Sampling

- As the sample size increases,
  - sample attribute values concentrate about the population attribute (at least, we hope that happens),
  - this concentration reassures us that estimating the population attribute from a sample attribute may not be too misleading.
- For any particular sample, there is little to suggest whether it is good or bad in itself.

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

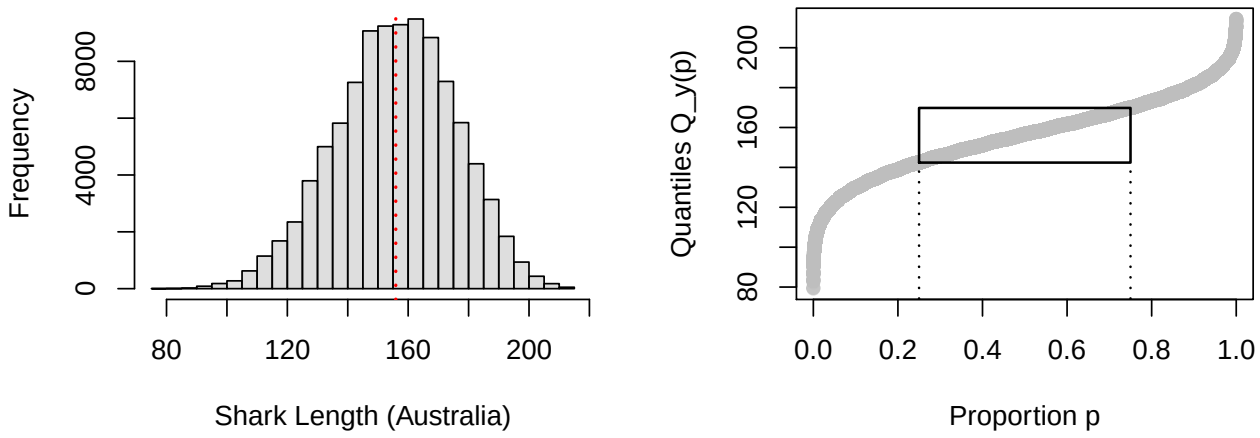
- Consider the population of M samples with size n.

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

- Any attribute $a(\mathcal{S}_i)$ is now just a variate on that unit!

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

- 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 $a(\mathcal{S})$.



Randomly selecting a Sample

- This is good news!
  - This means that by **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 $p(\mathcal{S}) = \frac{1}{M}$.

  - Read off many other probabilities about $a(\mathcal{S})$ from the histogram or the quantile plot.

Randomly selecting m Samples

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 $\sum_{k=1}^{K} M_k = M$.

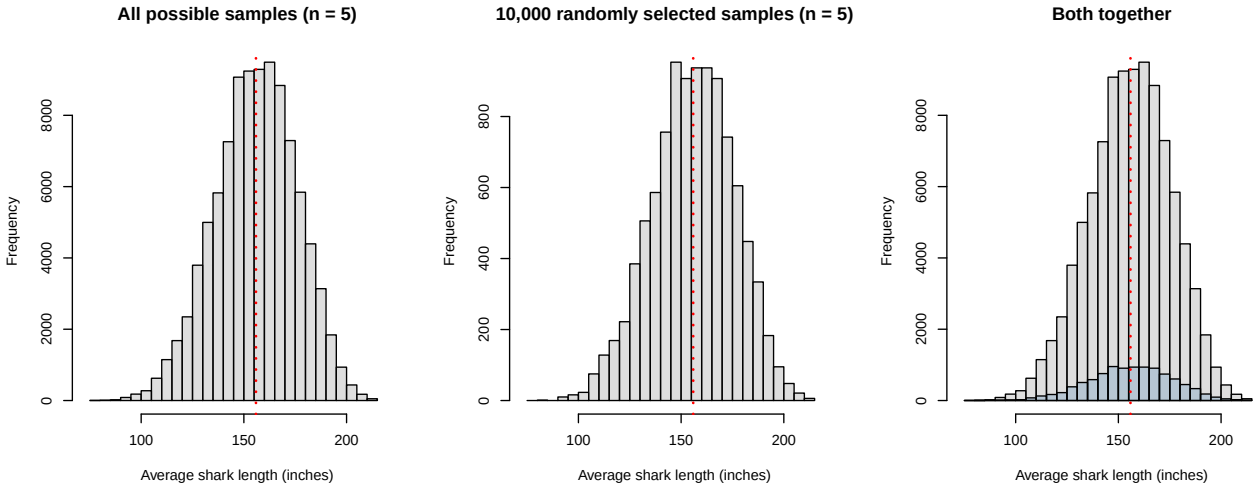


Figure 1: All versus 10,000 randomly selected samples (n = 5)

  - 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 $m_1, \ldots, m_K$.
- See more details in the notes.

