Selecting Samples

Contents


```
### Units in the large population of all encounters
popSharks <- rownames(sharks)
### get the sub-population that is just those encounters in Australian waters
popSharksAustralia <- popSharks[sharks$Australia == 1]
### the units in the sub-population are
```

```
samples <- combn(popSharksAustralia, 5)
N_s <- ncol(samples)
```

3.2 Selecting samples

- For any particular sample,
  - the attribute calculated based on the sample identical to the population attribute or
  - it might be so different we would be completely misled about the true nature of the population attribute from the sample attribute.
- This is why it is important to understand **how** the sample is selected, and if it is within our power to do so to have a hand in selecting the sample itself.
  - Even when the latter is possible, enormous care must be taken so that our own prejudices and pre-conceptions about the population do not render a sample that is misleading.

Population of Samples

- Consider the population $\mathcal{P}_{\mathcal{S}}$ of M samples each of size n.

$$\mathcal{P}_{\mathcal{S}} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_M\}$$

- Any attribute $a(\mathcal{S}_i)$ is now just a variate on that unit!
  - We then have a population of attributes

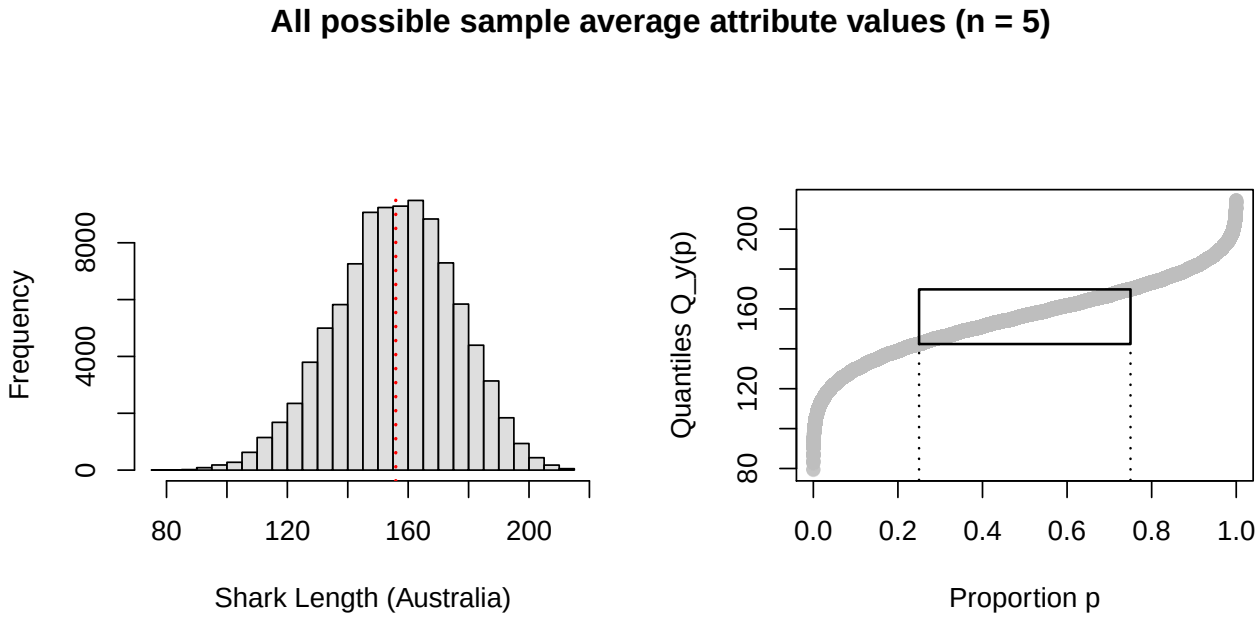
$$\mathcal{P}_{a(\mathcal{S})} = \{a(\mathcal{S}_1), a(\mathcal{S}_2), \ldots, a(\mathcal{S}_M)\}$$

- If we select our sample from $\mathcal{P}_{\mathcal{S}}$ with probability $\frac{1}{M}$ then the histogram shows the distribution for the variate values $a(\mathcal{S})$.

```
avePop <- mean(sharks[popSharksAustralia, "Length"])
avesSamp <- apply(samples, MARGIN = 2,
                  FUN = function(s){mean(sharks[s,"Length"])})
sampleErrors <- avesSamp - avePop
```

```
par(mfrow=c(1,2),oma=c(0,0,2,0))

hist(avesSamp, col=adjustcolor("grey", alpha = 0.5),
     main="",
     xlab="Shark Length (Australia)",
     breaks=25
     )
### Mark the population attribute in red
abline(v=avePop, col="red", lty=3, lwd=2)

qvals <- sort(avesSamp)
pvals <- ppoints(length(avesSamp))
plot(pvals, qvals, pch = 19, col=adjustcolor("grey", alpha = 0.5),
     xlim=c(0,1),
     xlab = "Proportion p",
```

```
     ylab = "Quantiles Q_y(p)",
     main = "")

y.plot = quantile(avesSamp,c(.25,.75,.75,.25,.25))
lines(y.plot~c(.25,.25,.75,.75,.25),lwd=1.5)
lines(c(0.25,0.25),c(0,y.plot[1]),lty=3,lwd=1.5)
lines(c(0.75,0.75),c(0,y.plot[1]),lty=3,lwd=1.5)


title("All possible sample average attribute values (n = 5)", outer=TRUE)
```



Good News!

- This means that because we **randomly selecting a sample** from $\mathcal{P}_{\mathcal{S}}$ we are able to make probability statements regarding the attribute $a(\mathcal{S})$ taking on any value.
  - If $n = 5$, we know that with probability $\frac{1}{2}$ the attribute that results will be within the range $[142.4, 169.8]$ inches, (IQR). i.e.

$$\Pr\left(\, a(\mathcal{S}) \in [142.4, 169.8] \, \right) = \frac{1}{2}$$

    because we are selecting $\mathcal{S}$ from $\mathcal{P}_{\mathcal{S}}$ with probability $p(\mathcal{S}) = \frac{1}{M}$.
  - We can read off many other probabilities about $a(\mathcal{S})$ from the histogram or the quantile plot.

