Sampling

Mostly time, more data the better but in several cases are not

- 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.

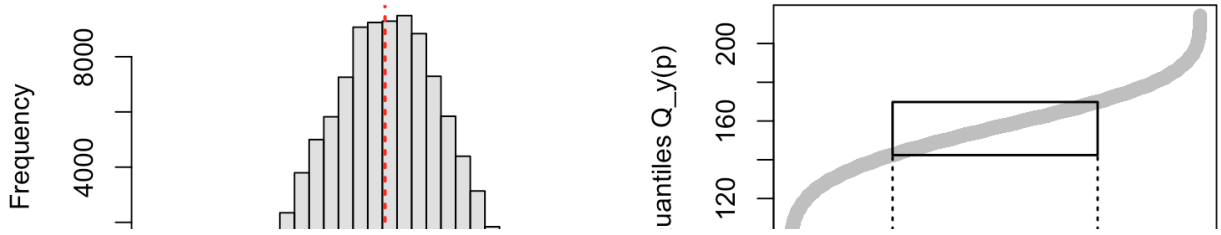
$$\mathcal{P}_{\mathcal{S}} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_M\}$$

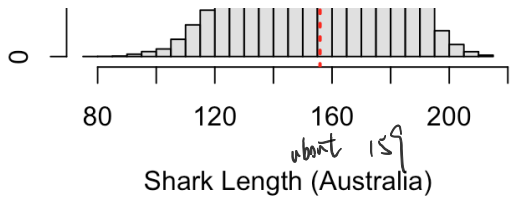
- Any attribute $a(\mathcal{S}_i)$ is now just a variate on that unit!

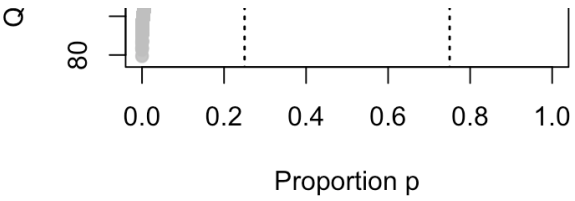
$$\mathcal{P}_{a(\mathcal{S})} = \{a(\mathcal{S}_1), a(\mathcal{S}_2), \ldots, a(\mathcal{S}_M)\}$$

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

All possible sample average attribute values (n = 5)







Randomly selecting a Sample

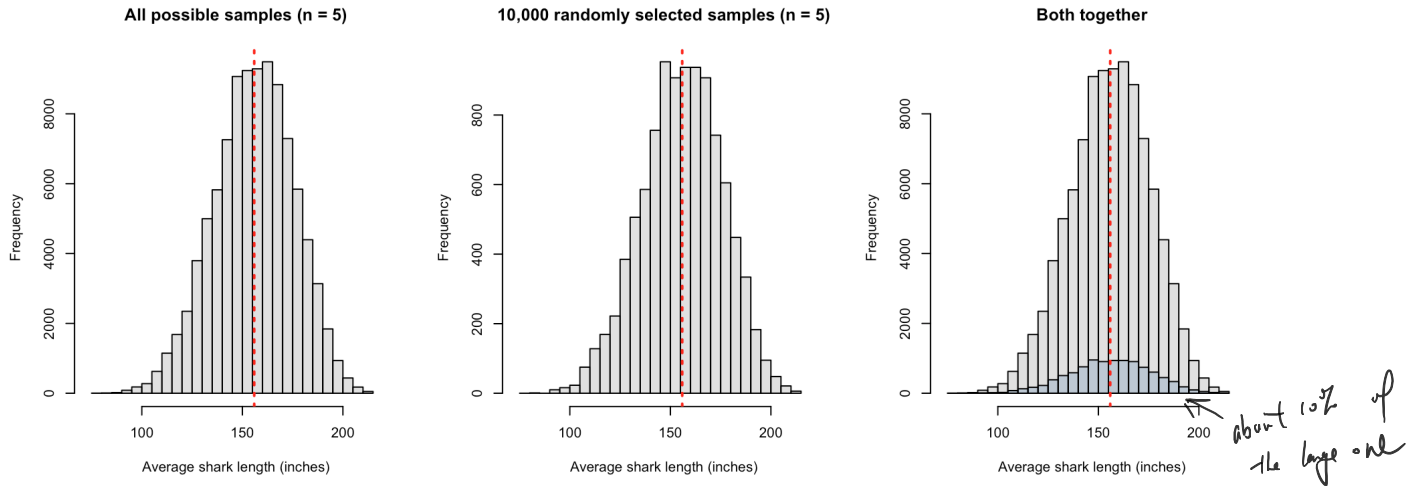
- This is good news!
  - This means that by **randomly selecting a sample** from $\mathcal{P}_S$ we are able to make probability statements regarding the attribute $a(S)$ taking on any value.
  - If $n = 5$, we know that with probability $\frac{1}{2}$ the attribute that results will be within the range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

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

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

Randomly selecting m Samples

Suppose we draw a sample of m = 10,000 samples $S_{u_1}, \ldots, S_{u_m}$ from $\mathcal{P}_S$ of $\binom{N}{n} = \binom{28}{5}$ = 98,280 possible samples.