Quantile Plot

Sampling Design

- 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**

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

  - For any collection of samples (or population of samples) $\mathcal{P}_{\mathcal{S}}$, 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**

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

  Sampling bias is just an **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 $a(\mathcal{P})$.

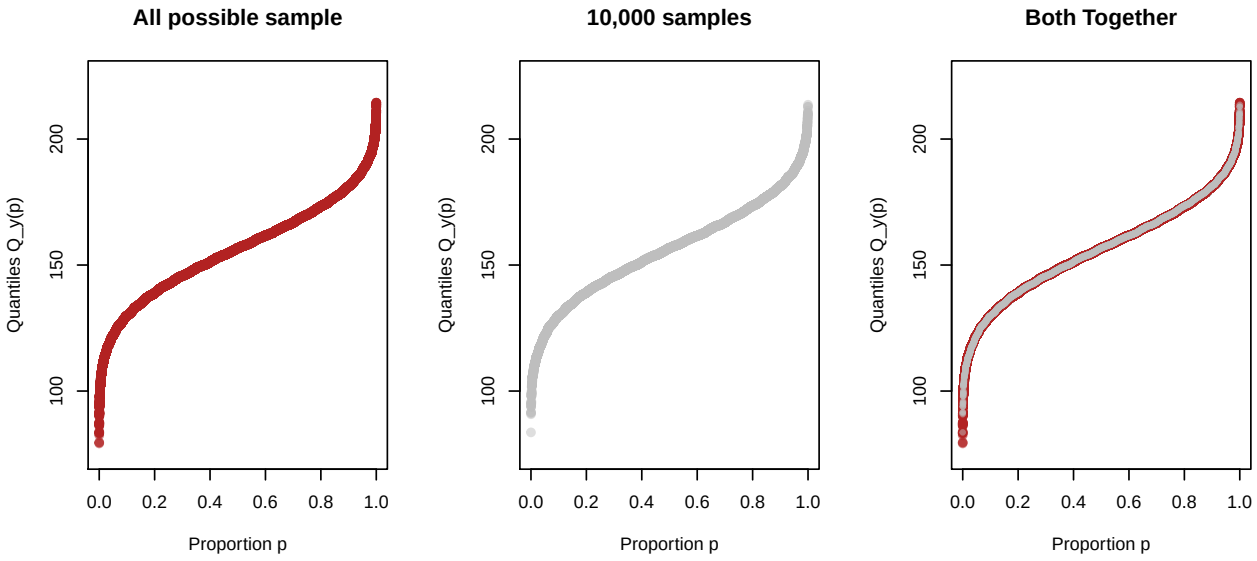


Figure 2: All possible sample average attribute values (n = 5)

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

$$MSE\left(a(\mathcal{S})\right) = Var\left(a(\mathcal{S})\right) + [\text{Sampling Bias}]^2$$

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 tunrs 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 show how the probability distirbution of samples $p(\mathcal{S})$, i.e. sampling design, affects the efficacy of samples.

Suppose that the population consists of five units

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

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

```
sam2 = combn(5,2)
sam2 # sam2 represent the units in all possible samples of size 2
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    1    1    1    1    2    2    2    3    3     4
## [2,]    2    3    4    5    3    4    5    4    5     5
```

```
# We now order the columns of sam2 with respect to the value of the sample mean
# to introduce a new sampling design.

a2 <- apply(sam2, MARGIN = 2, FUN = function(s){mean(x[s])})
sam2 = sam2[,order(a2)]
a2 = sort(a2)
sam2
```