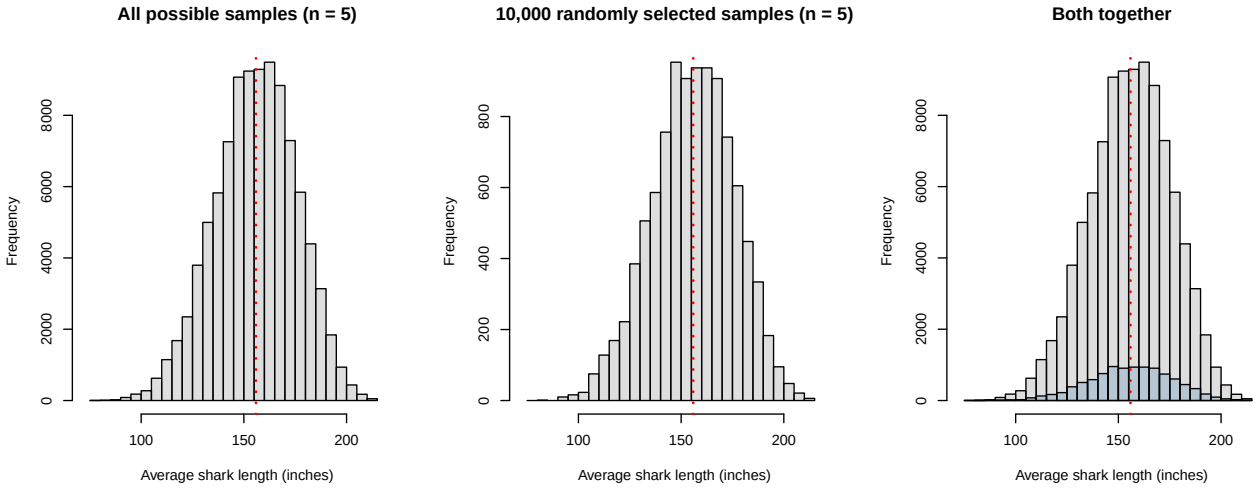


Figure 1: All versus 10,000 randomly selected samples (n = 5)

Randomly selecting m Samples

- In general, it computationally expensive to calculate all possible samples.
- Instead, suppose we draw a sample of m = 10,000 samples $\mathcal{S}_{u_1}, \ldots, \mathcal{S}_{u_m}$ from $\mathcal{P}_{\mathcal{S}}$ of $\binom{N}{n} = \binom{28}{5} =$ 98,280 possible samples.

Exercise: Regenerate the plots above. The argument `add=TRUE` in the `hist` function will be handy.

Distribution of a Histogram

- Suppose the histograms have K bins

$$B_1 = (b_0, b_1], B_2 = (b_1, b_2], \ldots, B_K = (b_{K-1}, b_K]$$

  and
  - the kth bin B_k contains $M_k \geq 0$ of the attribute values $a(S_i)$ $i = 1, \ldots, M$.
  - The bins contain the attribute values of all of the $\mathcal{S}_i \in \mathcal{P}_{\mathcal{S}}$ so that $\sum_{k=1}^{K} M_k = M$.
  - Let m_k be the number of the m selected samples whose attribute value falls in B_k, with $m = \sum_{k=1}^{K} m_k$.

- With this notation,
  - the histogram using all the data has heights $M_1, \ldots, M_K$ and
  - the sampled histogram has heights $m_1, \ldots, m_K$.

- What is the distribution of the sampled heights?

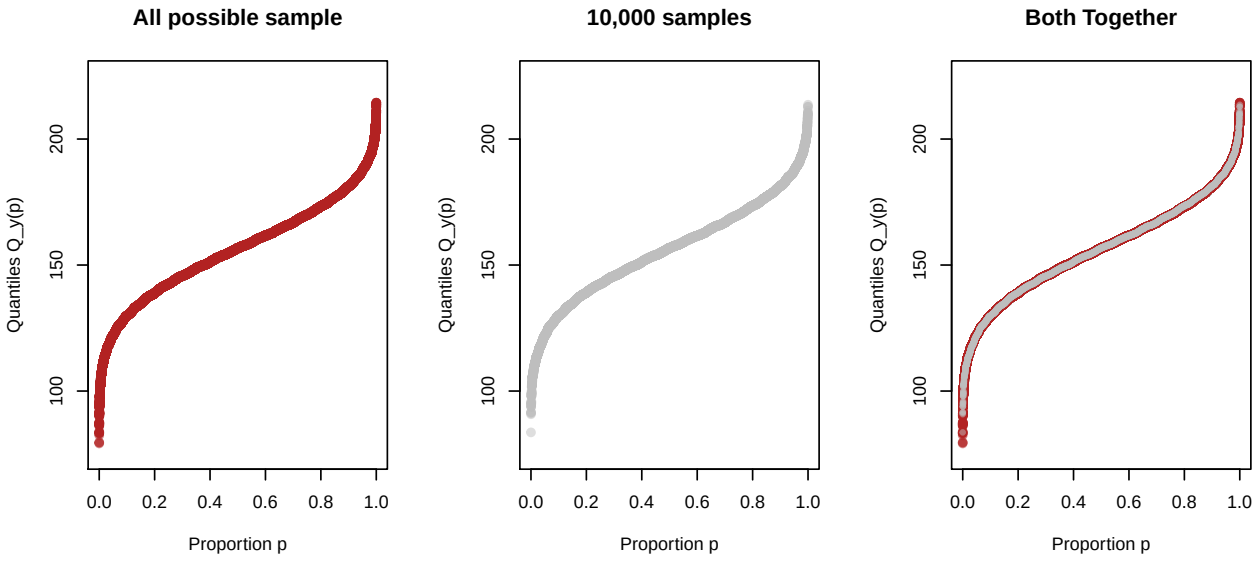


Figure 2: All possible sample average attribute values (n = 5)

Quantile Plot

- What do you learn from these quantile plots?

3.2.1 Sampling

- We select a sample $\mathcal{S}$ from the population $\mathcal{P}_{\mathcal{S}}$ of size M containing all available samples,
  - according to some probability $p(\mathcal{S}) \geq 0$ of being selected. We require of course that

$$\sum_{\mathcal{S} \in \mathcal{P}_{\mathcal{S}}} p(\mathcal{S}) = 1.$$

- For any sample $\mathcal{S} \in \mathcal{P}_{\mathcal{S}}$, we have its **sample error**

$$\text{Sample Error} = a(\mathcal{S}) - a(\mathcal{P}).$$

  - For any collection of samples (or population of samples) $\mathcal{P}_{\mathcal{S}}$ containing M samples, we have the **average sample error**

$$\text{Average Sample Error} \quad = \frac{1}{M} \sum_{\mathcal{S} \in \mathcal{P}_{\mathcal{S}}} \left(a(\mathcal{S}) - a(\mathcal{P}) \right).$$

- By sampling $\mathcal{S}$ randomly from $\mathcal{P}_{\mathcal{S}}$, we also have the **sampling bias**

$$\begin{aligned} \text{Sampling Bias} &= E(a(\mathcal{S})) - a(\mathcal{P}) \\ &= \textstyle\sum_{\mathcal{S}\in\mathcal{P}_{\mathcal{S}}} a(\mathcal{S})p(\mathcal{S}) - a(\mathcal{P}) \\ &= \textstyle\sum_{\mathcal{S}\in\mathcal{P}_{\mathcal{S}}} \left(a(\mathcal{S}) - a(\mathcal{P})\right) p(\mathcal{S}) \end{aligned}$$

Sampling bias is just the **expected** sample error induced by the repeated random sampling of $\mathcal{S}$ from $\mathcal{P}_{\mathcal{S}}$. If $p(\mathcal{S}) = \frac{1}{M}$, the sampling bias is identical to the average sample error of $a(\mathcal{P})$.