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

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

Bin 1, ..., k -> K bins
# of elements from sample in each bin = x
Mi = # of all possible sample attributes in bin i, i=1,...,k
mi = # of 10000 possible
x ~ Multinomial ((N n) ; p1, p2, ..., pk), pi = Mi/...

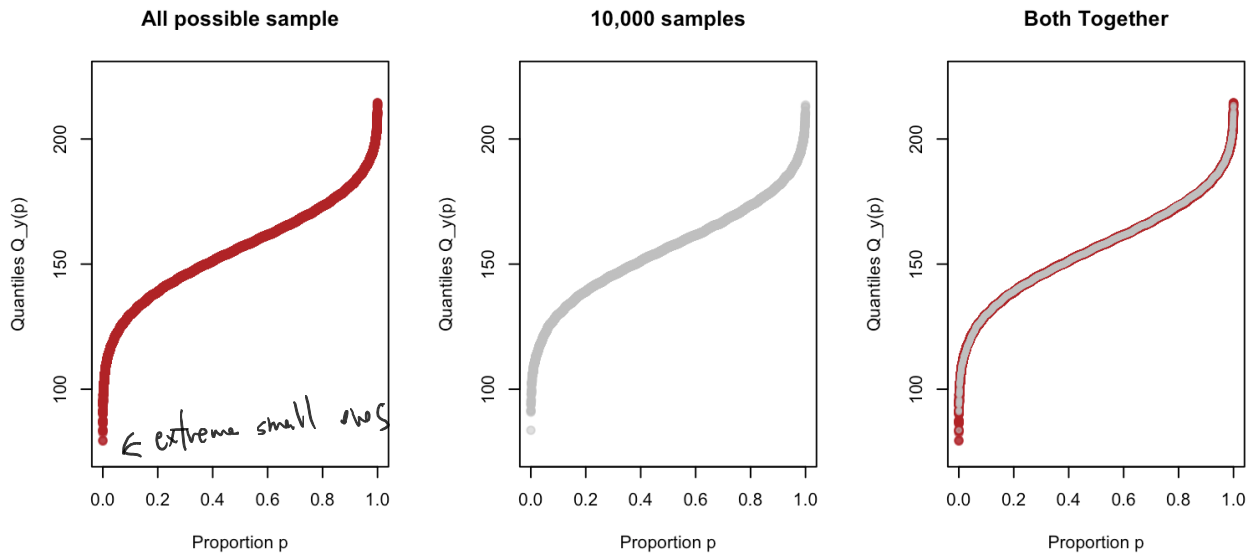
$$B_1 = (b_0, b_1], B_2 = (b_1, b_2], \dots, B_K = (b_{K-1}, b_K]$$

and

- the kth bin B_k contains $M_k \geq 0$ of the attribute values $a(S_i)$ $i = 1, \dots, M$.
- The bins contain the attribute values of all of the $S_i \in \mathcal{P}_S$ 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 $M_1, \dots, M_K$ and
  - the sampled histogram has heights $m_1, \dots, m_K$.
- See more details in the notes.

Quantile Plot



All possible sample average attribute values (n = 5)

Sampling Design

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

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

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

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

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

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

sampling bias
$E[\underbrace{a(S)}_{\text{random variable}} - \underbrace{a(P)}_{\text{number}}]$
$= \sum_i (a_i - a(P))\, P(a(S)=a_i)$
if $P(a(S)=a_i) \neq 0 \;\forall\, i$:
then we get the average sample error
if $P(a(S)=a_i) \neq 0 \;\forall\, i$:

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

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

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 $a(\mathcal{P})$.

Sampling Variance

lack of accuracy = bias
lack of precison = high variance

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

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

  - where all expectations are taken with respect to the probabilities $p(\mathcal{S})$ of the samples $\mathcal{S}$ from $\mathcal{P}_{\mathcal{S}}$.
- The sampling bias depends on the attribute $a(\cdot)$, the set of possible samples $\mathcal{P}_{\mathcal{S}}$, and the sample probabilities $p(\mathcal{S})$.
  - 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

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

$\Rightarrow$ unbiased estimator when bias = 0

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

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

$= P(\text{all samples which result in } A = a)$
$= \sum_{\text{all samples such that } A=a} p(S) = \sum_{\text{all } S} p(S) \cdot \underbrace{I(a(S)=a)}_{\text{1 or 0}}$ ($A = a(S)$)

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

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

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

Suppose that the population consists of five units

```
set.seed(341)
x = round(rnorm(5),2)
x = sort(x)
x
```

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

```
sam2 = combn(5,2)
sam2
```

```
##      [,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
```

```
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
```

Two sampling designs:

- `p1` assigns same probability to the 10 possible samples (1/10 each).
- `p2` assigns probabilities almost proportional to how far the units are in the order of observation.

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

```
##  [1] 0.0 0.1 0.0 0.2 0.3 0.1 0.2 0.0 0.1 0.0
```

Sampling bias

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

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

Sampling Variance

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

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

Mean Square Error

- Sampling Mean Square Error (MSE)

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

Two sampling designs

```
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
```

Sampling MSE

```
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.

Example Plot

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

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, typ
e='s',ylim=c(0,1))
```