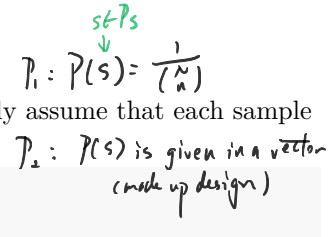
```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    1    1    2    1    1    2    2    3    3     4
## [2,]    2    3    3    4    5    4    5    4    5     5
```

```
a2
```

```
##  [1] -1.025 -0.685 -0.650 -0.115 -0.095 -0.080 -0.060  0.260  0.280  0.850
```

Example cont'd: Two sampling designs :

- p1 assigns same probability to the 10 possible samples (1/10 each).
- p2 a *biased* design: there is, really, no intuition behind this design, so simply assume that each sample of size 2 is chosen with probabilities p2

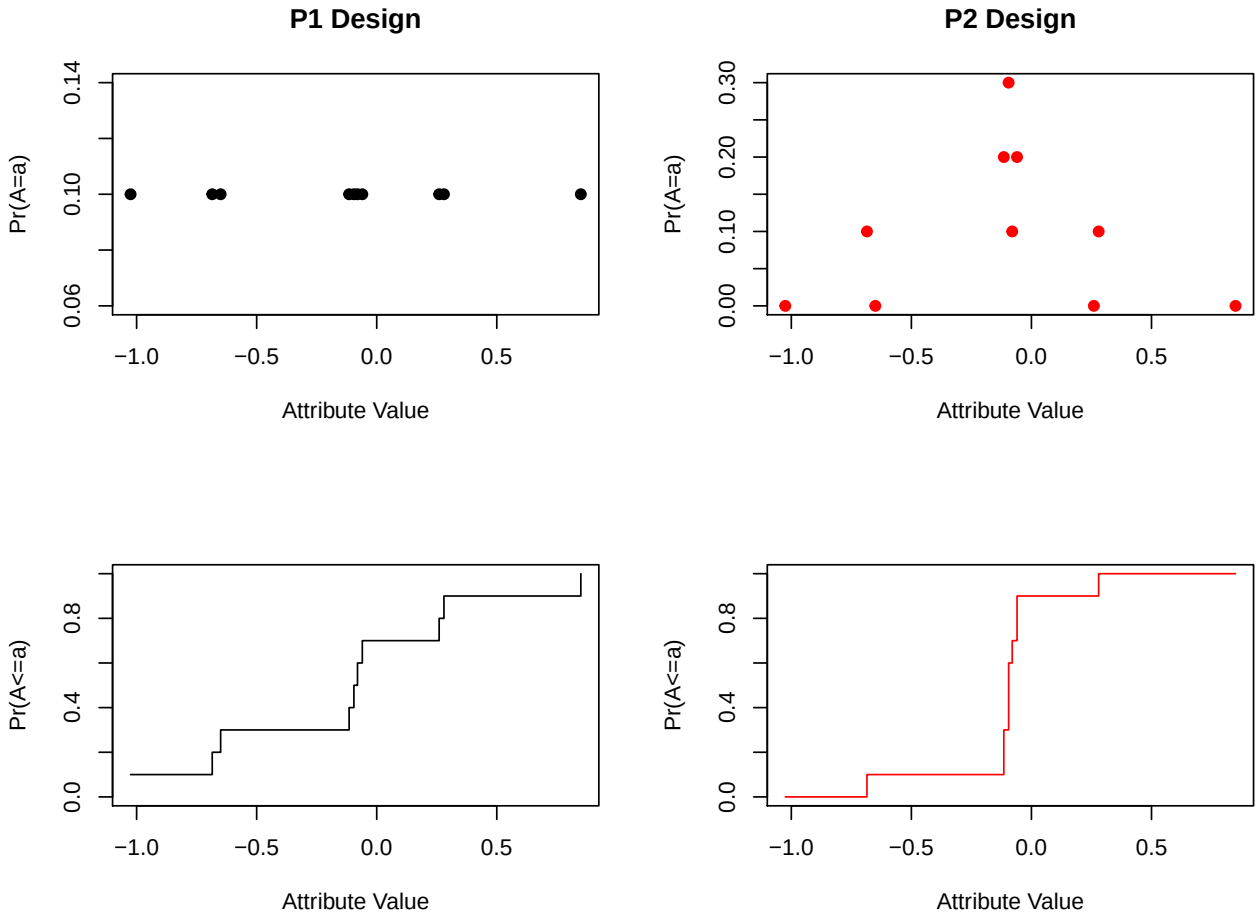



```
p1 = rep(1/10,10)
p2 = abs(apply(sam2, 2, diff))-1
p2 = p2/sum(p2)

rbind(p1,p2)
```

```
##    [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## p1  0.1  0.1  0.1  0.1  0.1  0.1  0.1  0.1  0.1   0.1
## p2  0.0  0.1  0.0  0.2  0.3  0.1  0.2  0.0  0.1   0.0
```