- The sampling bias depends on the attribute $a(\cdot)$, the set of possible samples $\mathcal{P}_{\mathcal{S}}$, and the sample probabilities $p(\mathcal{S})$.
  - **Note:** If sampling bias is 0, then $a(\mathcal{S})$ is called an **unbiased** estimator of $a(\mathcal{P})$.

Sampling Variance

- We could similarly define other characteristics of the sampling such as the **sampling variance**

$$Var\left(a(\mathcal{S})\right) = E\left(\left[\, a(\mathcal{S}) - E(a(\mathcal{S}))\, \right]^2 \right)$$

  - where all expectations are taken with respect to the probabilities $p(\mathcal{S})$ of the samples $\mathcal{S}$ from $\mathcal{P}_{\mathcal{S}}$.
- Given a sample $\mathcal{S}$, we would like $a(\mathcal{S})$ and $a(\mathcal{P})$ to be as close as possible, hence we look at the expected distance between these two quantities, i.e.

$$\begin{aligned} MSE\left(a(\mathcal{S})\right) &= E\left([a(\mathcal{S}) - a(\mathcal{P})]^2\right) \\ &= Var\left(a(\mathcal{S})\right) + [\text{Sampling Bias}]^2 \end{aligned}$$

- Ideally, we would like to choose $p(\mathcal{S})$ and/or $\mathcal{P}_{\mathcal{S}}$, so that both the square of sampling bias and the sampling variance are as small as possible.

Attribute as a Random Variable

- We can introduce a **random variate**, say A, that takes values a from the distinct values of $a(\mathcal{S})$ for all $\mathcal{S} \in \mathcal{P}_{\mathcal{S}}$. The induced probability distribution has

$$Pr(A = a) = \sum_{\mathcal{S}\in\mathcal{P}_{\mathcal{S}}} p(\mathcal{S}) \times I_{\{a\}}(a(\mathcal{S}))$$

  where $I_X(x)$ is the usual indicator function defined for any x and set X as

$$I_X(x) = \begin{cases} 1 & \text{if } \; x \in X \\ 0 & \text{otherwise.} \end{cases}$$

- It turns out that A is a discrete random variate.
- Probability statements about its values can be made using its distribution, including its expectation, variance, etc.

Exercise: If there are only $K \leq M$ distinct values, say $a_1, \ldots, a_K$ (M is the total number of possible samples defined above), then show that A, as defined above, is a discrete random variate with probabilities $Pr(A = a_i)$. Express the sampling bias and the sampling variance in terms of this random variate.

Example

In this example we illustrate two different sampling designs.

- Suppose that the population consists of five units

```
set.seed(341)
pop5 = round(rnorm(5),2)
pop5 = sort(pop5)
pop5 # our population
```

```
## [1] -1.06 -0.99 -0.31  0.83  0.87
```

All possible samples of size 2.

```
sam2 = combn(5,2)
colnames(sam2) = paste("S", 1:10, sep="")
sam2
```

```
##      S1 S2 S3 S4 S5 S6 S7 S8 S9 S10
## [1,]  1  1  1  1  2  2  2  3  3   4
## [2,]  2  3  4  5  3  4  5  4  5   5
```

Note that the elements of the matrix above are the indices of the samples, not the actual values of the units. Next, we calculate the attribute (the average) on theses samples.

```
sam.avg <- apply(sam2, MARGIN = 2, FUN = function(s){mean(pop5[s])})
round(sam.avg,3)
```

```
##     S1     S2     S3     S4     S5     S6     S7     S8     S9    S10
## -1.025 -0.685 -0.115 -0.095 -0.650 -0.080 -0.060  0.260  0.280  0.850
```

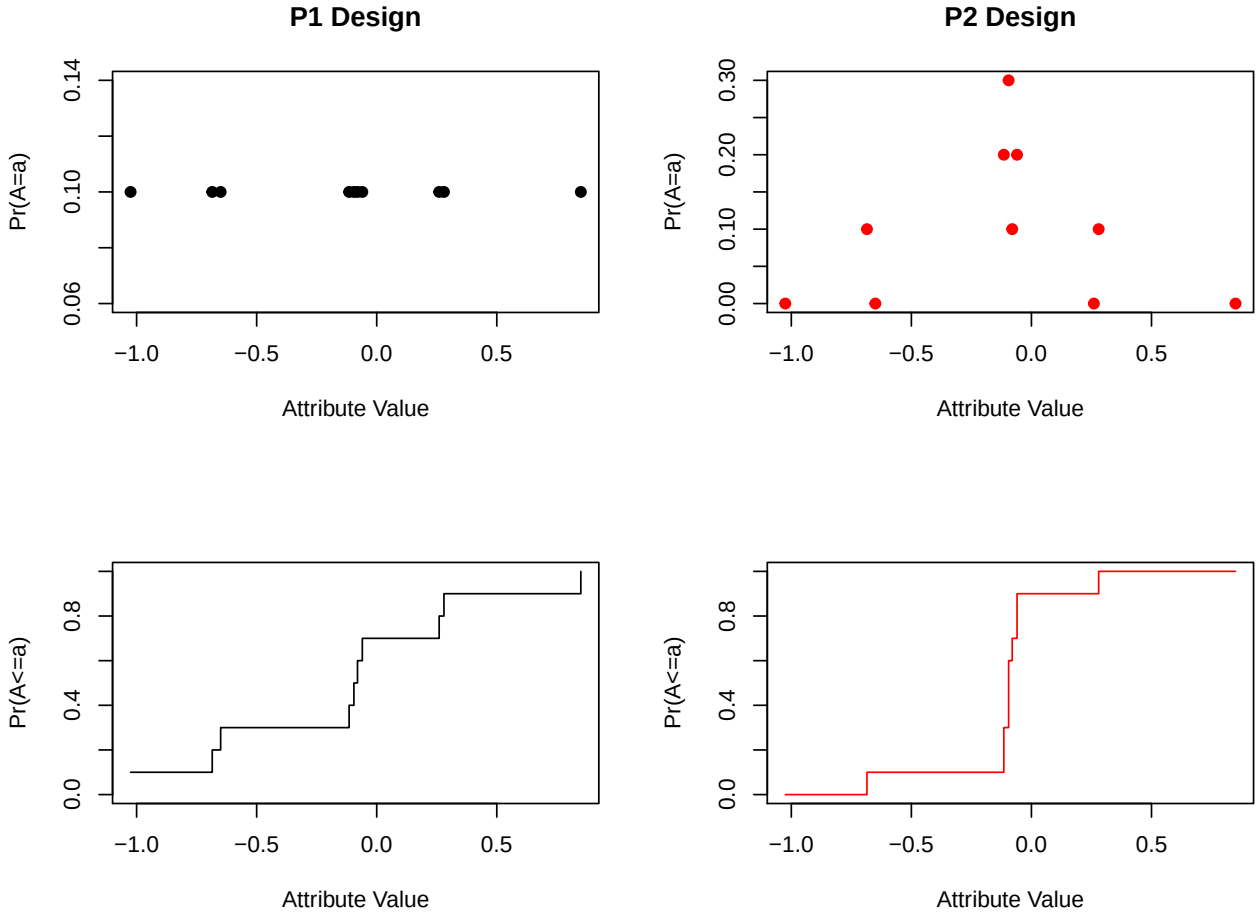
- Now create two sampling designs
  - `d1` assigns same probability to the 10 possible samples (1/10 each).
  - `d2` is a *biased* design: there is, really, no intuition behind this design, so simply assume that each sample of size 2 is chosen with probabilities `d2`

```
d1 = rep(1/10,10)
d2 = 2*(abs(apply(sam2, 2, diff))-1)
d2 = d2/sum(d2)
designs = rbind(d1,d2)
colnames(designs) = paste('S', 1:10, sep="")
round( designs,2)
```

```
##      S1  S2  S3  S4  S5  S6  S7  S8  S9 S10
## d1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1
## d2 0.0 0.1 0.2 0.3 0.0 0.1 0.2 0.0 0.1 0.0
```

