Samples

Contents

3 Samples

- It may not be possible to calculate an attribute for the population. For example,
  - we might not have access to the entire population,
  - the population is too large,
  - or the attribute too complex.

- If we have a **sample** or a subset $\mathcal{S}$ of $n << N$ units,
  - Then the attribute $a(\mathcal{S})$ calculated based on this sample is an **estimate** of its population counterpart $a(\mathcal{P})$.
$$a(\mathcal{S}) = \widehat{a}(\mathcal{P}) = a(\widehat{\mathcal{P}})$$
  - The second equality emphasizes that $\mathcal{S}$ as an estimate of $\mathcal{P}$.

- Two things we might consider are
  - sample error, and
  - Fisher consistency.

Sample error

- Any difference between the actual values of the estimate $a(\mathcal{S})$ and the quantity being estimated (the **estimand**) $a(\mathcal{P})$ is an **error**.

$$\text{sample error} \quad = \quad a(\mathcal{S}) - a(\mathcal{P})$$

- The error will depend on the sample and the attribute.

- Quantifying error;
  - for numerical attributes, this is determined mathematically;
  - for graphical attributes, it is not precise and meant to be taken notionally.

Example - Agriculture Data

Load the data and obtain a sample of size $n = 100$

```
directory <- "../Data"
dirsep <-"/"
filename <- paste(directory, "agpop_data.csv", sep=dirsep)
agpop <- read.csv(filename, header=TRUE)

set.seed(341)
s = sample(length(agpop$farms87), 100)
```

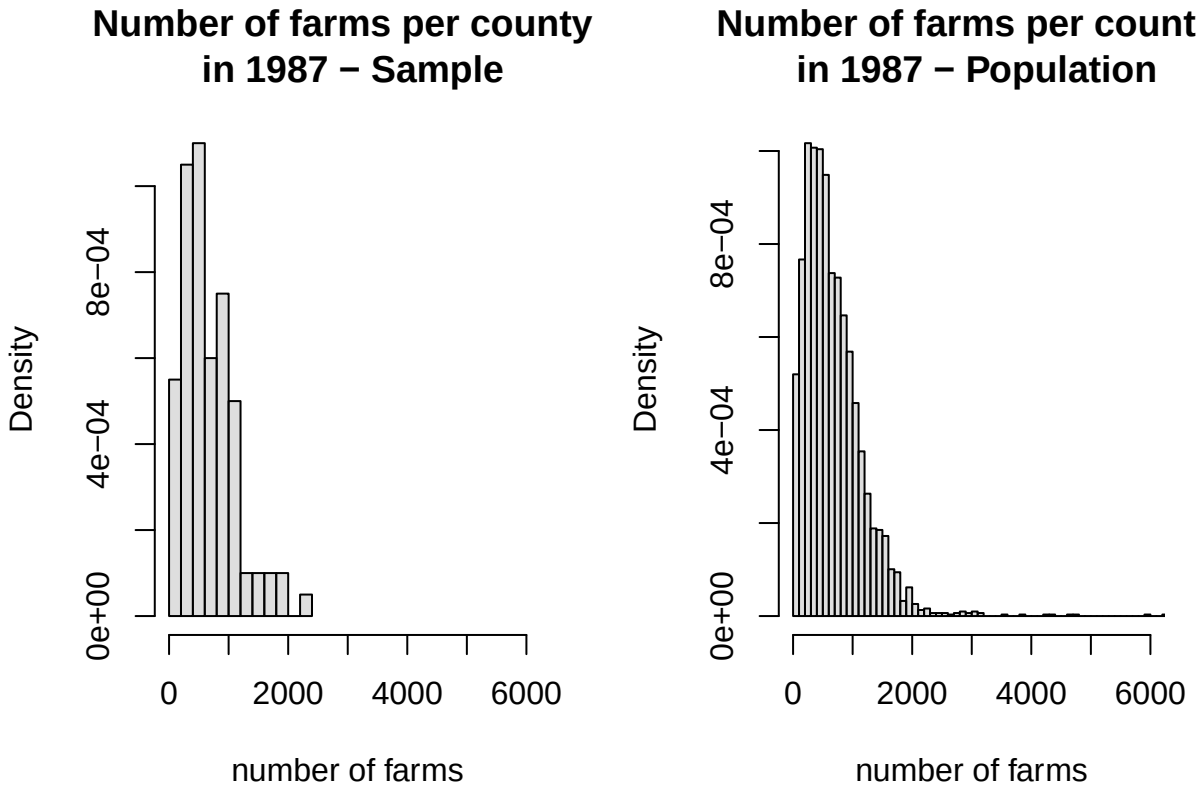
Since we have the population we can calculate the differences among some attributes.

```r
c(mean(agpop$farms87[s]) - mean(agpop$farms87),
median(agpop$farms87[s]) - median(agpop$farms87),
sd(agpop$farms87[s]) - sd(agpop$farms87),
IQR(agpop$farms87[s]) - IQR(agpop$farms87) )
```

```
## [1] -10.21428  -8.50000 -86.64667 -34.00000
```

We can compare difference from histograms.

```r
par(mfrow=c(1,2))
hist(agpop$farms87[s], breaks='FD',col=adjustcolor("grey", alpha = 0.5), main="Number of farms per coun
hist(agpop$farms87, breaks='FD',col=adjustcolor("grey", alpha = 0.5), main="Number of farms per county
```



- For obvious reasons, an attribute with lower sampling error is **often** preferable.

Fisher Consistency

- If the sample $\mathcal{S}$ is equal to the population $\mathcal{P}$ then the sample error should be zero (or non-existent), i.e. $a(\mathcal{P}) = a(\mathcal{S})$.

- This would mean that the estimation is in some sense **consistent**.
  - This type of consistency is sometimes called **Fisher consistency** in the statistical literature,

  - Named after the statistical scientist Ronald A. Fisher who in 1922 identified this consistency as an important criterion for estimation.

