Stat 138: Introduction to Sampling Designs
Problem Set 4

Anne Christine Amores

April 25, 2025

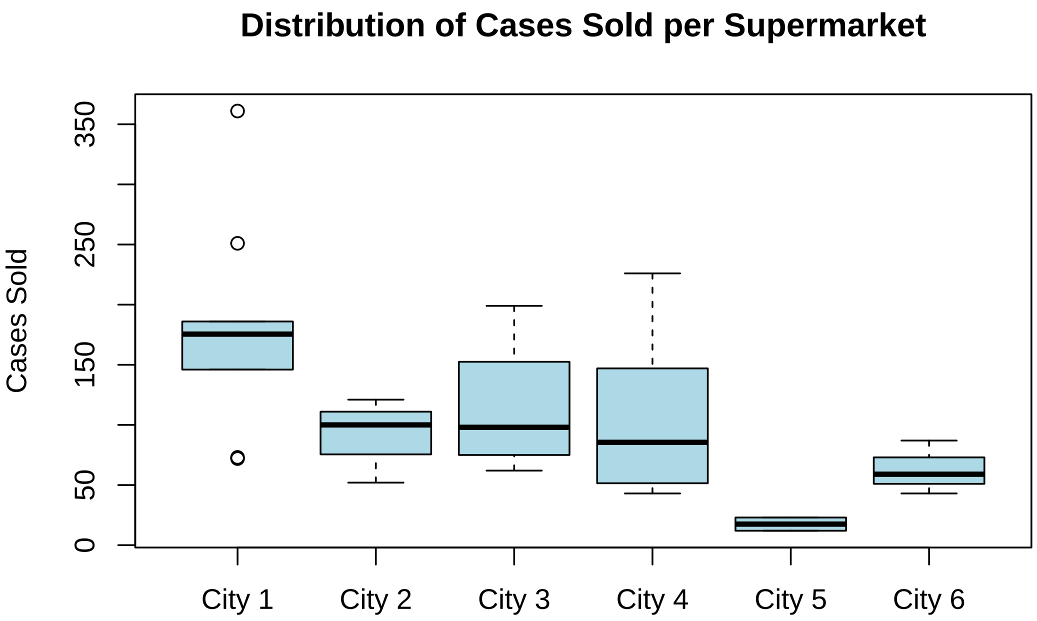
Green Globules

The new candy Green Globules is being test-marketed in an area of upstate New York. The market research firm decided to sample 6 cities from the 45 cities in the area and then to sample supermarkets within cities, wanting to know the number of cases of Green Globules sold.

City	Number of Supermarkets	Number of Cases Sold
1	52	146, 180, 251, 152, 72, 181, 171, 361, 73, 186
2	19	99, 101, 52, 121
3	37	199, 179, 98, 63, 126, 87, 62
4	39	226, 129, 57, 46, 86, 43, 85, 165
5	8	12, 23
6	14	87, 43, 59

Obtain summary statistics for each cluster. Plot the data, and estimate the total number of cases sold, and the average number sold per supermarket, along with the standard errors of your estimates.

```r
# Plotting the data
cases <- list(
  c(146, 180, 251, 152, 72, 181, 171, 361, 73, 186),
  c(99, 101, 52, 121),
  c(199, 179, 98, 63, 126, 87, 62),
  c(226, 129, 57, 46, 86, 43, 85, 165),
  c(12, 23),
  c(87, 43, 59)
)
boxplot(cases,
        names = paste("City", 1:6),
        col = "lightblue",
        main = "Distribution of Cases Sold per Supermarket",
        ylab = "Cases Sold")
```



```
# Creating a base dataframe w/ the given info
city <- 1:6
M_i <- c(52, 19, 37, 39, 8, 14)
m_i <- c(10, 4, 7, 8, 2, 3)

gg1 <- data.frame(city, M_i, m_i)
print(gg1)
```

```
##   city M_i m_i
## 1    1  52  10
## 2    2  19   4
## 3    3  37   7
## 4    4  39   8
## 5    5   8   2
## 6    6  14   3
```

```
# Specifying y_ij, the jth no. of cases sold per ith city
city1 <- c(146, 180, 251, 152, 72, 181, 171, 361, 73, 186)
city2 <- c(99, 101, 52, 121)
city3 <- c(199, 179, 98, 63, 126, 87, 62)
city4 <- c(226, 129, 57, 46, 86, 43, 85, 165)
city5 <- c(12, 23)
city6 <- c(87, 43, 59)
cities <- list(city1, city2, city3, city4, city5, city6)

# Updating the dataframe with summary statistics
gg1 <- gg1 %>%
  mutate(
    y_bar_i = sapply(cities, mean),
    s2_i = sapply(cities, var),
    t_hat_i = M_i / m_i *sapply(cities, sum)
  )

print(gg1)
```

```
##   city M_i m_i  y_bar_i      s2_i  t_hat_i
## 1    1  52  10 177.3000 6988.9000 9219.600
## 2    2  19   4  93.2500  854.9167 1771.750
## 3    3  37   7 116.2857 2974.5714 4302.571
## 4    4  39   8 104.6250 4172.2679 4080.375
## 5    5   8   2  17.5000   60.5000  140.000
## 6    6  14   3  63.0000  496.0000  882.000
```

