# GR5234 - HW5

*Mathieu Sauterey - UNI: mjs2364*

*October 27, 2018*

## Problem 1

An accounting firm is interested in estimating the error rate in a compliance audit it is conducting. The population contains 828 claims, and the firm audits a simple random sample of 85 of those claims. In each of the 85 sampled claims, 215 fields are checked for errors. One claim has errors in 4 of the 215 fields, 1 claim has 3 errors, 4 claims have 2 errors, 22 claims have 1 error, and the remaining 57 claims have no errors.

### (a) Treating the claims as psus and the observations for each field as ssus, estimate the error rate, defined to be the average number of errors per field, along with the standard error for your estimate. Give a 95% confidence interval.

```
N = 828

n = 85

M = rep(215, n)

m = rep(215, n)

y = data.frame(y4 = rep(c(1,0), times = c(4,211)),
               y3 = rep(c(1,0), times = c(3,212)),
               y2 = rep(c(1,0), times = c(2,213)),
               y1 = rep(c(1,0), times = c(1,214)),
               y0 = rep(c(0), times = c(215)))

ybar = rep(apply(y, 2, mean), c(1,1,4,22,57))

s2 = rep(apply(y, 2, var), c(1,1,4,22,57))

ybar_hat_r = sum(M * ybar) / sum(M)

s2.r <- var(M * (ybar - ybar_hat_r))

V.term2 <- sum(M^2 * s2/m * (1 - m/M))

V.hat.ybar <- 1/mean(M)^2 * (s2.r/n * (1 - n/N) + 1/(n*N) * V.term2)

SE.ybar <- sqrt(V.hat.ybar)


print(ybar_hat_r)
```

```
## [1] 0.002024624
```

```r
print(SE.ybar)
```

```
## [1] 0.0003570679
```

```r
# 95% CI for the average number of errors per field
( ybar_hat_r + c(-1,1) * 1.96 * SE.ybar )
```

```
## [1] 0.001324771 0.002724477
```

$\mathrm{E}[\hat{\bar{y}}_r] = 0.002024624$

$\mathrm{SE}[\hat{\bar{y}}_r] = 0.0003570679$

[ 0.001324771 ; 0.002724477 ] is a 95% CI for the average number of errors per field

### (b) Estimate (with standard error) the total number of errors in the 828 claims. Give a 95% confidence interval.

```r
t_hat_unb = (N/n) * sum(M * ybar)

s2.t <- var(M * ybar)

V.hat.t <- N^2 * s2.t/n * (1 - n/N) + (N/n) * V.term2

SE.t <- sqrt(V.hat.t)

print(t_hat_unb)
```

```
## [1] 360.4235
```

```r
print(SE.t)
```

```
## [1] 63.56523
```

```r
# 95% CI for the average number of errors per field
( t_hat_unb + c(-1,1) * 1.96 * SE.t )
```

```
## [1] 235.8357 485.0114
```

$\mathrm{E}[\hat{t}_{unb}] = 360.42$ errors

$\mathrm{SE}[\hat{t}_{unb}] = 63.57$ errors

[236 ; 485 ] is a 95% CI for the total number of errors in the 828 claims.

## Problem 2

An inspector samples cans from a truckload of canned creamed corn to estimate the prevalence of worm fragments in the product. The truck has 580 cases; each case contains 24 cans. The inspector samples 12 cases at random, and subsamples 3 cans randomly from each selected case.