"The statistician cannot evade the responsibility for understanding the process he applies or recommends."

Ronald Fisher

The Sample as a Population

- In every respect the sample could be considered a population itself and might even sensibly be called a "sample population".

- We avoid this nomenclature because, unfortunately, it flies in the face of traditional statistical language and common English usage – it is to be avoided therefore and will not be used here.
  - In some applications (e.g. bootstrap, which will discuss later), we use the term "pseudo population" in reference to the sample.

- Nevertheless, treating $\mathcal{S}$ as a population allows us to evaluate any population attribute on the sample in the same way we would for $\mathcal{P}$.
- Some samples will have a small sample error and some will have a large one.
  - To quantify this we could look at all possible samples of size n.

3.1 All possible samples

- Suppose the population $\mathcal{P}$ was of size N and that the sample $\mathcal{S}$ was of size n.
  - Then there are $\binom{N}{n}$ different possible samples $\mathcal{S}$ of size n.

Shark Data

- Consider the population $\mathcal{P}$ of all the great white shark encounters reported from 1999 to 2014 worldwide.
  - There are $N = 65$ such encounters in our population.

Table 1: Number of samples of size n

n = 5	n=10	n=15	n=20
8259888	179013799328	2.073747e+14	2.83396e+16

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

Year	Sex	Age	Time	Australia	USA	Surfing	Scuba	Fatality	Injury	Length
2014	M	35	AM	1	0	1	0	0	0	180
2013	M	19	AM	0	0	1	0	0	1	140
2013	M	74	AM	0	0	0	0	1	1	144
2013	M	45	AM	0	1	1	0	0	1	95
2013	M	46	PM	0	0	0	0	1	1	156
2012	M	24	AM	1	0	1	0	1	1	196

- Even for $N = 65$, generating all possible samples of size $n = 5$ can be computationally prohibitive.
  - To reduce the computation, we focus on a sub-population of these encounters, just those which occurred in Australian waters (`sharks$Australia == 1`).

Year	Sex	Age	Time	Australia	USA	Surfing	Scuba	Fatality	Injury	Length
2014	M	35	AM	1	0	1	0	0	0	180

Year	Sex	Age	Time	Australia	USA	Surfing	Scuba	Fatality	Injury	Length
2013	M	19	AM	0	0	1	0	0	1	140
2013	M	74	AM	0	0	0	0	1	1	144
2013	M	45	AM	0	1	1	0	0	1	95
2013	M	46	PM	0	0	0	0	1	1	156
2012	M	24	AM	1	0	1	0	1	1	196

Australian waters

- This population contains only $N = 28$ units. There are now only 98,280 possible samples of size $n = 5$ from this population, a still large but much more manageable number.

```
### 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
popSharksAustralia
```

```
##  [1] "1"  "6"  "7"  "9"  "10" "11" "14" "16" "18" "19" "20" "21" "22" "24"
## [15] "25" "30" "33" "34" "37" "38" "40" "41" "48" "54" "55" "58" "59" "61"
```

Generating All Samples

- We can generate the indices of all possible samples of size n from a population of size N in `R` using the combination function `combn(...)`.
  - For example, we could construct all subsets of size 2, from the population of $\{A, B, C, D\}$

```
combn(LETTERS[1:4], 2)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,] "A"  "A"  "A"  "B"  "B"  "C"
## [2,] "B"  "C"  "D"  "C"  "D"  "D"
```

- Generating All Australia Shark Samples

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

```
## [1] 98280
```

Table 4: First five samples & the last sample (size=5)

first	second	third	fourth	fifth	last
1	1	1	1	1	54
6	6	6	6	6	55
7	7	7	7	7	58

first	second	third	fourth	fifth	last
9	9	9	9	9	59
10	11	14	16	18	61

A Population of Attributes

- For every sample we can calculate any attribute, e.g. the average shark length.
  - The attribute (average) calculated on all possible samples is

```
avePop <- mean(sharks[popSharksAustralia, "Length"])

### Because the samples are stored in a matrix,
### use the apply function to apply FUN over its columns
### (i.e. its second dimension; margin = 2)
### Each column provides the row indices
### for that sample in the original population

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

The average on the first 8 samples is

```
avesSamp[1:8]
```

```
## [1] 142.6 146.6 129.8 142.2 142.2 161.8 154.0 158.0
```

- We now have a population of attributes, i.e. a population of sample means.
  - We can calculate attributes on this population to summarize it.

- The histogram of the sample attributes:
  - The red dotted line is the value of the attribute on the population, $a(\mathcal{P}) = 156$.
  - The attributes (sample averages) range from 79 to 214 inches.
- Comments
  - There are a few samples that produce a value far from the population value.
  - Concentration near the population value.
  - *nearly* symmetric about the population value, and bell-shaped.
- A numerical summary of the sample averages is

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    79.2   142.4   156.8   155.9   169.8   214.4
```

- Half of the samples will produce an average shark length between 142.4 and 169.8 inches.