From this, we can now compute an estimate for the total number of cases sold by plugging values into the following formula:

$$\begin{aligned}\hat{t}_{unb} &= \sum_{i\in S} \frac{N}{n}\hat{t}_i \\ &= \frac{45}{6}(9219.6 + 1771.750 + 4302.571429 + 4080.375 + 140 + 882) \\ &= \boxed{152972.2232}\end{aligned}$$

Now, to compute the standard error of this estimate, we need to compute for

$$\begin{aligned}SE(\hat{t}_{unb}) &= \sqrt{\hat{V}(\hat{t}_{unb})} \\ &= \sqrt{N^2(1-\frac{n}{N})\frac{s_t^2}{n} + \frac{N}{n}\sum_{i\in S}(1-\frac{m_i}{M_i})M_i^2\frac{s_i^2}{m_i}}\end{aligned}$$

where $s_t^2 = \frac{1}{n-1}\sum_{i\in S}\left(\hat{t}_i - \frac{\hat{t}_{unb}}{N}\right)^2$.

The values we are missing to compute $\hat{V}(\hat{t}_{unb})$ are s_t^2 and $\sum_{i\in S}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}$.

Computing for the term $(\hat{t}_i - \frac{\hat{t}_{unb}}{N})^2$ (which is needed for s_t^2) for the first city:

$$\begin{aligned}\left(\hat{t}_1 - \frac{\hat{t}_{unb}}{N}\right)^2 &= \left(9219.6 - \frac{152972.2}{45}\right)^2 \\ &= 33874934.98\end{aligned}$$

Computing the rest using R, we will label the column containing these values **"forvart".**

```
t_hat_unb <- 152972.2232
N <- 45

gg2 <- gg1 %>%
  mutate(
    forvart = (gg1$t_hat_i - t_hat_unb / N)^2
  ) %>%
  select(-M_i, -m_i, -y_bar_i, -s2_i, -t_hat_i)
print(gg2)
```

```
##   city    forvart
## 1    1 33874929.0
## 2    2  2649188.3
## 3    3   815749.8
## 4    4   463750.5
## 5    5 10623575.8
## 6    6  6337215.8
```

Summing up the values of the column "forvart", we will get

$$\sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}_{unb}}{N} \right)^2 = 54764409.26$$

and we can now compute for s_t^2 as

$$\begin{aligned} s_t^2 &= \frac{1}{n-1} \sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}_{unb}}{N} \right)^2 \\ &= \frac{1}{6-1}(54764409.26) \\ &= 10952881.85 \end{aligned}$$

Now, to compute for $\sum_{i \in S} \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{s_i^2}{m_i}$ (which is needed for $\hat{V}(\hat{t}_{unb})$), we will label the column containing the values of $\left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{s_i^2}{m_i}$ as **"forvarthat".**

```
gg2 <- gg2 %>%
  mutate(
    forvarthat = (1 - m_i / M_i) * M_i^2 * gg1$s2_i / m_i
  ) %>%
  select(-forvart)
print(gg2)
```

```
##   city forvarthat
## 1    1 1526375.76
## 2    2   60912.81
## 3    3  471682.04
## 4    4  630533.98
## 5    5    1452.00
## 6    6   25461.33
```

Summing up the values of the column "forvarthat", we will get

$$\sum_{i \in S} \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{s_i^2}{m_i} = 2716417.927$$

Thus, the standard error of $\hat{t}_{unb}$ is given by

$$\begin{aligned} SE(\hat{t}_{unb)} &= \sqrt{\hat{V}(\hat{t}_{unb})} \\ &= \sqrt{N^2(1 - \frac{n}{N})\frac{s_t^2}{n} + \frac{N}{n}\sum_{i \in S}(1 - \frac{m_i}{M_i})M_i^2 \frac{s_i^2}{m_i}} \\ &= \sqrt{45^2(1 - \frac{6}{45})\frac{10952881.85}{6} + \frac{45}{6}(2716417.927)} \\ &= \boxed{56781.0803} \end{aligned}$$

To estimate the average number of cases sold per supermarket, we use the following formula:

$$\begin{aligned}\hat{\bar{y}}_r &= \frac{\hat{t}_{unb}}{\hat{M}_o} \\ &= \frac{\hat{t}_{unb}}{\sum_{i\in S}\frac{N}{n}M_i} \\ &= \frac{152972.2232}{\frac{45}{6}(52+19+37+39+8+14)} \\ &= \boxed{120.6881}\end{aligned}$$

