colnames(dat)<- c("Repeat","Routine")
coltot<-apply(dat,2,sum); #should be same as samples
risk<-dat["TC Pos",]/coltot
risk.ratio<-risk/risk[2]
odds<-risk/(1-risk)
odds.ratio<-odds/odds[2]
#Display results
dat
rbind(risk, risk.ratio, odds, odds.ratio)
```