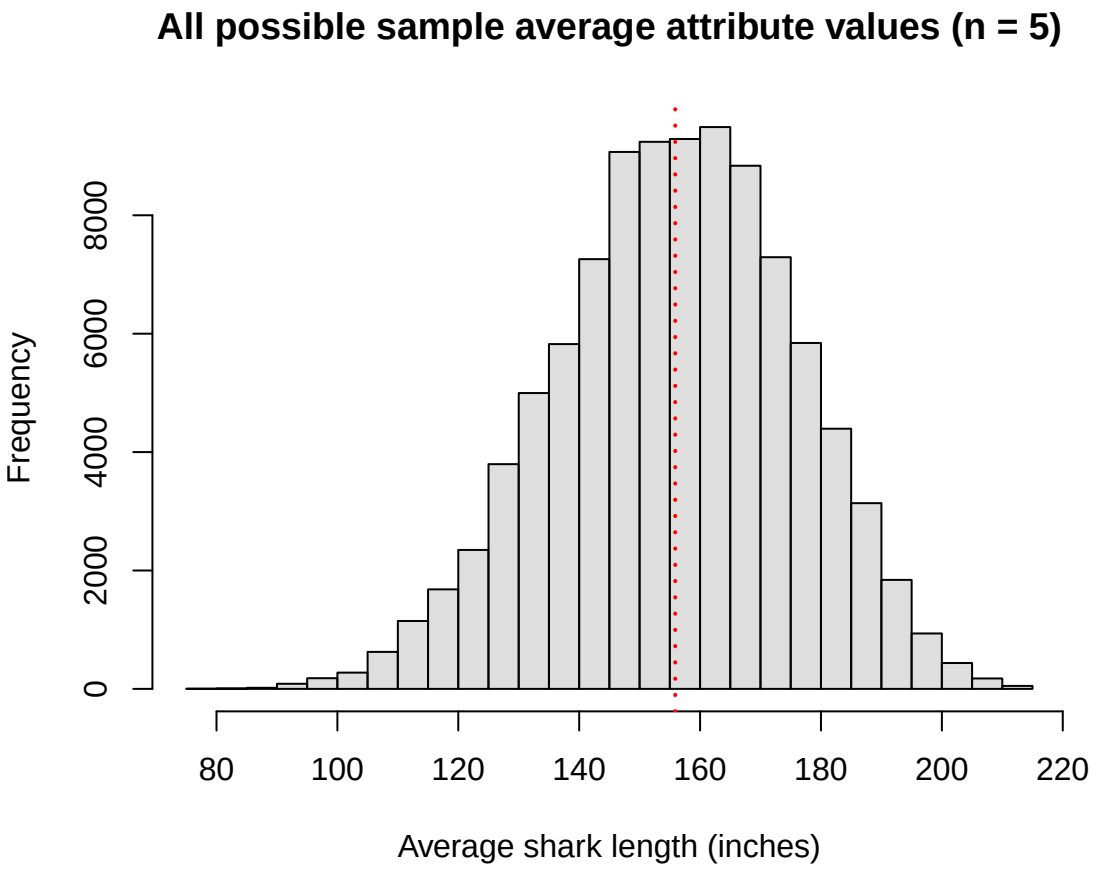


Figure 1: All possible samples: great white encounters in Australia

- This is somewhat reassuring, especially given the sample is of size 5 (which is little more than 1/7 the population size).

Sample Error

- Suppose we are interested in the average length (in inches) of great white sharks encountering humans in Australian waters.
- The sample error for a sample $\mathcal{S}$ of size n is

$$a(S) - a(\mathcal{P}_{Australia}) = \frac{1}{n}\sum_{u\in\mathcal{S}} y_u - \frac{1}{N}\sum_{u\in\mathcal{P}_{Australia}} y_u.$$

- We can calculate the sample error for all possible samples and print off the first 8 sample errors.

```
sampleErrors <- avesSamp - avePop
sampleErrors[1:8]
```

```
## [1] -13.292857  -9.292857 -26.092857 -13.692857 -13.692857   5.907143
## [7]  -1.892857   2.107143
```

- The sample error ranges from -77 to 59 inches.
- A numerical summary of the sample errors is