### (a) Estimate the mean number of worm fragments per can, along with the standard error for your estimate.

```r
N = 580

n = 12

M = rep(24, n)

m = rep(3, n)

y = data.frame(can1 = c(1,4,0,3,4,0,5,3,7,3,4,0),
               can2 = c(5,2,1,6,9,7,5,0,3,1,7,0),
               can3 = c(7,4,2,6,8,3,1,2,5,4,9,0))

ybar = apply(y, 1, mean)

s2 = apply(y, 1, var)

ybar_hat_r = sum(M * ybar) / sum(M)

s2.r <- var(M * (ybar - ybar_hat_r))

V.term2 <- sum(M^2 * s2/m * (1 - m/M))

V.hat.ybar <- 1/mean(M)^2 * (s2.r/n * (1 - n/N) + 1/(n*N) * V.term2)

SE.ybar <- sqrt(V.hat.ybar)


print(ybar_hat_r)
```

```
## [1] 3.638889
```

```r
print(SE.ybar)
```

```
## [1] 0.6101924
```

$\mathrm{E}[\hat{\bar{y}}_r] = 3.64$

$\mathrm{SE}[\hat{\bar{y}}_r] = 0.61$

### (b) Estimate the total number of worm fragments in the truckload, along with the standard error for your estimate.

```r
t_hat_unb = (N/n) * sum(M * ybar)

s2.t <- var(M * ybar)

V.hat.t <- N^2 * s2.t/n * (1 - n/N) + (N/n) * V.term2

SE.t <- sqrt(V.hat.t)

print(t_hat_unb)
```

```
## [1] 50653.33
```

```r
print(SE.t)
```

```
## [1] 8493.878
```

$E[\hat{t}_{unb}] = 50653$ worm fragments

$SE[\hat{t}_{unb}] = 8494$ worm fragments

# Problem 3

A researcher took a simple random sample of 4 high schools from a region with 29 high schools for a study on the prevalence of smoking among female high school students in the region. The results were as follows.

### (a) Estimate the percentage of female high school students in the region who smoke, along with a 95% confidence interval.

```r
N = 29

n = 4

M = c(792, 447, 511, 800)

m = c(25,15,20,40)

y = list(school1 = rep(c(1,0), times = c(10,15)),
               school2 = rep(c(1,0), times = c(3,12)),
               school3 = rep(c(1,0), times = c(6,14)),
               school4 = rep(c(1,0), times = c(27,13)))

ybar = sapply(y, mean)

s2 = sapply(y, var)

ybar_hat_r = sum(M * ybar) / sum(M)

s2.r <- var(M * (ybar - ybar_hat_r))

V.term2 <- sum(M^2 * s2/m * (1 - m/M))

V.hat.ybar <- 1/mean(M)^2 * (s2.r/n * (1 - n/N) + 1/(n*N) * V.term2)

SE.ybar <- sqrt(V.hat.ybar)


print(ybar_hat_r)
```

```
## [1] 0.4311765
```

```r
print(SE.ybar)
```

```
## [1] 0.09910716
```

```r
# 95% CI for the average number of errors per field
( ybar_hat_r + c(-1,1) * 1.96 * SE.ybar )
```

```
## [1] 0.2369264 0.6254265
```

$E[\hat{\bar{y}}_r] = 43.12$ %

$SE[\hat{\bar{y}}_r] = 9.91$ %

[23.69 % ; 62.54 % ] is a 95% CI for the percentage of female high school students in the region who smoke.

### (b) Estimate the total number of female high school students in the region who smoke, along with a 95% confidence interval.

```r
t_hat_unb = (N/n) * sum(M * ybar)

s2.t <- var(M * ybar)

V.hat.t <- N^2 * s2.t/n * (1 - n/N) + (N/n) * V.term2

SE.t <- sqrt(V.hat.t)

print(t_hat_unb)
```

```
## [1] 7971.375
```

```r
print(SE.t)
```

```
## [1] 2725.67
```

```r
# 95% CI for the average number of errors per field
( t_hat_unb + c(-1,1) * 1.96 * SE.t )
```

```
## [1]  2629.061 13313.689
```

$E[\hat{t}_{unb}] = 7971$ female smokers

$SE[\hat{t}_{unb}] = 2726$ female smokers

[2629 ; 13314 ] is a 95% CI for the total number of female high school students in the region who smoke