- The distribution of the attribute (here, the mean) induced by the sampling design, i.e. $P(A = a_i)$:

```
avg.ord = order(sam.avg)

par(mfrow=c(2,2),oma=c(0,0,0,0))
plot(sam.avg[avg.ord], d1[avg.ord], xlab="Attribute Value",
     ylab="Pr(A=a)", pch=19,main="P1 Design")
plot(sam.avg[avg.ord], d2[avg.ord], xlab="Attribute Value",
     ylab="Pr(A=a)", pch=19,col=2,main="P2 Design")

plot(sam.avg[avg.ord], cumsum(d1[avg.ord]), xlab="Attribute Value",
     ylab="Pr(A<=a)", pch=19, type='s',ylim=c(0,1))
plot(sam.avg[avg.ord], cumsum(d2[avg.ord]), xlab="Attribute Value",
     ylab="Pr(A<=a)", pch=19,col=2, type='s',ylim=c(0,1))
```



- Note that the distribution of the attribute with respect to design `d2` is more concentrated.

Mean Square Error

- We can compare these two sampling designs `d1` and `d2` numerically using the Sampling Mean Square Error (MSE)

$$\begin{aligned} \text{Sampling MSE} &= \text{Sampling Variance} + (\text{Sampling Bias})^2 \\ &= Var\left[a(\mathcal{S})\right] + \left(E\left[a(\mathcal{S}) - a(\mathcal{P})\right]\right)^2 \end{aligned}$$

Sampling bias

```
exp1 = sum(sam.avg*d1)
exp2 = sum(sam.avg*d2)

sam.bias=  c(exp1, exp2) - mean(pop5)
round(sam.bias,3)
```

```
## [1] 0.00 0.02
```

Sampling Variance

```
sam.var = c( sum( (sam.avg-exp1)^2 * d1 ), sum( (sam.avg- exp2)^2*d2 ) )
round(sam.var,3)
```

```
## [1] 0.267 0.049
```

Sampling MSE

```
designs.MSE = rbind( sam.bias, sam.var, MSE=sam.var + sam.bias^2)
colnames(designs.MSE) = c("d1", "d2")
round(designs.MSE,5)
```

```
##                d1      d2
## sam.bias 0.00000 0.02000
## sam.var  0.26689 0.04893
## MSE      0.26689 0.04933
```

Alternatively, we could use the formula

$$MSE\left(a(\mathcal{S})\right) = E\left(\left[a(\mathcal{S}) - a(\mathcal{P})\right]^2\right)$$

to calculate the MSE, i.e.

```
MSE = c( sum( (sam.avg-mean(pop5))^2 * d1 ), sum( (sam.avg- mean(pop5))^2*d2 ) )
round(MSE,3)
```

```
## [1] 0.267 0.049
```

Note: Although the `d2` scheme is biased, it has a lower sampling MSE.

3.2.2 Large Populations

- As the population size increases, constructing all possible samples becomes prohibitive
  - Rather than constructing all possible samples, we might randomly select m samples.

- For example, consider the agricultural census of US counties whose population consists of only $N = 3078$ counties.
  - For $n = 100$, there are $\binom{3078}{100}$ or about 1.4×10^{190} possible samples.

- The combinatorial explosion is avoided if we examine only m, say $m = 10{,}000$, samples.
  - Unfortunately, if we have to enumerate all possible samples just to select from them we are no farther ahead (scheme $d2$ in the example above).

3.2.3 Sampling mechanisms

- Rather than constructing a probability measure on all possible samples, each unit u in a sample $\mathcal{S}$ is selected one at a time from the population $\mathcal{P}$.
- A *sequence* of the first k units u_i selected from $\mathcal{P}$ is

$$s_k = (u_{i_1}, u_{i_2}, \ldots, u_{i_k})$$

- A **sampling mechanism** is defined by the probabilities

$$\Pr(u \mid k, s_{k-1}) \quad \text{and} \quad \Pr(u).$$

  - The first unit is selected with probability $\Pr(u)$
  - and the probability of the sequence of the first k units selected is

$$\Pr(s_k) = \Pr(u_{i_1}) \times \Pr(u_{i_2} \mid 2, s_1) \times \Pr(u_{i_3} \mid 3, s_2) \times \cdots \times \Pr(u_{i_k} \mid k-1, s_{k-1}).$$

- To determine $p(\mathcal{S})$ from a sampling mechanism,
  - the order in which the units appear does not matter, i.e. any permutation of the elements of s_n counts as $\mathcal{S}$
  - $p(\mathcal{S})$ is simply the sum of $\Pr(s_n)$ over all permutations s_n.

Simple Random Sampling without Replacement

- The **sampling mechanism** is

$$\Pr(u) = \frac{1}{N} \quad \text{and} \quad \Pr(u \mid k, s_{k-1}) = \frac{1}{N-k+1}$$

  - The probability of a sequence is

$$\Pr(s_n) = \frac{1}{N} \times \frac{1}{N-1} \times \frac{1}{N-2} \times \cdots \times \frac{1}{N-n+1}$$

– This is the same for all $n!$ permutations, so the

$$p(\mathcal{S}) = \frac{n!}{N(N-1)(N-2)\cdots(N-n+1)} = \frac{1}{\binom{N}{n}},$$

– This probability is the same we had before for selecting n distinct units from a population of N distinct units.

