MVE155 – Statistical inference, 31st January 2013

Computer exercise 1

Simon Sigurdhsson, 900322–0291

This is a report on exercise 7.65 in Rice (2007), which concerns a data set of breast cancer mortality in the adult caucasian female population of three U.S. states during the 1950s. The exercise consists of 14 parts labeled (a) through (n). The exercise has been solved using the R statistical software and the code used is included throughout the report. The preamble of this code simply loads a couple of packages along with the data set:

```r
library(ggplot2)
library(tikzDevice)
cancer <- read.csv("cancer.txt", header=FALSE,
                   col.names=c("mortality", "population"))
```

Here, the `ggplot2` library is used to generate figures and the `tikzDevice` library is used to export them as Ti*k*Z code.

a. *Histogram of the population values.*

```r
tikz("tikz/histogram.tex", width=4, height=2.5)
qplot(mortality, data=cancer, geom="histogram", binwidth=5) + xlim(0,400)
dev.off()
```

A histogram showing the distribution of mortalities is shown in figure 1 on the following page. As the histogram shows, the mean seems to be somewhere near ≈ 50 deaths.

b. *Key parameters of the data.*

```r
b.mean <- mean(cancer$mortality)
b.total <- sum(cancer$mortality)
b.variance <- var(cancer$mortality)
b.stddev <- sd(cancer$mortality)
```

The mean mortality of the population is 39.85 deaths, which is hinted at by the histogram in figure 1 on the next page. The total mortality in the data set is 11 997 persons. The population variance of the mortality is 2598.74, which gives a standard deviation of 50.98 deaths.

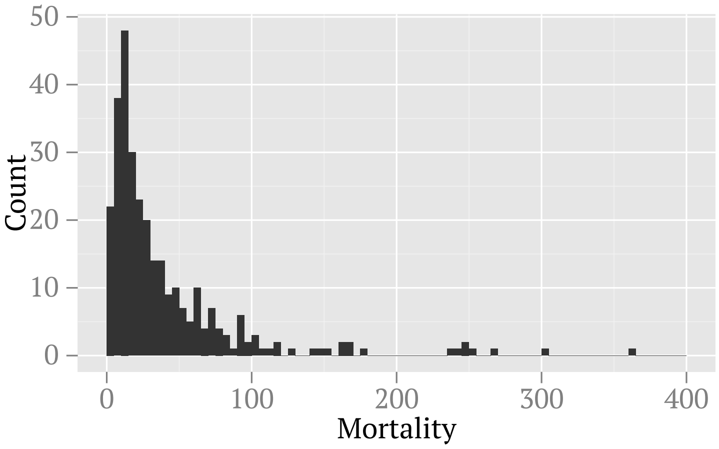

Figure 1: *Histogram of mortality rates.*

c. *Simulated sampling distribution.*

```
simulated.mean <- replicate(500, mean(sample(cancer$mortality, 25)))
tikz("tikz/simulated-histogram.tex", width=4, height=2.5)
qplot(simulated.mean, geom="histogram", binwidth=5) + xlim(0,400)
dev.off()
```

The sampling distribution is simulated by drawing 500 IID samples of 25 observations from the data set and calculating their mean, resulting in the histogram shown in figure 2 on page 5. unlike like the data itself, the sampling distribution seems to correspond well to a normal distribution. As with the actual data, the mean seems to be around 50.

d. *A simple random sample.*

```
random.sample <- sample(cancer$mortality, 25, replace=TRUE)
d.mean <- mean(random.sample)
d.total <- d.mean*301
```

A simple random sample of 25 observations is drawn from the data set. Using this sample, the population mean can be estimated to 29.24 deaths and the total mortality is estimated to 8801 persons.

e. *Variance estimate using the random sample.*

```
e.variance <- var(random.sample)/25*(1-24/300)
e.stddev <- sqrt(e.variance)
```