To obtain the standard error of $\hat{\bar{y}}_r$,

$$\begin{aligned}SE(\hat{\bar{y}}_r) &= \sqrt{\hat{V}(\hat{\bar{y}}_r)} \\ &= \sqrt{\frac{1}{\bar{M}^2}\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{1}{nN\bar{M}^2}\sum_{i\in S}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}},\end{aligned}$$

where $s_r^2 = \dfrac{1}{n-1}\displaystyle\sum_{i\in S}(M_i\bar{y}_i - M_i\hat{\bar{y}}_r)^2$.

The values we are missing to compute $\hat{V}(\hat{\bar{y}}_r)$ are $\bar{M}$ and s_r^2.

First,

$$\bar{M} = \frac{\sum_{i\in S} M_i}{n} = \frac{169}{6}$$

Next, to compute for $\sum_{i\in S}(M_i\bar{y}_i - M_i\hat{\bar{y}}_r)^2$ (which is needed for $\hat{V}(\hat{\bar{y}}_r)$), we will label the column containing the values of $(M_i\bar{y}_i - M_i\hat{\bar{y}}_r)^2$ as **"forvarybar".**

```
y_bar_hat_r <-
gg2 <- gg2 %>%
  mutate(
    forvarybar= (M_i * gg1$y_bar_i - M_i * 120.6881)^2
  ) %>%
  select(-forvarthat)
print(gg2)
```

```
##   city forvarybar
## 1    1 8666069.13
## 2    2  271778.61
## 3    3   26532.59
## 4    4  392453.26
## 5    5  681458.17
## 6    6  652271.71
```

Summing up the values of the column "forvarybar", we will get

$$\sum_{i\in S}(M_i\bar{y}_i - M_i\hat{\bar{y}}_r)^2 = 10690563.47$$

and we can now compute for s_r^2 as

$$\begin{aligned}s_r^2 &= \frac{1}{n-1}\sum_{i\in S}(M_i\bar{y}_i - M_i\hat{\bar{y}}_r)^2 \\ &= \frac{1}{6-1}(10690563.47) \\ &= 2138112.694\end{aligned}$$

Thus, the standard error of $\hat{\hat{y}}_r$ is given by

$$\begin{aligned}
SE(\hat{\hat{y}}_r) &= \sqrt{\hat{V}(\hat{\hat{y}}_r)} \\
&= \sqrt{\frac{1}{\left(\frac{169}{6}\right)^2}\left(1-\frac{6}{45}\right)\frac{2138112.694}{6}+\frac{1}{(6)(45)\left(\frac{169}{6}\right)^2}(2716417.927)} \\
&= \sqrt{401.9598426} \\
&= \boxed{20.0489}
\end{aligned}$$

Measles

The file measles.dat contains data consistent with that obtained in a survey of parents whose children had not been immunized for measles during a recent campaign to immunize all children between the ages of 11 and 15. During the campaign, 7633 children from the 46 schools in the area were immunized; 9962 children whose records showed no previous immunization were not immunized. In a follow-up survey to explore why the children had not been immunized during the campaign, Roberts et al. (1955) sent questionnaires to the parents of a cluster sample of the 9962 children. Ten schools were randomly selected, then a sample of the M_i nonimmunized children from each school was selected and the parents of those children were sent a questionnaire. Not all parents responded to the questionnaire.

(a) **Estimate, separately for each school, the percentage of parents who returned a consent form (variable *returnf*). For this exercise, treat the "no answer" responses (value 9) as not returned.**

```r
# Importing dataset
measles <- read_excel("measles.xlsx")
head(measles)
```

```
## # A tibble: 6 x 13
##   school  form returnf consent hadmeas previmm sideeff    gp noshot
##    <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>  <dbl>
## 1      1     1       1       0       1       0       0     0      0
## 2      1     1       1       0       1       1       0     0      0
## 3      1     1       1       0       1       0       0     0      0
## 4      1     1       0       0       1       1       0     0      0
## 5      1     1       1       0       0       0       1     0      0
## 6      1     0       9       9       0       0       0     0      0
## # i  more variables: Mitotal <dbl>, mi <dbl>, notser <dbl>, gpadv <dbl>
```

```r
# Updating the dataset
measles_cleaned <- measles %>%
  mutate(returned = ifelse(returnf == 1, 1, 0))

# Computing for the % of parents who returned a consent form per school
measles_returned <- measles_cleaned %>%
  group_by(school) %>%
  summarise(
    M_i = n(),
    t_i = sum(returned),
    percent = round(mean(returned) * 100, 4))
print(measles_returned)
```

```
## # A tibble: 10 x 4
##    school   M_i   t_i  percent
##      <dbl> <int> <dbl>   <dbl>
##  1       1    40    19    47.5
##  2       2    38    19    50
##  3       3    19    13    68.4
##  4       4    30    18    60
##  5       5    30    12    40
##  6       6    25    13    52
##  7       7    23    15    65.2
##  8       8    43    21    48.8
##  9       9    38    23    60.5
## 10      10    21     7    33.3
```

The rightmost column in the table directly above gives the percentage of parents who returned a consent form.