- However, we now have a mechanism that allows us to select a sample *without first enumerating* all $\binom{M=N}{n}$ possible samples in $\mathcal{P}_{\mathcal{S}}$.
- In `R`, the indices of a simple random sample without replacement of size n from indices $1, \ldots, N$ is returned from the function call `sample(N,n)`.

```
set.seed(341)
pop10 = round(rnorm(10),2)
pop10
```

```
##  [1] -1.06 -0.31  0.87 -0.99  0.83  0.47 -0.66 -0.05  1.46 -0.72
```

```
set.seed(341)
sample2 = sample(10, 2)
sample2
```

```
## [1] 2 9
```

```
pop10[sample2]
```

```
## [1] -0.31  1.46
```

- If rather than indices, the units were identified by the (assumed unique) contents of a vector `Pop`, then `sample(Pop,n)` would return the vector of units in the sample.

```
set.seed(341)
sample(pop10, 2)
```

```
## [1] -0.31  1.46
```

Simple Random Sampling with Replacement

- The **sampling mechanism** is

$$\Pr(u) = \frac{1}{N} = \Pr(u \mid k, s_{k-1})$$

and a sample, $\mathcal{S}$, can have one or more units repeated in the sample.

- Using the equation above in

$$\Pr(s_k) = \Pr(u_{i_1}) \times \Pr(u_{i_2} \mid 2, s_1) \times \Pr(u_{i_3} \mid 3, s_2) \times \cdots \times \Pr(u_{i_k} \mid k-1, s_{k-1}).$$

we get

$$p(\mathcal{S}) = \frac{1}{N^n}$$

where the population of all samples $\mathcal{P}_{\mathcal{S}}$ contains $M = N^n$ different samples.

- To generate simple random samples with replacement in `R`, the previous calls are adjusted to include the argument `replace = TRUE` as in `sample(N, m, replace = TRUE)`.

```
set.seed(341)
pop10 = round(rnorm(10),3)

set.seed(341)
sample5 = sample(10, 5, replace=TRUE)
sample5
```

```
## [1] 2 9 4 1 9
```

```
pop10[sample5]
```

```
## [1] -0.308  1.462 -0.993 -1.060  1.462
```

Comparing Sampling Mechanisms

- For a population of size N
  - there exist $\binom{N}{n}$ samples **without replacement**
  - there exist N^n samples **with replacement**

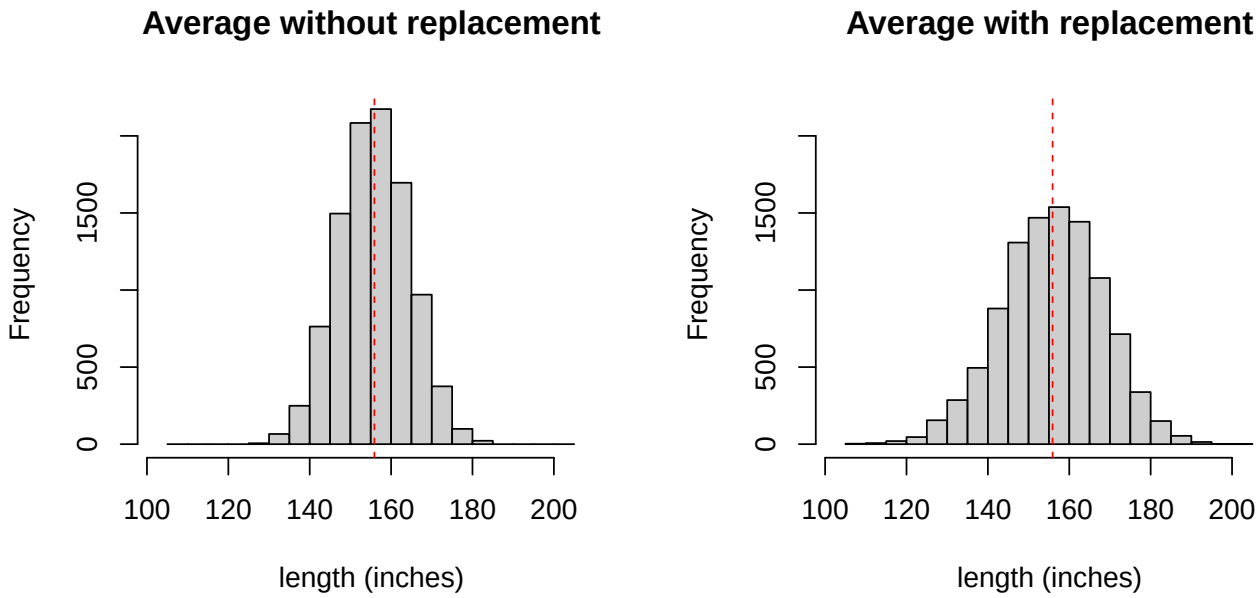
- Using the Australian shark encounter population, if we take $n = 15$ samples,
  1. sampling without replacement yields a population $\mathcal{P}_{\mathcal{S}}$ of size $M = \binom{28}{15} = 37{,}442{,}160$.
  2. for sampling with replacement, $\mathcal{P}_{\mathcal{S}}$ is much larger, containing $M = 28^{15} = 5.097655 \times 10^{21} = 5{,}097{,}655{,}000{,}000{,}000{,}000{,}000$ different possibilities.

- Using each mechanism we construct $m = 10,000$ samples and for each sample calculate the average (R codes in the notes).

```
directory = "../Data"
dirsep= "/"
sharkfile <- paste(directory, "Sharks", "sharks.csv", sep=dirsep)
sharks <- read.csv(sharkfile)

popSharks <- rownames(sharks)
popSharksAustralia <- popSharks[sharks$Australia == 1]
avePop <- mean(sharks[popSharksAustralia, "Length"])
```

```
### sample size
n <- 15
### number of samples
m <- 10000

### reproducibiity
set.seed(123415)
```

```
### samples without replacement
sampsWithout <- Map(function(i){sample(popSharksAustralia, size=n, replace = FALSE)},
                    1:m)
### attribute evaluated on each sample
aveWithout <- Map(function(s){mean(sharks[s,"Length"])}, sampsWithout)


### samples with replacement
sampsWith <- Map(function(i){sample(popSharksAustralia, size=n, replace = TRUE)},
                 1:m)
### attribute evaluated on each sample
aveWith <- Map(function(s){mean(sharks[s,"Length"])}, sampsWith)

### Note that in both cases, there are so many samples to choose from
### that we are not going to worry about whether we have repeated any
### in the m we have selected from M
###

### Now prepare to plot histograms
###
### Use the same x scale in the plots
xlim <- extendrange(c(aveWith, aveWithout))
### and bins
bins <- hist(c(as.numeric(aveWithout), as.numeric(aveWith)),
             breaks = 30, plot=FALSE)
### And heights
ylim <- c(0, 2200)

### Without replacement
###
par(mfrow=c(1,2))
hist(as.numeric(aveWithout), main = "Average without replacement",
     xlim = xlim, xlab = "length (inches)", ylim = ylim,
     breaks = bins$breaks, col =  adjustcolor("grey", 0.75))
abline(v=avePop, col="red", lty=2)

### and with
hist(as.numeric(aveWith), main = "Average with replacement",
     xlim = xlim, xlab = "length (inches)", ylim = ylim,
     breaks = bins$breaks, col =  adjustcolor("grey", 0.75))
abline(v=avePop, col="red", lty=2)
```