The variance of the population is estimated as $s^2 = \frac{\sigma^2}{n}(1 - \frac{n-1}{N-1})$, with $n = 25$ and $N = 300$. The result is an estimated variance of 49.66 and an estimated standard deviation of 7.05 deaths.

f. *Confidence intervals.*

```r
s.X <- e.stddev
f.mean.low <- d.mean - 1.96*s.X
f.mean.high <- d.mean + 1.96*s.X
f.total.low <- f.mean.low*301
f.total.high <- f.mean.high*301
```

Using the data calculated in (d) and (e), confidence intervals for the mean mortality and total mortality are calculated. These confidence intervals reveal that with 95 % certainty, the mean mortality of the population will be in [15.43, 43.05], while the total mortality will be in the interval [4644, 12 959]. Referring to (b), it is clear that the actual population parameters fall within these confidence intervals.

g. *A larger sample size.*

```r
random.sample <- sample(cancer$mortality, 100, replace=TRUE)
g.mean <- mean(random.sample)
g.total <- g.mean*301
g.variance <- var(random.sample)/100*(1-99/300)
g.stddev <- sqrt(g.variance)
s.X <- g.stddev
g.mean.low <- g.mean - 1.96*s.X
g.mean.high <- g.mean + 1.96*s.X
g.total.low <- g.mean.low*301
g.total.high <- g.mean.high*301
```

Repeating the procedure of (d)–(f) for a sample of size 100, a more accurate result should be obtained. The estimated mean is now 37.10 deaths, with an estimated total of 11 167 persons which is fairly close to the actual population total. The variance is calculated to 12.57 yielding a standard deviation of 3.54 deaths. The confidence intervals, with 95 % certainty, are [30.15, 44.04] for the population mean and [9076, 13 258] for the population total. Again, the actual values are within these ranges.

h. *A ratio estimator.*

Given the purpose and design of a ratio estimator, and the fact that information about the total number of adult females is available, it is likely that a ratio estimator will do a fairly good job of improving the estimates.

i. *Sampling distribution of the ratio estimator.*

```r
ratio.estimator.sample <- function(){
    temp <- colMeans(cancer[sample(1:dim(cancer)[1], 25),])
    unname(mean(cancer$population)*temp[1]/temp[2])
}
simulated.ratio <- replicate(500, ratio.estimator.sample())
tikz("tikz/simulated-ratio-histogram.tex", width=4, height=2.5)
```

```
qplot(simulated.ratio, geom="histogram", binwidth=2.5) + xlim(0,400)
dev.off()
```

As shown by figure 3 on the next page when compared to figure 2 on the following page, the ratio estimator does a fairly good job. The mean seems to be roughly the same, and the histogram itself implies a similar if not lower variance.

j. *A ratio estimator sample.*

```
random.sample <- cancer[sample(1:dim(cancer)[1], 25, replace=TRUE),]
sample.ratio <- mean(random.sample$mortality)/mean(random.sample$population)
j.ratio.mean <- mean(cancer$population)*sample.ratio
j.ratio.total <- j.ratio.mean*301
j.normal.mean <- mean(random.sample$mortality)
j.normal.total <- j.normal.mean*301
```

Drawing a sample of size 25 and calculating both ratio estimates and the statistics previously calculated in (d) illustrates the difference between these. The ratio estimate mean mortality is 38.88 compared to the “regular” mean 44.52, and the total mortality as calculated from the ratio estimate is 11 702 compared to the total mortality calculated from the regular mean, 13 401.

k. *Confidence interval of the ratio estimator.*

```
s.normal.X <- sqrt(var(random.sample$mortality)/25*(1-24/300))
k.normal.ci.low <- j.normal.mean - 1.96*s.normal.X
k.normal.ci.high <- j.normal.mean + 1.96*s.normal.X
s.xy <- sum((random.sample$mortality-mean(random.sample$mortality)) *
            (random.sample$population-mean(random.sample$population)))/24
s.x <- var(random.sample$population)/25*(1-24/300)
s.y <- var(random.sample$mortality)/25*(1-24/300)
s.r.square <- 1/24*(1-24/300)*1/(mean(random.sample$population)^2) *
            (j.ratio.mean^2*s.x + s.y - 2*j.ratio.mean*s.xy)
s.ratio.X <- sqrt(s.r.square)
k.ratio.ci.low <- j.ratio.mean - 1.96*s.ratio.X
k.ratio.ci.high <- j.ratio.mean + 1.96*s.ratio.X
```