Distribution of sampling design, i.e. $p(A = a_i)$**:**

```
par(mfrow=c(2,2),oma=c(0,0,0,0))
plot(a2, p1, xlab="Attribute Value", ylab="Pr(A=a)", pch=19,main="P1 Design")
plot(a2, p2, xlab="Attribute Value", ylab="Pr(A=a)", pch=19,col=2,main="P2 Design")

plot(a2, cumsum(p1), xlab="Attribute Value", ylab="Pr(A<=a)", pch=19, type='s',ylim=c(0,1))
plot(a2, cumsum(p2), 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 p2 design is more concentrated.

Example cont'd: Mean Square Error

- 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 Variance

```
c( sum( ( a2 -  sum(a2*p1) )^2*p1 ), sum( ( a2 -  sum(a2*p2) )^2*p2 ) )
```

```
## [1] 0.266886 0.048931
```

Sampling bias

$\mu = E[a(s)] = \sum_i a_i P(a(s)=a_i)$

$Var(a(s)) = \sum_i (a_i - \mu)^2 P(a(s)=a_i)$

$Bias = E[a(s)] - \mu^* \rightarrow mean(x)$

```r
mean(a2) - c(sum(a2*p1), sum(a2*p2))
```

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

Sampling MSE

```r
bias = mean(a2) - c(sum(a2*p1), sum(a2*p2))
samp.var = c( sum( ( a2 -  sum(a2*p1) )^2*p1 ), sum( ( a2 -  sum(a2*p2) )^2*p2 ) )

rbind( bias, samp.var, MSE=samp.var + bias^2)
```

```
##              [,1]      [,2]
## bias     0.000000 -0.020000
## samp.var 0.266886  0.048931
## MSE      0.266886  0.049331
```

Note: Although the p2 scheme is biased, it has a lower sampling MSE.

Large Populations

- We randomly selectd a sample, $\mathcal{S}$, from $\mathcal{P}_{\mathcal{S}}$ according to the probability measure $p(\mathcal{S})$
  - Rather than constructing all $\binom{N}{n}$ 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 sample sizes $n = 100$, 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 $p2$ in the example above).

Sampling mechanisms

- Rather than constructing a probability measure on all possible samples
  - we form $\mathcal{S}$ by selecting n units $u_{i_1}, u_{i_2}, \ldots, u_{i_n}$ directly from the population of units $\mathcal{P} = \{u_1, u_2, \ldots, u_N\}$.
  - i.e. 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 appeared 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 $\mathcal{P}_{\mathcal{S}}$.

Simple Random Sampling without Replacement

- In `R`, the indices of a simple random sample of size n from indices $1, \ldots, N$ generated without replacement is returned from the function call `sample(N,n)`.

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

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

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

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

```
x[samplex]
```

```
## [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(x, 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.

- Check the notes for another way of looking at this sampling scheme based on permutations.

Simple Random Sampling with Replacement (R code)