- **Comment**
  - Simple random sampling without replacement produces a more concentrated histogram.
  - Numerical summary using a five number summary

```
##                       Min 1st Qu. Median 3rd Qu.   Max
## Without Replacement 128.0   149.8  155.7   161.7 183.6
## With Replacement    106.8   147.5  156.0   164.4 203.6
```

3.2.3.1 A curious sampling mechanism

- The following mechanism was first explored by Basu (1958).
- Suppose we perform simple random sampling with replacement except that we *remove* any duplicate units.
  - The samples produced will have sizes anywhere from 1 to n according to how many distinct units were selected in a sample (sampling with replacement).

```
set.seed(341)
pop10 = round(rnorm(10),3)
set.seed(341)
sample5 = sample(10, 5, replace=TRUE)
sample5
```

```
## [1] 2 9 4 1 9
```

- Simple random sample with replacement yields

```
pop10[sample5]
```

```
## [1] -0.308  1.462 -0.993 -1.060  1.462
```

- Simple random sample with replacement removing duplicate units yields

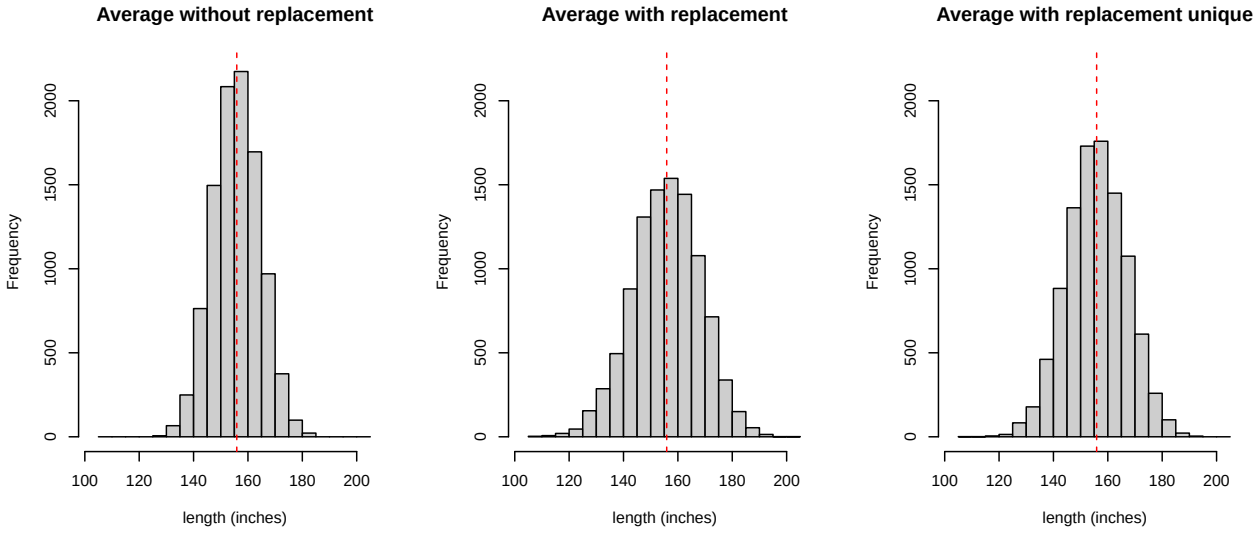
```
unique(pop10[sample5])
```

```
## [1] -0.308  1.462 -0.993 -1.060
```

Note that since the number of duplicates is a random variable, the actual sample size (n minus the number of duplicates) is also a random variable here!

Australian shark encounter population

- Let's compare these sampling plans.




```
##                                        Min 1st Qu. Median 3rd Qu.   Max
## Without Replacement                  128.0   149.8  155.7   161.7 183.6
## With Replacement                     106.8   147.5  156.0   164.4 203.6
## With Replacement but no duplicates 107.9   148.4  155.8   163.4 192.2
```

- Comparing these sampling numerically we might use the MSE of the 3 designs.

```
population=sharks[popSharksAustralia,]$Length

# the 10,000 averages based on different sampling designs
aveWithout = unlist(aveWithout)
aveWith = unlist(aveWith)
aveWithUnique = unlist(aveWithUnique)

average = matrix(0, nrow=length(aveWith), ncol=3)
average[,1] <- aveWithout
average[,2] <- aveWith
average[,3] <- aveWithUnique

temp = rbind(apply(average,2,mean) -mean(population),
apply(average,2,sd), (apply(average,2,mean)-mean(population))^2 + apply(average,2,var) )
```

```r
dimnames(temp)[[1]] = c("Bias", "StDev", "MSE")
dimnames(temp)[[2]] = c("Without Replacement",
                        "With Replacement", "With Replacement but no duplicates")
round(t(temp),4)
```

```
##                                      Bias   StDev      MSE
## Without Replacement                -0.0704  8.6538  74.8937
## With Replacement                   -0.0365 12.4550 155.1272
## With Replacement but no duplicates -0.0340 10.9842 120.6528
```

Selecting n balls

- Suppose that we had a box containing N different balls that are either white or black.
  - We would like to estimate the proportion of balls in the box which are black by drawing n balls at random from the box.

1. Simple random sampling **without** replacement.
   - Randomly draw n balls from the box one after another, **without replacing** any at any time.
   - The estimate is the proportion of black balls.
2. Simple random sampling **with** replacement.
   - Randomly draw n balls from the box one after another, **each time replacing** the ball.
   - Again the estimate is the proportion of black balls.
3. Randomly varying sample sizes.
   - Select one ball at a time and record its score before returning it to the box mark the ball with an X.
   - If a ball drawn already has an X marked on it, then it counts as a draw, is returned to the box.
   - Continue in this way until n draws have been made.
   - The estimate is the proportion of black balls from the number of unmarked balls.

The sampling distribution

Suppose that we have N balls in total; M black and $N - M$ white and $N > n$ & $M > n$. Let X be the number of black balls and the number of balls selected.