The confidence interval of the ratio estimate is [35.34, 42.41], which very good compared to the regular confidence interval, [24.31, 64.73].

l. *A stratified population mean.*

```
strata.1 <- cancer[sort(cancer$population, index.return=TRUE)$ix,][1:75,   ]
strata.2 <- cancer[sort(cancer$population, index.return=TRUE)$ix,][76:150, ]
strata.3 <- cancer[sort(cancer$population, index.return=TRUE)$ix,][151:225,]
strata.4 <- cancer[sort(cancer$population, index.return=TRUE)$ix,][226:301,]
sample.1 <- sample(strata.1$mortality, 6, replace=TRUE)
sample.2 <- sample(strata.2$mortality, 6, replace=TRUE)
sample.3 <- sample(strata.3$mortality, 6, replace=TRUE)
sample.4 <- sample(strata.4$mortality, 6, replace=TRUE)
W.1 <- length(strata.1[,1])/301
```

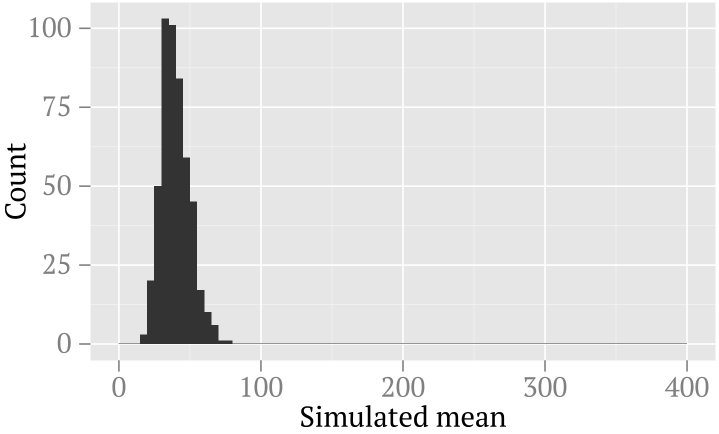

Figure 2: *Histogram of the simulated sampling distribution of the mean.*

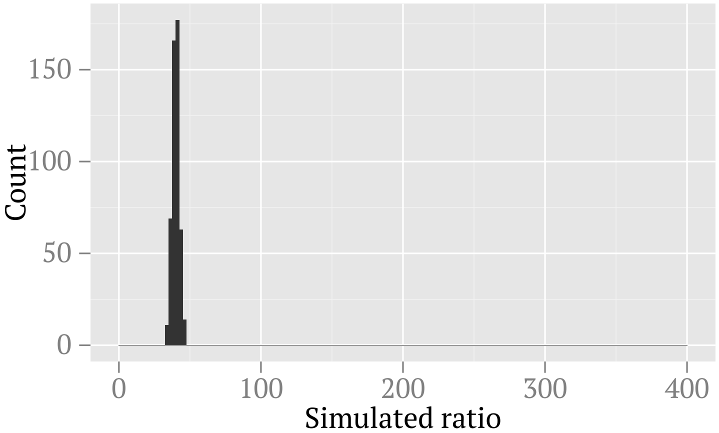

Figure 3: *Histogram of the simulated sampling distribution of the ratio estimate.*

```
W.2 <- length(strata.2[,1])/301
W.3 <- length(strata.3[,1])/301
W.4 <- length(strata.4[,1])/301
l.mean <- W.1*mean(sample.1)/301 + W.2*mean(sample.2) +
          W.3*mean(sample.3)/301 + W.4*mean(sample.4)
l.total <- l.mean*301
```