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -76.6929 -13.4929   0.9071   0.0000  13.9071  58.5071
```

Average Sample Error

- The **average sample error** over all possible samples of size n is

$$\text{Average sample error} \quad = \frac{\sum_{i=1}^{N_s} a(\mathcal{S}_i)}{N_s} - a(\mathcal{P})$$

  where N_s (= 98,280 here) is the number of possible samples $\mathcal{S}_i$.

  - For the average shark length, the average sample error was actually `round(mean(avesSamp) - avePop,5)` = 0.
  - At least for this attribute, the sample error is zero on average.

3.1.1 Effect of sample size

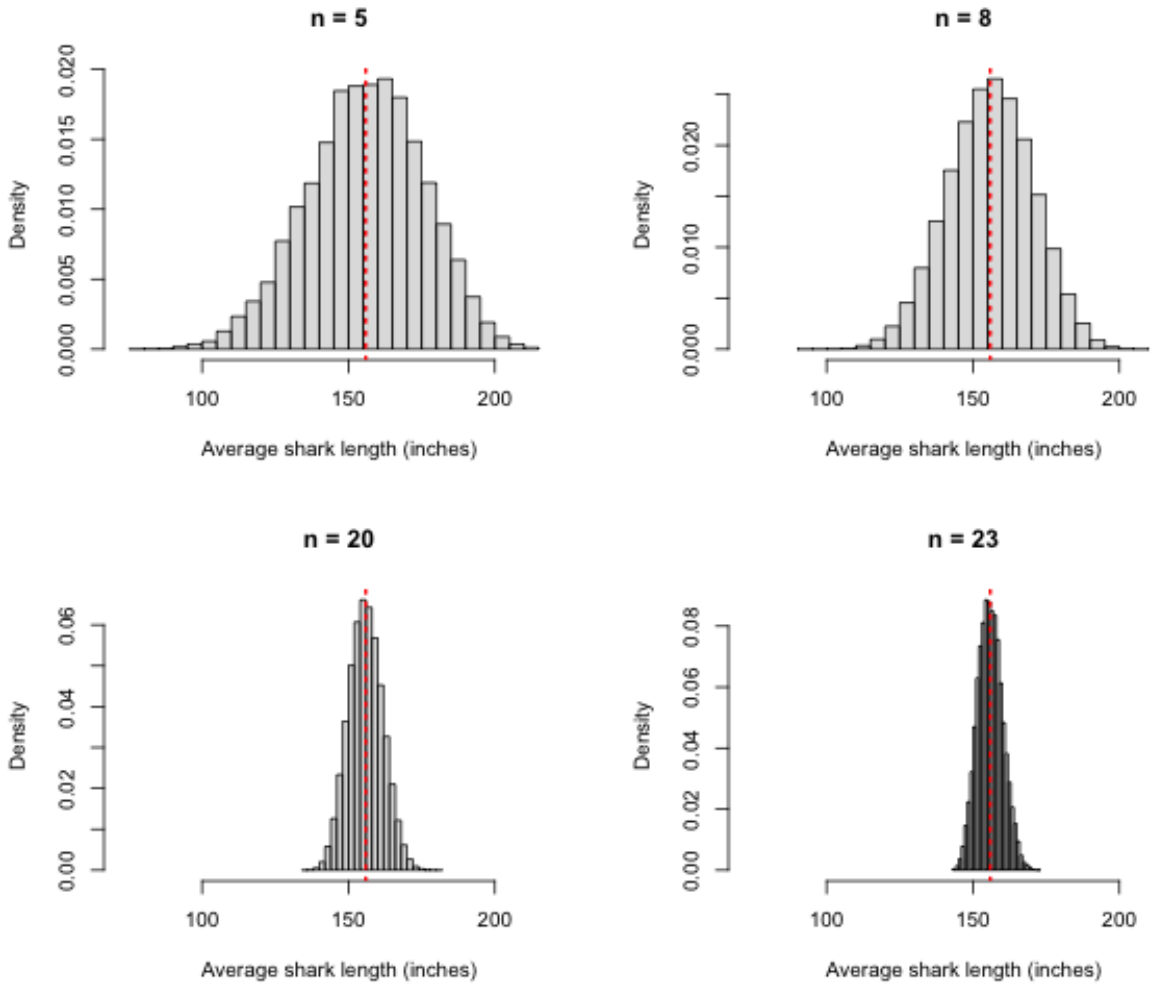


Figure 2: All possible samples for different sample sizes

- **Note:** These plots demonstrate the effect of sample size on a particular attribute: the sample mean.
- The concentration around the true value (red line) indicates some kind of **consistency** for the particular attribute here (viz. the arithmetic average).

- To quantify this concentration we look at

$$\mid a(S) - a(\mathcal{P}_{Australia}) \mid = \left| \frac{1}{n} \sum_{u \in \mathcal{S}} y_u - \frac{1}{N} \sum_{u \in \mathcal{P}_{Australia}} y_u \right| < c$$

  for some $c > 0$

  – Then we could calculate the proportion of samples that satisfy this.

Consistency

- This definition of consistency is different and separate from Fisher consistency.
- Consider a population $\mathcal{P}$ of size $N < \infty$.
  - For each n, we can construct the set of all possible samples.

$$\mathcal{P}_{\mathcal{S}}(n) = \{\mathcal{S} \ : \ \mathcal{S} \subset \mathcal{P} \text{ and } |S| = n\}$$

  - For any $c > 0$,

$$\mathcal{P}_a(c, n) = \{\mathcal{S} \ : \ \mathcal{S} \subset \mathcal{P}_{\mathcal{S}}(n) \text{ and } \mid a(\mathcal{S}) - a(\mathcal{P}) \mid < c\}$$

    and define the proportion

$$p_a(c, n) = \frac{|\mathcal{P}_a(c, n)|}{|\mathcal{P}_{\mathcal{S}}(n)|}$$

    for all $c > 0$, and $n \leq N$.

- Plotting the absolute sample error versus the proportion $p_a(c, n)$ for varying sample size n, for a fixed $c > 0$, we see that $p_a(c, n)$ increases with n.

3.1.2 Consistency for Other attributes

We will focus on two types of attributes to study consistency:

- Location attributes (central tendency)
  - Mean
  - Median
  - Trimmed Mean: 100p% trimmed average is the arithmetic average of the central 100(1-2p)% of the sorted values

- Scale attributes (spread)
  - Range: $|y_{max} - y_{min}|$
  - Interquartile Range: $Q_y(0.75) - Q_y(0.25)$
  - Standard Deviation: $\sqrt{\frac{\sum_{u \in \mathcal{P}} (y_u - \overline{y})^2}{N}}$

We will use the length of the shark variable in encounters in Australian waters population ($N = 28$)