- Sampling without replacement (hypergeometric distribution)

$$P(X=k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \qquad k=0,1,2$$

in which

$$E\left(\frac{X}{n}\right)=\frac{M}{N} \qquad \text{and} \qquad Var\left(\frac{X}{n}\right)=\frac{1}{n}\frac{M}{N}\frac{(N-M)}{N}\left[\frac{1-n/N}{1-1/N}\right]$$

- Sampling with replacement (Binomial distribution)

$$\Pr(X=k)=\binom{n}{k}p^k(1-p)^{n-k}, \qquad k=0,1,2$$

where $p=M/N$. Hence

$$E\left(\frac{X}{n}\right)=\frac{M}{N} \qquad \text{and} \qquad Var\left(\frac{X}{n}\right)=\frac{p(1-p)}{n}=\frac{1}{n}\frac{M}{N}\frac{(N-M)}{N}$$

- Sampling with replacement but no duplicates. Here the sample size n is random as well, hence we have a joint distribution for (X,n). We have

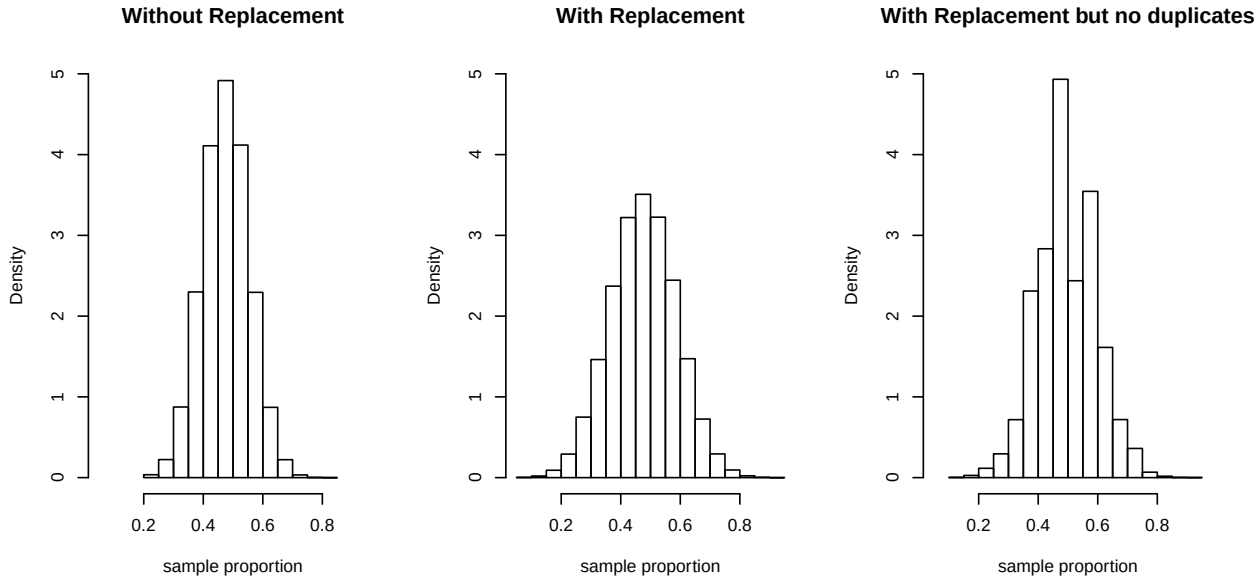
$$E\left(\frac{X}{n}\right)=\sum_{x=1}^{k}\sum_{k=1}^{n}\frac{x}{n}\,Pr(X=x\,,\,n=k) \qquad \text{and} \qquad Var\left(\frac{X}{n}\right)=E\left(\frac{X}{n}\right)^2-\left[E\left(\frac{X}{n}\right)\right]^2$$

A simulation

- Let us simulate a population of $N=40$ balls, $M=20$ of which are black.
  - We select $n=20$ balls and as such
  - in our the third scheme the sample size will vary from 1 to 20 depending how many unique balls were selected.

```
N = 40
M = 20
x = rep(1:0, times=c(M, N-M)) #x=1 means black

n = 20
m = 10^5

sam = matrix(0, nrow=m, ncol=3)
sam[,1] <- as.numeric(Map( function(i) { mean(x[sample(N, size=n, replace = FALSE) ]) }, 1:m))
sam[,2] <- as.numeric(Map( function(i) { mean(x[sample(N, size=n, replace = TRUE) ]) }, 1:m))
sam[,3] <- as.numeric(Map( function(i) { mean( x[unique(sample(N, size=n, replace = TRUE)) ]) }, 1:m))


par(mfrow=c(1,3))
nam = c("Without Replacement", "With Replacement",
"With Replacement but no duplicates")
```

```
for (i in 1:3){
  hist(sam[,i], xlim=range(sam), main=nam[i], prob=TRUE, ylim=c(0,5),xlab='sample proportion')
}
```




```
#Notice that in this example, mean(x)=M/N
temp = rbind( apply(sam,2,mean)-mean(x),
apply(sam,2,sd), (apply(sam,2,mean)-mean(x))^2 + apply(sam,2,var) )
dimnames(temp)[[1]] = c("Bias", "Std. Dev.", "Sampling of MSE")
dimnames(temp)[[2]] = c("Without Replacement",
                        "With Replacement", "With Replacement but no duplicates")
round(t(temp),5)
```

```
##                                      Bias Std. Dev. Sampling of MSE
## Without Replacement                -0.00005   0.08008         0.00641
## With Replacement                    0.00037   0.11159         0.01245
## With Replacement but no duplicates  0.00034   0.09935         0.00987
```

Suppose the task is to estimate the population mean and we assume

- $D1$: sampling without replacement
- $D2$: sampling with replacement
- $D3$: sampling with replacement but no duplicates

then it can be shown that

$$MSE_{D1} < MSE_{D3} < MSE_{D2}$$

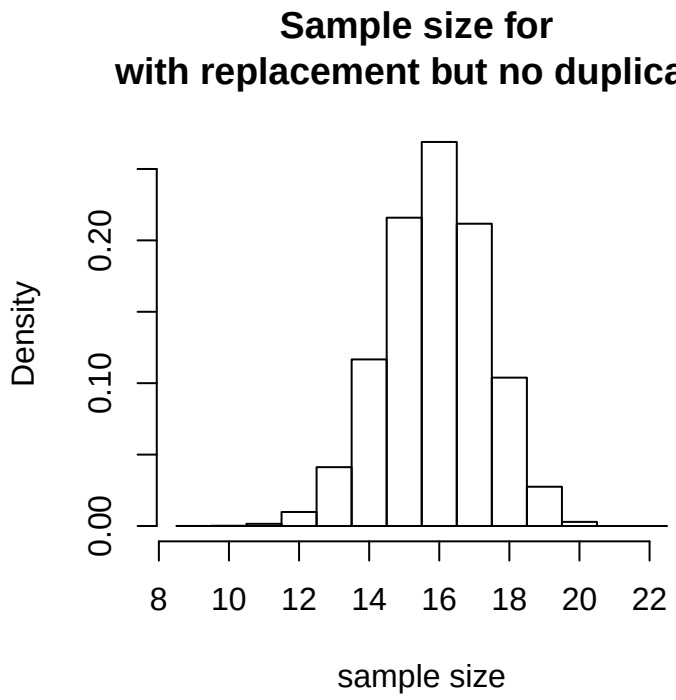
and since they are all unbiased designs,

$$Var_{D1} < Var_{D3} < Var_{D2}$$

- Histogram of the sample size

```
m = 10^5

n <- as.numeric(Map( function(i) { length( x[unique(sample(N, size=n, replace = TRUE)) ]) }, 1:m))
```

```
nam = c("Sample size for \n with replacement but no duplicates")

hist(n, breaks=seq(+8.5, 22.5, 1), main=nam, prob=TRUE,
     xlab='sample size')
```




```
summary(n)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   15.00   16.00   15.89   17.00   20.00
```