- To generate simple random amples 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)
x = round(rnorm(10),3)
samplex = sample(10, 5, replace=TRUE)
samplex
```

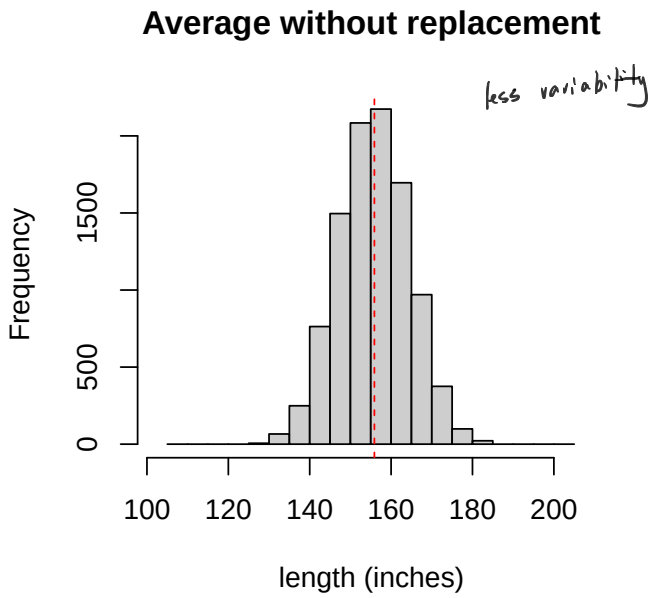
```
## [1]  8  6 10  6 10
```

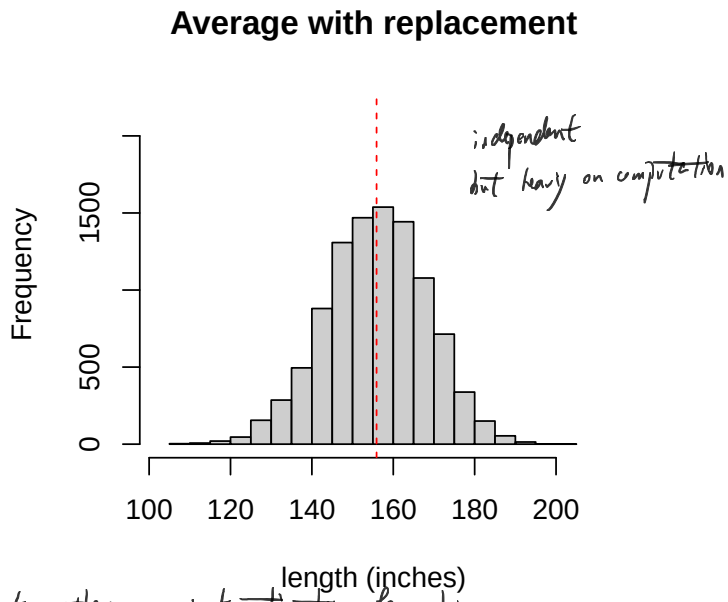
```
x[samplex]
```

```
## [1] -0.051  0.473 -0.720  0.473 -0.720
```

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).





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

```
##                       Min 1st Qu. Median 2nd 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
```

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)
x = round(rnorm(10),3)
samplex = sample(10, 5, replace=TRUE)
samplex
```

```
## [1]  8  6 10  6 10
```

- Simple random sample with replacement yields

```
x[samplex]
```

```
## [1] -0.051  0.473 -0.720  0.473 -0.720
```

- Simple random sample with replacement removing duplicate units yields

```
unique(x[samplex])
```

