# EDA A3 Q2: Finite populations and simple random sampling

To demonstrate results numerically, define an example population and its variate values for $N = 1000$ as follows

```
set.seed(314159)
N <- 1000
y <- rchisq(N, df = 5)
#y
g <- sample(c("A", "B"), size = N, replace = TRUE)
#g
pop <- 1:N
A <- g == "A"
B <- g == "B"
N_A <- sum(A)
N_B <- sum(B)
data <- data.frame(u = pop, y = y, g = g)
#data
```

a. *(2 marks)* Write down how $\mu_y$ can be determined mathematically from $\mu_A$ and $\mu_B$. In R demonstrate this holds for the population values given in 'data' above. Show your code.

Let $N_A$ be the number of elements in group A and $N_B$ be the number of elements in group B.

$\mu_y = \frac{N_A}{N} * \mu_A + \frac{N_B}{N} * \mu_B$

```
mean(data$y)
```

```
## [1] 4.891386
```

```
ua<-mean(data$y[g=="A"])
ub<-mean(data$y[g=="B"])
(N_A/N)*ua + (N_B/N)*ub
```

```
## [1] 4.891386
```

As we can see, both values match.

b. *(12 marks)* Show mathematically how $\sigma_y^2$ can be calculated from $\sigma_A^2$ and $\sigma_B^2$, the difference in the group averages $(\bar{y}_A - \bar{y}_B)$, and the known group sizes $N_A$ and $N_B$.

Demonstrate numerically that the derived formula holds by applying it to the population values given in 'data' above. Show your code.

$\sigma_y^2 = \frac{1}{N} \sum_{k=A,B} \sum_{j=1}^{N_k} (y_{k,j} - \mu)^2$

$= \frac{1}{N} \sum_{k=A,B} \sum_{j=1}^{N_k} (y_{k,j} - \mu_k + \mu_k - \mu_y)^2$

$= \frac{1}{N} \sum_{k=A,B} \sum_{j=1}^{N_k} ((y_{k,j} - \mu_k) + (\mu_k - \mu))^2$

$= \frac{1}{N} \sum_{k=A,B} \sum_{j=1}^{N_k} ((y_{k,j} - \mu_k)^2 + (\mu_k - \mu_y)^2 + 2(y_{k,j} - \mu_k)(\mu_k - \mu_y))$

Since $\sum_{j=1}^{N_k} (y_{k,j} - \mu_k) = 0$

$\sigma_y^2 = \frac{1}{N} \sum_{k=A,B} \sum_{j=1}^{N_k} ((y_{k,j} - \mu_k)^2 + (\mu_k - \mu_y)^2)$

$= \frac{1}{N} \sum_{k=A,B} ((N_k - 1)\sigma_k^2 + N_k(\mu_k - \mu_y)^2)$

Here we see:

$\mu_A - \mu_y = \mu_A - (\frac{N_A}{N}\mu_A + \frac{N_B}{N}\mu_B)$

$= \frac{N_B}{N}(\mu_A - \mu_B)$

Similarly

$\mu_B - \mu_y = \frac{N_A}{N}(\mu_A - \mu_B)$

Substituting them gives the following equation:

$\frac{N_A\sigma_A^2 + N_B\sigma_B^2}{N_A+N_B} + \frac{N_A N_B(\mu_A-\mu_B)^2}{(N_A+N_B)^2}$

```
var(data$y) * (N-1)/N
```

```
## [1] 10.10849
```

```
u<-mean(data$y)
#sum(data$y)
#sum((data$y-u)^2)/N
#sum(c(6,7,8)-5)
ua<-mean(data$y[g=="A"])
ub<-mean(data$y[g=="B"])
va<-var(data$y[g=="A"])* (N_A-1)/N_A
vb<-var(data$y[g=="B"])* (N_B-1)/N_B
(N_A/N)*va + (N_B/N)*vb+((N_A*N_B*((ua-ub)^2)))/(N^2)
```

```
## [1] 10.10849
```

As we can see, both values match.

c. Simple random sampling (without replacement)

d. *(4 marks)* Prove that $\tilde{\mu}_y$ is **unbiased** for $\mu_y$.