```
hist(sharks[popSharksAustralia, "Length"], col=adjustcolor("grey", alpha = 0.5),main="Shark Encounters
     xlab="shark length (inches)",breaks=25)
```

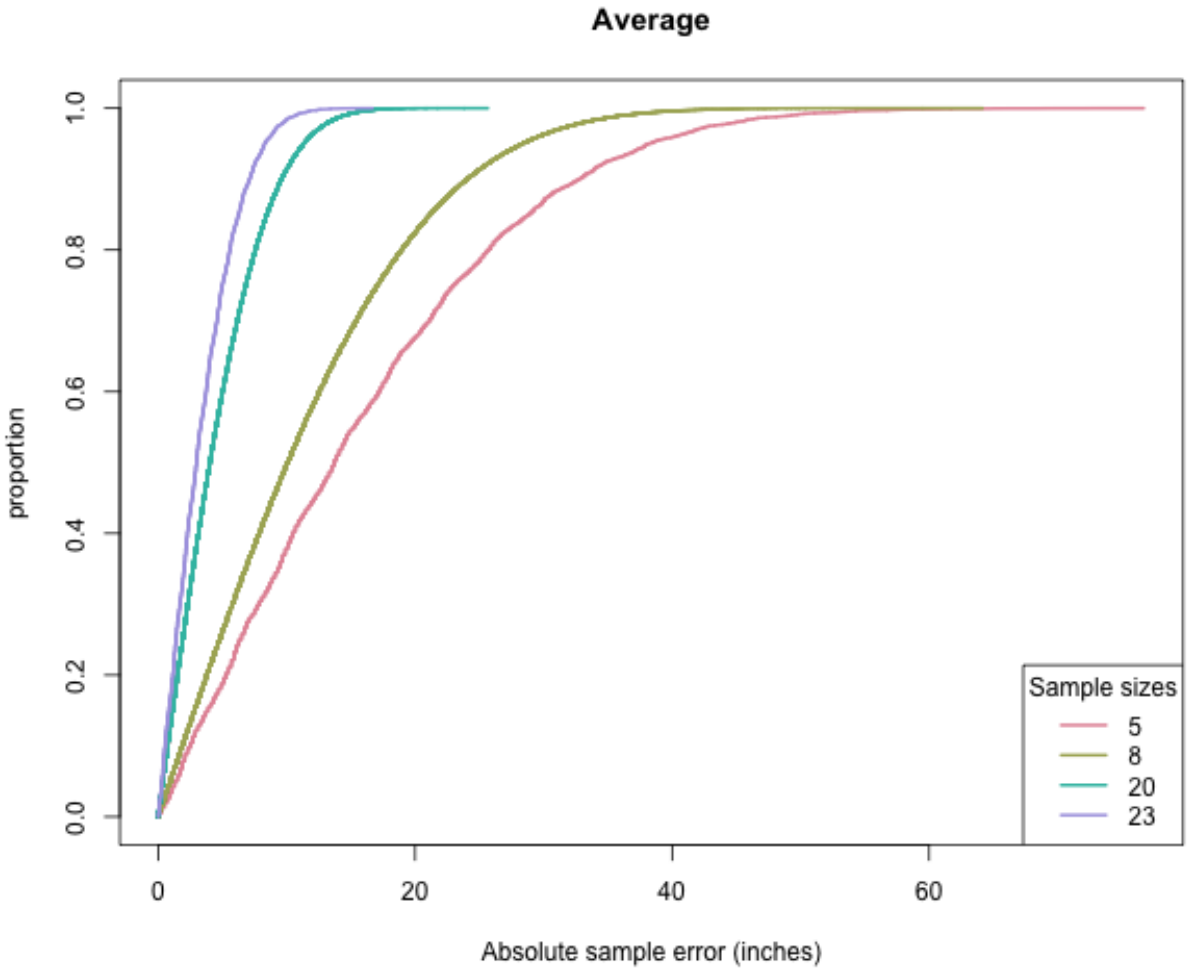
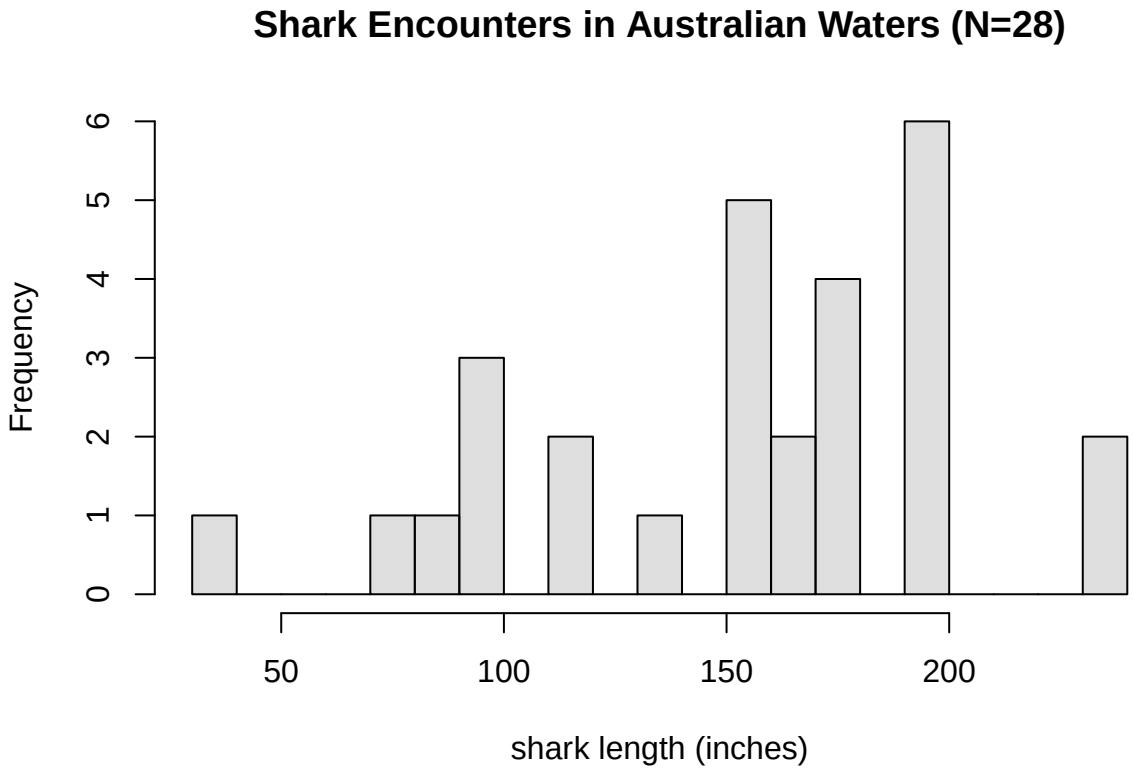

Average
proportion
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
Absolute sample error (inches)
Sample sizes
5
8
20
23

Figure 3:



Location attributes

- The location attributes for samples of size $n = 5$.
  - Note that these are all plotted on the same scale to aid the comparisons.
  - The value of the attribute calculated on the whole population is marked with red dotted line.
- **Average:** as before.
- **Trimmed average:** in this case, for all sample calculations, any trim below 20% will be exactly the same as the regular average (why?)