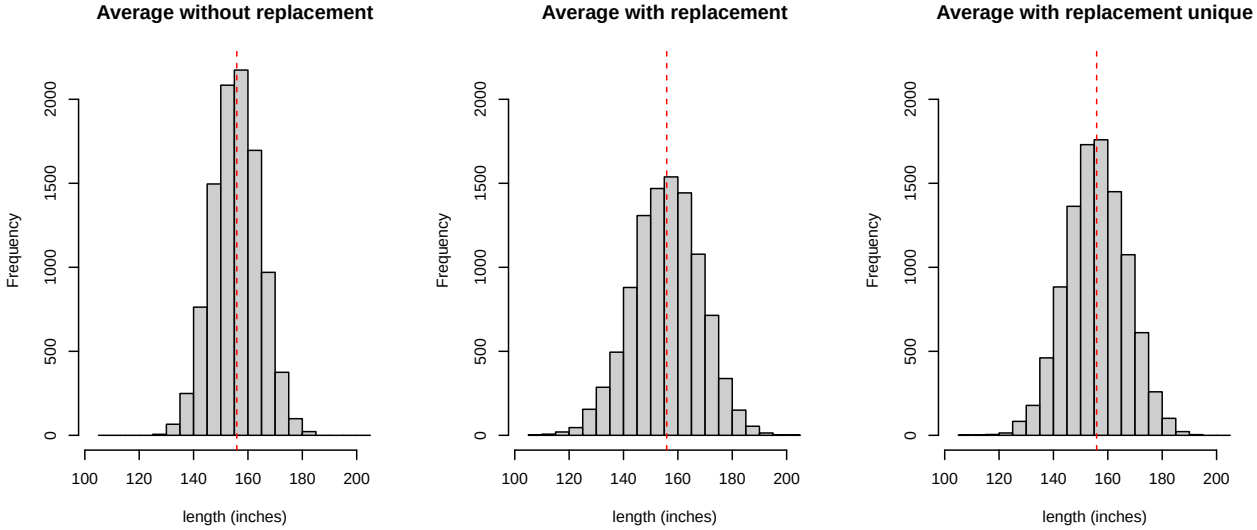
```
## [1] -0.051  0.473 -0.720
```

Curious design:

n is a random variable

In the shark example, avg. length of Australian shark attacks, n is a r.v. taking values from 1 to 5.

Australian shark encounter population




```
##                                        Min 1st Qu. Median 2nd 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
```

Why?

- 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. $\hat{p} = \frac{b}{n}$
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. $\hat{p} = \frac{b}{n}$
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, but ignore it 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.

Selecting two balls

- Suppose that we have N balls in totals; 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 is $n = 2$.

- Sampling without replacement (hypergeometric distribution)

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

and $\mathbb{E}[X] = nM/N$ and then

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

- Sampling with replacement (Binomial distribution)

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

where $p = M/N$ and $\mathbb{E}[X] = np = nM/N$. Then

$$\mathbb{E}[X/n] = M/N \qquad \text{and} \qquad \mathbb{V}\text{ar}[X/n] = \frac{p(1-p)}{n} = \frac{1}{n}\frac{M}{N}\frac{(M-N)}{N}$$

- Sampling with replacement but no duplicates. Here the sample size n is random as well. Then the joint probabilities can be represented by a table.

$\Pr(n, X)$	$X=0$	$X=1$	$X=2$	$P(n=)$
$n=1$				
$n=2$				
$\Pr(X=)$				

$$\mathbb{E}[X/n] = \sum_{x=}^{2}\sum_{n=1}^{2} \frac{x}{n}\Pr(n, X)$$

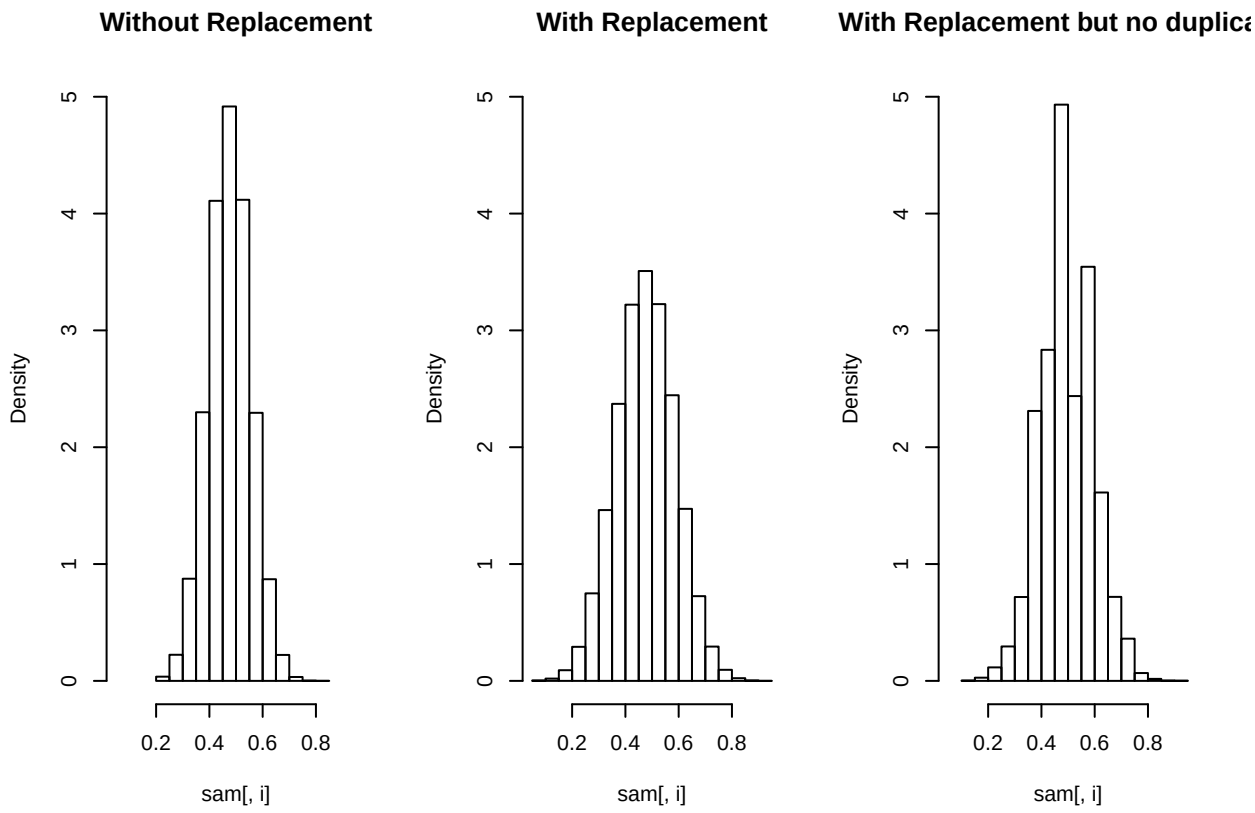
$$\mathbb{V}\text{ar}[X/n] = \frac{1}{2}\frac{M(N-M)}{N^2}$$