$\tilde{\mu}_y$ will be unbiased if $E(\tilde{\mu}_y) = \mu_y$

$E(\tilde{\mu}_y) = E(\frac{1}{n}\Sigma_{u\in P} y_u)$

$= \frac{1}{n}\Sigma E(y_u)$

$= \frac{1}{n}\Sigma \mu_y$

$= \frac{1}{n} n\mu_y$

$= \mu_y$

Since $E(\tilde{\mu}_y) = \mu_y$ , $\tilde{\mu}_y$ is unbiased.

ii. *(10 marks)* Prove that

$E(\tilde{\sigma}^2_{n-1}) = \frac{1}{N-1}\sum_{u\in P}(y_u - \widehat{\mu}_y)^2$

and hence that $\tilde{\sigma}^2_{n-1}$ is **biased** for the finite population variance $\sigma^2_y$.

$\widehat{\mu}_y^2 = \left\{\frac{1}{n}\sum_{u=1}^n y_u\right\}^2 = \frac{1}{n^2}(\sum_{u=1}^n y_u^2 + \sum_{u\neq v} y_u y_v)$

$\tilde{\sigma}^2_{n-1} = \frac{1}{n-1}\sum_{u\in P}(y_u - \widehat{\mu}_y)^2$

$= \frac{1}{n-1}\sum_{u=1}^n (y_u^2 + \widehat{\mu}_y^2 - 2y_u\widehat{\mu}_y)$

$= \frac{1}{n-1}(\sum_{u=1}^n y_u^2 - n\widehat{\mu}_y)$

replacing value of $\widehat{\mu}_y$

$= \frac{1}{n-1}\{\sum_{u=1}^n y_u^2 - \frac{1}{n}(\sum_{u=1}^n y_u^2 + \sum_{u\neq v} y_u y_v)\}$

$= \frac{1}{n(n-1)}\left\{(n-1)\sum_{u=1}^{n} y_u^2 - \sum_{u\neq v} y_u y_v\right\}$

$= \frac{1}{2n(n-1)}\left\{(n-1)\sum_{u=1}^{n} y_u^2 + (n-1)\sum_{v=1}^{n} y_v^2 - 2\sum_{u\neq v} y_u y_v\right\}$

$= \frac{1}{n(n-1)}\sum_{u\neq v}\frac{(y_u-y_v)^2}{2}$

using this form of variance, we will calculate $E(\tilde{\sigma}^2_{n-1})$.

$E(\tilde{\sigma}^2_{n-1}) = \frac{1}{n(n-1)}\sum_{u\neq v} E\left(\frac{(y_u-y_v)^2}{2}\right)$

$= \frac{1}{n(n-1)}\sum_{u\neq v}\left(\frac{1}{2}2E(y_u^2) - 2E(y_u y_v)\right)$

$= \frac{1}{n(n-1)}n(n-1)\big(E(y_u^2) - E(y_u y_v)\big)$

$= E(y_u^2) - E(y_u y_v)$

$= \sigma_y^2 - Cov(y_u, y_v)$

since it is sampling with replacement, all pairs of $y_u s$ have covariance of $\frac{-\sigma_y^2}{N-1}$

so replacing its value gives us:

$= \sigma_y^2 - \frac{-\sigma_y^2}{N-1}$

$= \frac{N}{N-1}\sigma_y^2$

$= \frac{1}{N-1}\sum_{u\in P}(y_u - \widehat{\mu}_y)^2$

iii. *(2 marks)* Show how $\tilde{\sigma}^2_{n-1}$ can be corrected to become **unbiased** for the finite population variance $\sigma_y^2$. What happens to this correction as $N \to \infty$?

In order to be unbiased, we can multiply $\tilde{\sigma}^2_{n-1}$ by a factor of $\frac{(N-1)}{N}$ to make it unbiased and $E(\tilde{\sigma}^2_{n-1}) = \sigma_y^2$

When $N \to \infty$, the closer the estimated variance will be to the true variance.