- **Median:** quite different. When there are an odd number like $n = 5$ units in the sample, the median will be one of the y values in the sample.
  - This is why we see such distinct bars in the histogram.
  - Nevertheless, the sample attribute values do concentrate around the population value, even more so than does either average.

Scale Attributes

- **Range:**
  - sample values quite far from the population range, considerably underestimating its value (why?).

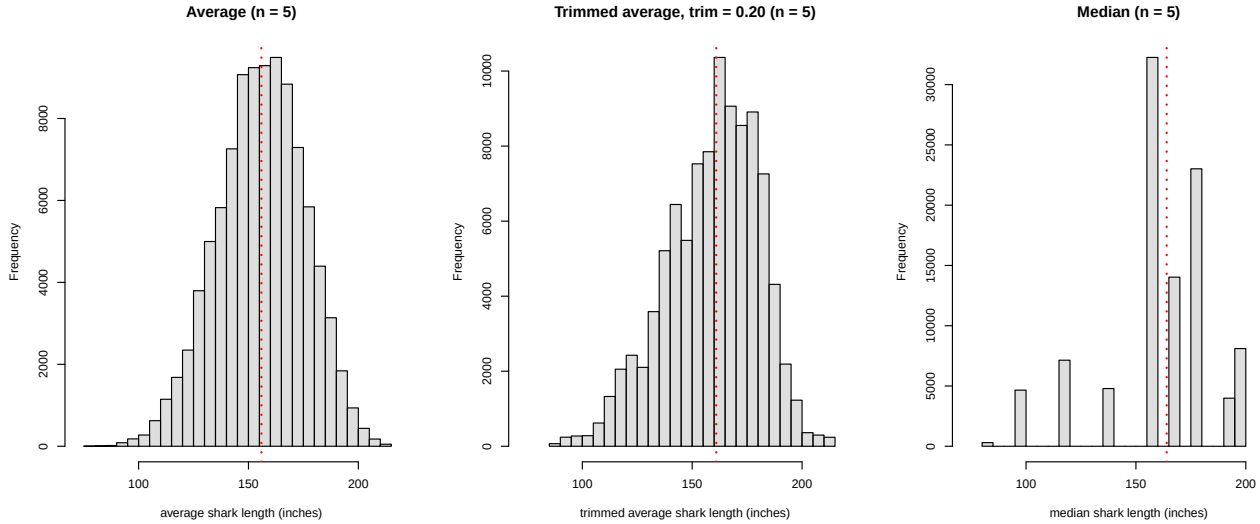


Figure 4: Different location attributes: over all possible samples (n = 5)

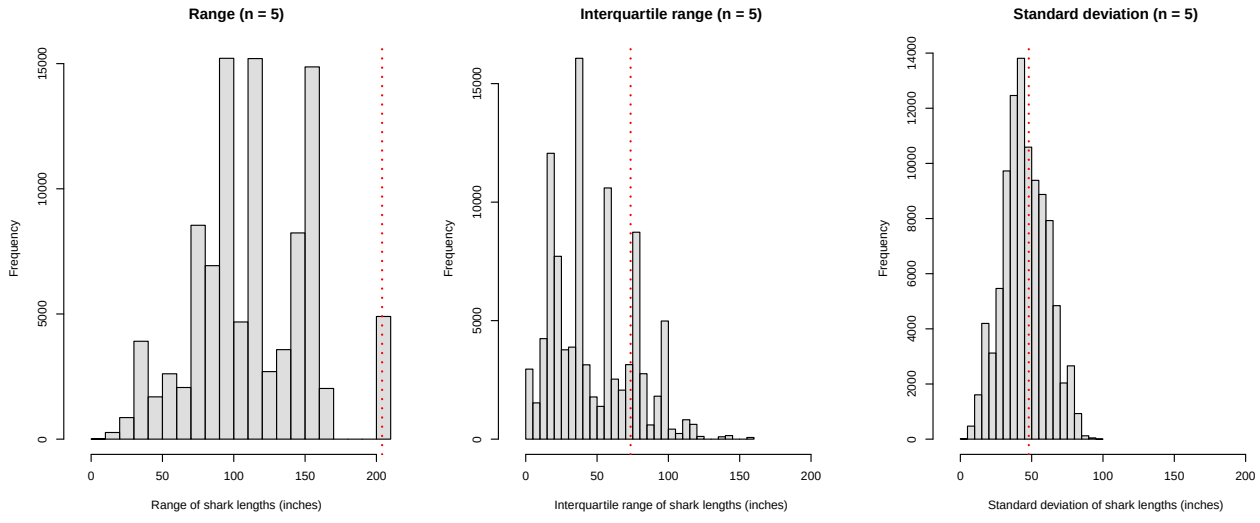


Figure 5: Different scale attributes: over all possible samples (n = 5)

  - sample errors would be mostly negative. Which samples does the bar on the right represent?