Selecting n balls

Let us simulate a population of N balls, M of which are black. We sample $n = 20$ balls, but for the third scheme, 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))
}
```




```
temp = rbind( apply(sam,2,mean) -mean(x),
apply(sam,2,sd), (apply(sam,2,mean)-M/N)^2 + apply(sam,2,var) )
dimnames(temp)[[1]] = c("Bias", "Srd. Dev.", "Sampling of MSE")
dimnames(temp)[[2]] = c("Without Replacement",
                        "With Replacement", "With Replacement but no duplicates")
t(temp)
```

```
##                                          Bias  Srd. Dev.
## Without Replacement                -0.0000480000 0.08008191
## With Replacement                    0.0003665000 0.11159106
## With Replacement but no duplicates  0.0003441454 0.09934775
##                                    Sampling of MSE
## Without Replacement                    0.006413114
## With Replacement                       0.012452700
## With Replacement but no duplicates     0.009870094
```

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.

```
set.seed(354661)
### A sample of size 5
srswor(5)
```

```
## [1] "45" "30" "5"  "46" "40"
```

```
### of size 10
srswor(10)
```

```
##  [1] "1"  "50" "54" "31" "61" "28" "55" "32" "23" "26"
```

```
### Of size 30
srswor(30)
```

```
##  [1] "22" "4"  "3"  "8"  "58" "20" "52" "32" "31" "49" "30" "47" "33" "23"
## [15] "39" "28" "41" "15" "44" "21" "62" "57" "10" "56" "9"  "19" "54" "51"
## [29] "63" "36"
```

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

Probability of a unit being in a sample

- In addition to the probability, $p(\mathcal{S})$, of selecting a sample $\mathcal{S}$ from $\mathcal{P}_{\mathcal{S}}$,
  - it can 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 $\mathcal{S}$

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

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

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

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

The Sample Size

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

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

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 we have

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

and we have

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

Exercise: Prove the last equation.

Sampling without replacement

- When the sampling mechanism is simple random sampling without replacement
  – the probability that a unit u will be in the sample is

$$\pi_u = \frac{n}{N}.$$

  – For joint inclusion probabilities, we have for simple random sampling *without* replacement (assuming $u \neq v$)

$$\pi_{uv} = \frac{n(n-1)}{N(N-1)}.$$

Sampling with replacement

- More challenging is the determination of the inclusion probabilities for simple random sampling *with* replacement.
- As an **Exercise** show that
  - the inclusion probability is
$$\pi_u = 1 - \left(\frac{N-1}{N}\right)^n.$$
prob of single unit in sample
  - and the joint inclusion probabilities, we have for simple random sampling *with* replacement (assuming $u \neq v$) –>
$$\pi_{uv} = 1 - 2\left(\frac{N-1}{N}\right)^n + \left(\frac{N-2}{N}\right)^n.$$
prob of 2 units in sample
- 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.