3.2.3.2 Implementation of sampling mechanisms

We could implement any of the above sampling mechanisms as a single call to a creator function.

```
### This will create a sampling mechanism
createSamplingMechanism <- function (pop, method = c("withoutReplacement","withReplacement",
 "withUnique")) {

  method = match.arg(method)
  switch (
    method,
    "withReplacement"  = function (sampSize) {
      sample(pop, sampSize, replace=TRUE)
    },
    "withoutReplacement" = function (sampSize) {
      sample(pop, sampSize, replace=FALSE)
```

```
    },
    "withUnique" = function (sampSize) {
      unique(sample(pop, sampSize, replace=TRUE))
    },
    stop(paste("No sampling mechanism:", method))
      )
}
```

For example, for simple random sampling without replacement on the population of all sharks, we might define a function `srswor(sampSize)` as

```
### without replacement is the default method.
srswor <- createSamplingMechanism(popSharks)
```

which now allows us to generate a sample of any size containing **units selected without replacement** from the population of all sharks.

- A sample of size 5, 10 and 30

```
set.seed(341)
srswor(5)
```

```
## [1] "10" "58" "24" "4"  "50"
```

```
srswor(10)
```

```
##  [1] "50" "11" "16" "65" "5"  "41" "1"  "15" "45" "27"
```

```
srswor(30)
```

```
##  [1] "62" "60" "42" "15" "1"  "48" "31" "53" "34" "54" "63" "8"  "3"  "12"
## [15] "25" "56" "59" "35" "61" "21" "22" "16" "30" "29" "27" "20" "36" "23"
## [29] "2"  "44"
```

- The created function will only generate samples from the population `pop` which allows us to write different sampling mechanisms that might actually depend on some features of the population.

3.2.4 Probability of inclusion

- Probability of inclusion is the probability of a unit being in the sample
- In addition to $p(\mathcal{S})$, the probability of selecting a sample $\mathcal{S}$ from $\mathcal{P}_{\mathcal{S}}$,
  - it may be of interest to determine the probability that any unit u will appear in the sample. This can be derived from $p(\mathcal{S})$.

- Consider the indicator function

$$D(u) = \begin{cases} 1 & \text{if } u \in \mathcal{S} \\ 0 & \text{otherwise.} \end{cases}$$

$D(u)$ is a binary random variate that takes value 1 with probability $\Pr(\mathcal{S} \ni u)$ if the probability that the sample $\mathcal{S}$ contains u, and 0 otherwise.

- The probability that unit u is in $\mathcal{S}$

$$\begin{aligned}\pi_u &= E\left[D(u)\right] \\ &= 1 \times \Pr(D(u)=1) + 0 \times \Pr(D(u)=0) \\ &= \Pr(\mathcal{S} \ni u) \\ &= \textstyle\sum_{\mathcal{S} \ni u}\ p(\mathcal{S})\end{aligned}$$

  This is called the **inclusion probability** of u in the sample $\mathcal{S}$; it is the probability that the unit u will be in a sample $\mathcal{S}$ selected according to $p(\mathcal{S})$.

The joint inclusion probability

The probability that u and v are in the sample $\mathcal{S}$ is

$$\begin{aligned}\pi_{uv} &= Pr(\ \mathcal{S} \ni u \text{ and } \mathcal{S} \ni v\ ) \\ &= E\left(\ D(u) \times D(v)\ \right) \\ &= \textstyle\sum_{\mathcal{S} \ni u,v}\ p(\mathcal{S})\end{aligned}$$

The sums are over all $\mathcal{S} \in \mathcal{P}_{\mathcal{S}}$ containing the designated units.

Simple Random Sampling without Replacement

- The inclusion probability is

$$\pi_u = \Pr\left(u \in \mathcal{S}\right) = \frac{1 \times \binom{N-1}{n-1}}{\binom{N}{n}} = \frac{1 \times \binom{N-1}{n-1}}{\frac{N}{n} \times \binom{N-1}{n-1}} = \frac{n}{N}$$

- and the joint inclusion probabilities are(assuming $u \neq v$)

$$\pi_{uv} = \Pr\left(u \in \mathcal{S} \cap v \in \mathcal{S}\right) = \frac{1 \times 1 \times \binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

- and the joint inclusion probabilities, when $v = u$

$$\pi_{uv} = \pi_u = \frac{n}{N}$$

Some Results (sampling without replacement)

- Since the indicator function $D(u)$ is one if the unit is in the sample,
  - its sum over $u \in \mathcal{P}$ must be one, i.e.

$$\sum_{u \in \mathcal{P}} D(u) = n = \text{ the size of sample } \mathcal{S}$$

and then taking expectations we have

$$\mathbb{E}\left[\sum_{u \in \mathcal{P}} D(u)\right] = \sum_{u \in \mathcal{P}} \mathbb{E}\left[D(u)\right] = \mathbb{E}\left[n\right],$$

$$\sum_{u \in \mathcal{P}} \pi_u = \mathbb{E}\left[n\right],$$

and if n is fixed (provide an example when it is not), we have

$$\sum_{u \in \mathcal{P}} \pi_u = n = \text{ the size of } \mathcal{S}$$

and we have

$$\sum_{v \in \mathcal{P}} \pi_{uv} = n\pi_u.$$

Simple Random Sampling with Replacement

- It is more challenging to calculate the inclusion probabilities for simple random sampling *with* replacement.
- It can be shown that the inclusion probability is

$$\pi_u = 1 - \left(\frac{N-1}{N}\right)^n.$$

  - and the joint inclusion probabilities (assuming $u \neq v$)

$$\pi_{uv} = 1 - 2\left(\frac{N-1}{N}\right)^n + \left(\frac{N-2}{N}\right)^n.$$

Simple Random Sampling with Replacement But Only Unique Units

- The inclusion probabilities for sampling with replacement but using only the unique units selected (i.e. the "curious" mechanism discussed earlier and investigated by Basu) are identical to simple random sampling *with* replacement.