- **Interquartile range:**
  - performs much better with some positive sample errors in addition to the negative sample errors.
  - the population interquartile range appears to be far more frequently underestimated than over estimated.
- **Standard deviation:**
  - much more like an average... sample values concentrate (roughly symmetrically compared to the interquartile range) about the population value.

Trimmed Average

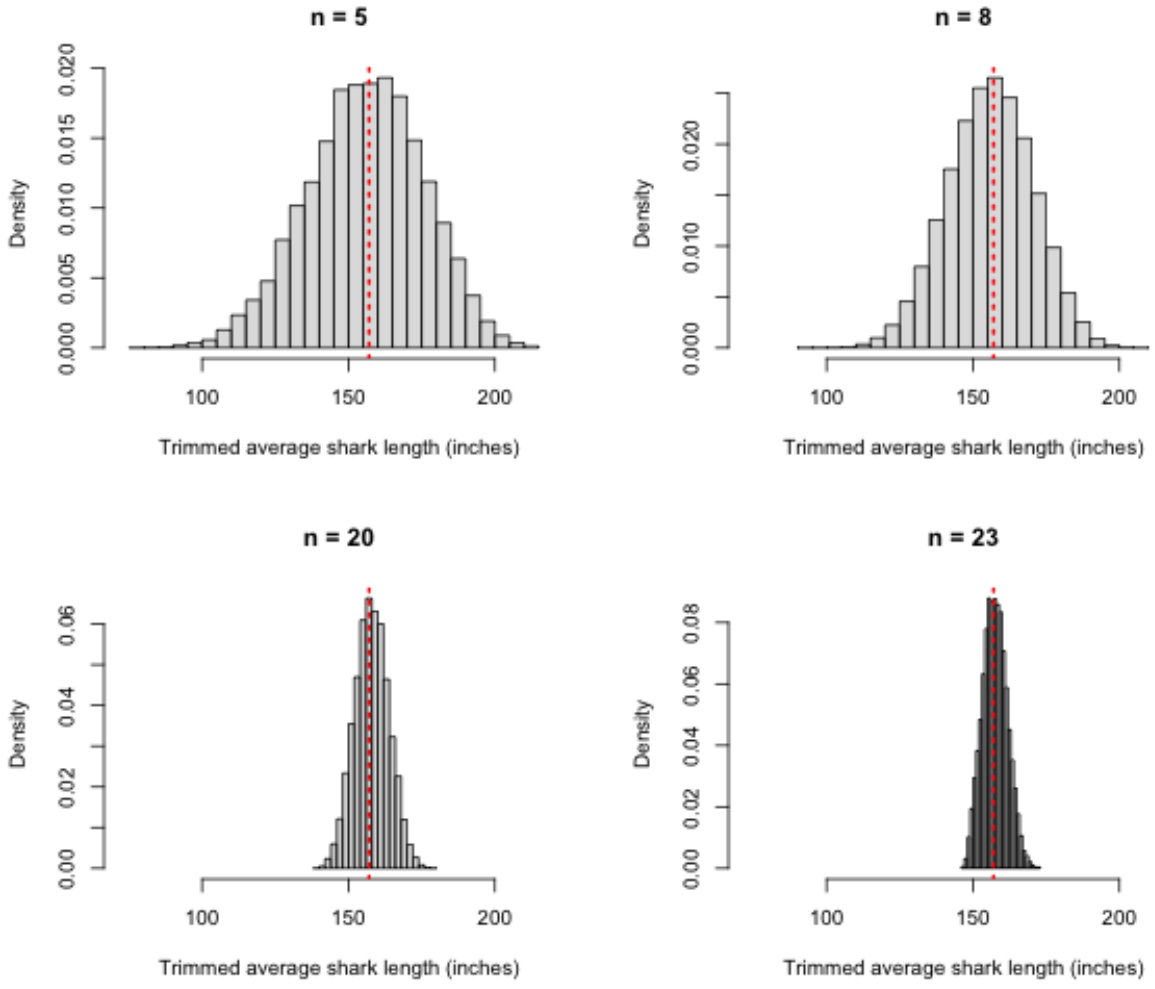


Figure 6: Trimmed averages (trim = 0.10) over all possible samples for different sample sizes

- The trimmed average behaves much like the ordinary average.

- Note the effect of increasing n on concentration

Median

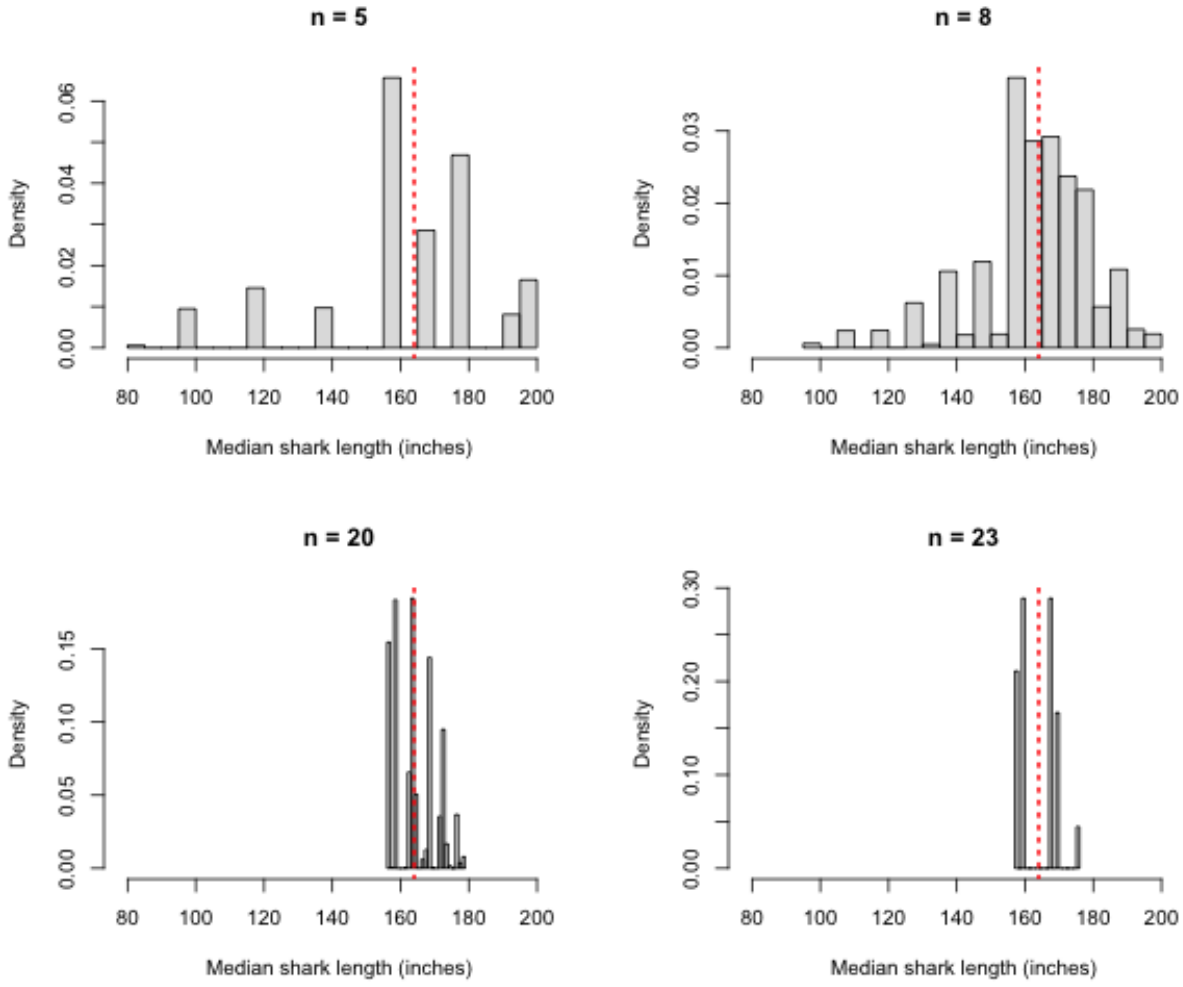