A stratified mean was calculated from samples of size 6 from the four stratas containing the first, second, third and fourth quarter of the data (sorted by population). This stratified mean was 29.09. The total mortality calculated from the stratified samples is 8758.

m. *Different allocations.*

```
srs.sample <- sample(cancer$mortality, 200, replace=TRUE)
m.srs.mean <- mean(srs.sample)
m.srs.variance <- var(srs.sample)/200*(1-199/300)
var.1 <- var(strata.1$mortality)
var.2 <- var(strata.2$mortality)
var.3 <- var(strata.3$mortality)
var.4 <- var(strata.4$mortality)
sum.w.var <- (W.1*var.1+W.2*var.2+W.3*var.3+W.4*var.4)
m.w.1 <- W.1*var.1/sum.w.var
m.w.2 <- W.2*var.2/sum.w.var
m.w.3 <- W.3*var.3/sum.w.var
m.w.4 <- W.3*var.4/sum.w.var
opt.sample.1 <- sample(strata.1$mortality, 1+round(m.w.1*199), replace=TRUE)
opt.sample.2 <- sample(strata.2$mortality, 1+round(m.w.2*199), replace=TRUE)
opt.sample.3 <- sample(strata.3$mortality, 1+round(m.w.3*199), replace=TRUE)
opt.sample.4 <- sample(strata.4$mortality, 1+round(m.w.4*199), replace=TRUE)
m.opt.mean <- (W.1*mean(opt.sample.1) + W.2*mean(opt.sample.2) +
               W.3*mean(opt.sample.3) + W.4*mean(opt.sample.4))
m.opt.variance <- (W.1*sd(opt.sample.1) + W.2*sd(opt.sample.2) +
                   W.3*sd(opt.sample.3) + W.4*sd(opt.sample.4))^2/200
prop.sample.1 <- sample(strata.1$mortality, 50, replace=TRUE)
prop.sample.2 <- sample(strata.2$mortality, 50, replace=TRUE)
prop.sample.3 <- sample(strata.3$mortality, 50, replace=TRUE)
prop.sample.4 <- sample(strata.4$mortality, 50, replace=TRUE)
m.prop.mean <- (W.1*mean(prop.sample.1) + W.2*mean(prop.sample.2) +
                W.3*mean(prop.sample.3) + W.4*mean(prop.sample.4))
m.prop.variance <- (W.1*var(prop.sample.1) + W.2*var(prop.sample.2) +
                    W.3*var(prop.sample.3) + W.4*var(prop.sample.4))/200
```

Proportional allocation of the four strata in (l) trivially yields sampling fractions $w_l \approx 1/4$. Optimal allocation instead yields sampling fractions $w_1 = 0.003$, $w_2 = 0.009$, $w_3 = 0.032$ and $w_4 = 0.943$. Table 1 on the next page shows the mean and variance for different sample allocations, with a total sample size of 200. The proportional allocation likely deviates because it is a bad allocation of samples (in this case, as indicated by the optimal allocation, extreme emphasis needs to

Table 1: *Mean and variance for different sample allocations.*

Method	**Mean**	**Variance**
Simple random	40.25	4.65
Stratified proportional	40.63	6.21
Stratified optimal	37.51	1.99

be put on the outliers in the fourth strata). All the samples are still very exact, but this is largely due to the large sample size, which is needed in order to get any samples at all from the first strata with optimal allocation.

n. *Introducing additional strata.*

Hopefully, as the number of strata grows larger the optimal allocation will become more even and the estimates more exact. However, as explained by Rice (2007, p. 238), in practice it is “rarely worthwile constructing more than a few strata”, which indicates that the gains of doing so are probably far outwegihed by the costs of having to increase the total sample size to get enough samples from every strata.

References

Rice, John (2007). *Mathematical statistics and data analysis*. Australia Belmont, CA: Thompson/Brooks/Cole. ISBN: 9780495118688.