This was computed by multiplying the mean of y_{ij} (which is a proportion in this case) by 100, where y_{ij} is defined as follows:

$$y_{ij} = \begin{cases} 1 & \text{if parent of child } j \text{ in school } i \text{ returned the form} \\ 0 & \text{otherwise} \end{cases}$$

(b) **Using the number of respondents in school i as m_i, construct the sampling weight for each observation.**

```
N = 46
n = 10

#Calculating the sample weights for each observation
measles_weights <- measles %>%
  group_by(school) %>%
  summarise(
    m_i = mi,
    M_i = Mitotal,
    samp_weight = (N * M_i) / (n * m_i)
  )
print(measles_weights)
```

```
## # A tibble: 307 x 4
## # Groups:   school [10]
##     school   m_i   M_i samp_weight
##      <dbl> <dbl> <dbl>       <dbl>
##  1       1    40    78        8.97
##  2       1    40    78        8.97
##  3       1    40    78        8.97
##  4       1    40    78        8.97
##  5       1    40    78        8.97
##  6       1    40    78        8.97
##  7       1    40    78        8.97
##  8       1    40    78        8.97
##  9       1    40    78        8.97
## 10       1    40    78        8.97
## # i 297 more rows
```

The following is a compressed version of the previous table, where the sampling weights are grouped by school, in order to show all sampling weights,

```
# Compressing into 10 rows
weights_comp <- measles %>%
  group_by(school) %>%
  summarise(
    m_i = unique(mi),
    M_i = unique(Mitotal),
    samp_weight = (N * M_i) / (n * m_i)
  )
print(weights_comp)
```

```
## # A tibble: 10 x 4
##    school   m_i   M_i samp_weight
##     <dbl> <dbl> <dbl>       <dbl>
##  1      1    40    78        8.97
##  2      2    38   238       28.8
##  3      3    19   261       63.2
##  4      4    30   174       26.7
##  5      5    30   236       36.2
##  6      6    25   188       34.6
##  7      7    23   113       22.6
##  8      8    43   170       18.2
##  9      9    38   296       35.8
## 10     10    21   207       45.3
```

(c) **Estimate the overall percentage of parents who received a consent form along with a 95% CI.**

```
# Revising the dataset
measles_new <- measles %>%
  mutate(
    form = ifelse(form == 1, 1, 0),
    weight = measles_weights$samp_weight
  ) %>%
  select(-returnf, -consent, -hadmeas, -previmm, -sideeff, -gp, -noshot,
         -notser, -gpadv, -Mitotal, -mi)

# Defining the survey design
design <- svydesign(
  id = ~school,
  weights = ~weight,
  data = measles_new
)

# Estimating overall percentage of parents who received the form (with 95%
  CI)
cluster_prop <- svymean(~form, design)
cluster_prop1 <- as.data.frame(cluster_prop)
cluster_prop_ci <- confint(cluster_prop, level = 0.95)
cat("Estimate: ", round(cluster_prop*100, 4), "%", "\n")
cat("95% CI: ", round(cluster_prop_ci, 4))
```

```
## Estimate:  90.9092 %
## 95% CI:  0.8731 0.945
```

(d) **How do your estimate and interval in part (c) compare with the results you would have obtained if you had ignored the clustering and analyzed the data as an SRS?**

Find the ratio:

$$\frac{\textbf{estimated variance from (c)}}{\textbf{estimated variance if the data were analyzed as an SRS}}$$

What is the effect of clustering?

```
n = nrow(measles_new)
N = 9962

# Computing the estimated variance using cluster sampling
cluster_v_hat <- (cluster_prop1$form)^2 # 'SE' has been renamed to 'form'
                                         in the cluster_prop dataframe

# Computing the estimated variance using SRSWOR
srs_prop <- sum(measles_new$form)/n
srs_v_hat <- (1 - n/N)*(srs_prop*(1 - srs_prop))/(n - 1)

# Computing the ratio
ratio = cluster_v_hat / srs_v_hat

cat("Ratio: ", round(ratio, 4))
```

```
## Ratio:  1.3706
```

Since the ratio is greater than 1, this means that the estimated variance computed under cluster sampling is greater than the estimated variance computed under SRSWOR. In general, while it is often more convenient and less costly, **cluster sampling generates estimators with a greater variance than estimators under SRSWOR**.