Figure 7: Medians over all possible samples for different sample sizes

- There is a greater variety of possible values when the sample size n is even (why?)
- As the sample size increases, there is a greater concentration of the sample values about the population values.

Range

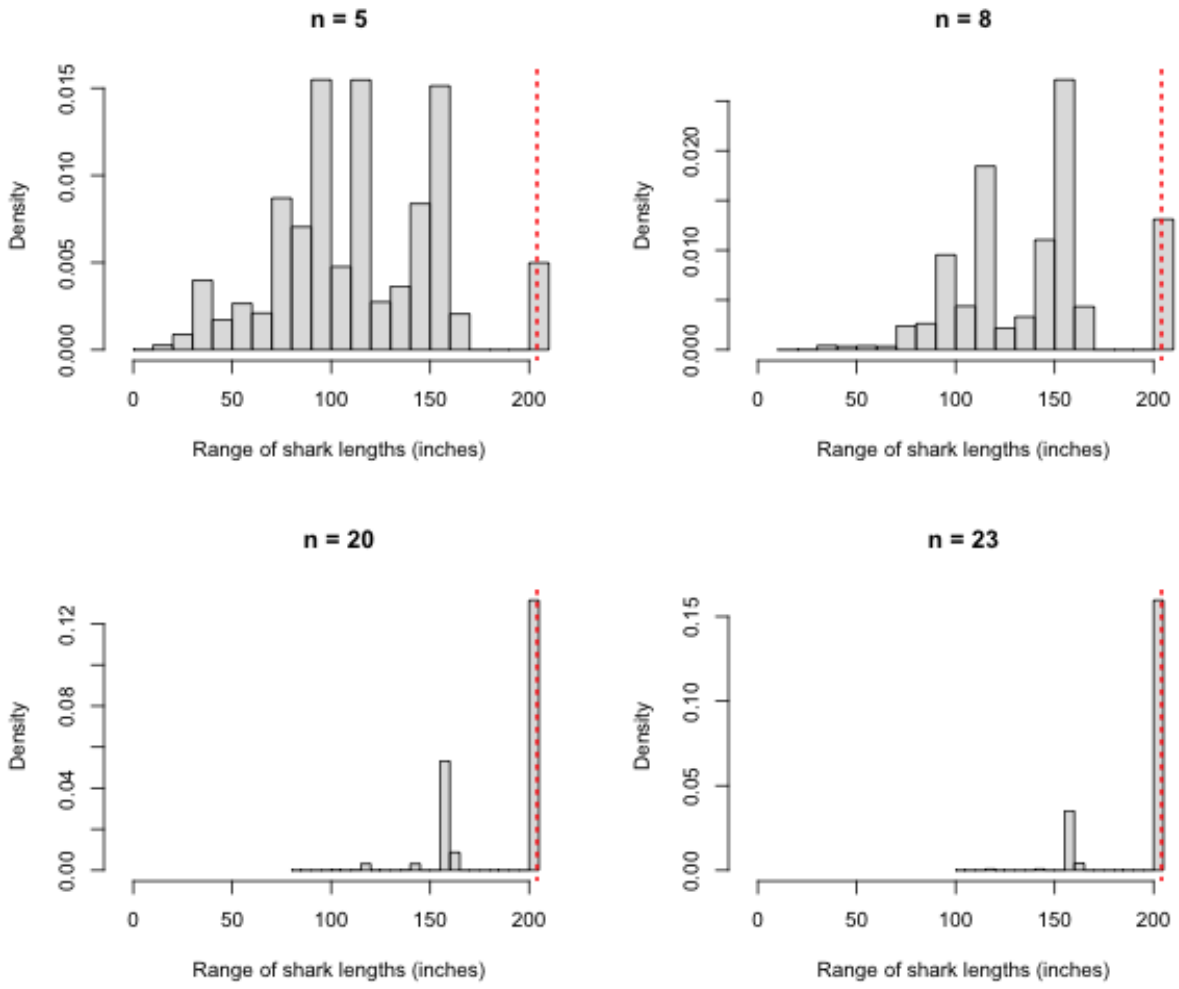


Figure 8: Ranges over all possible samples for different sample sizes

- The range shows a consistent underestimation of the population value. - The average sample error will be negative.
- As the sample size increases, more samples will contain both y_{min} and y_{max} and so will match the population value of the range.

Interquartile Ranges

- The interquartile range histogram becomes more symmetric and increasingly concentrated about the population value as n increases.

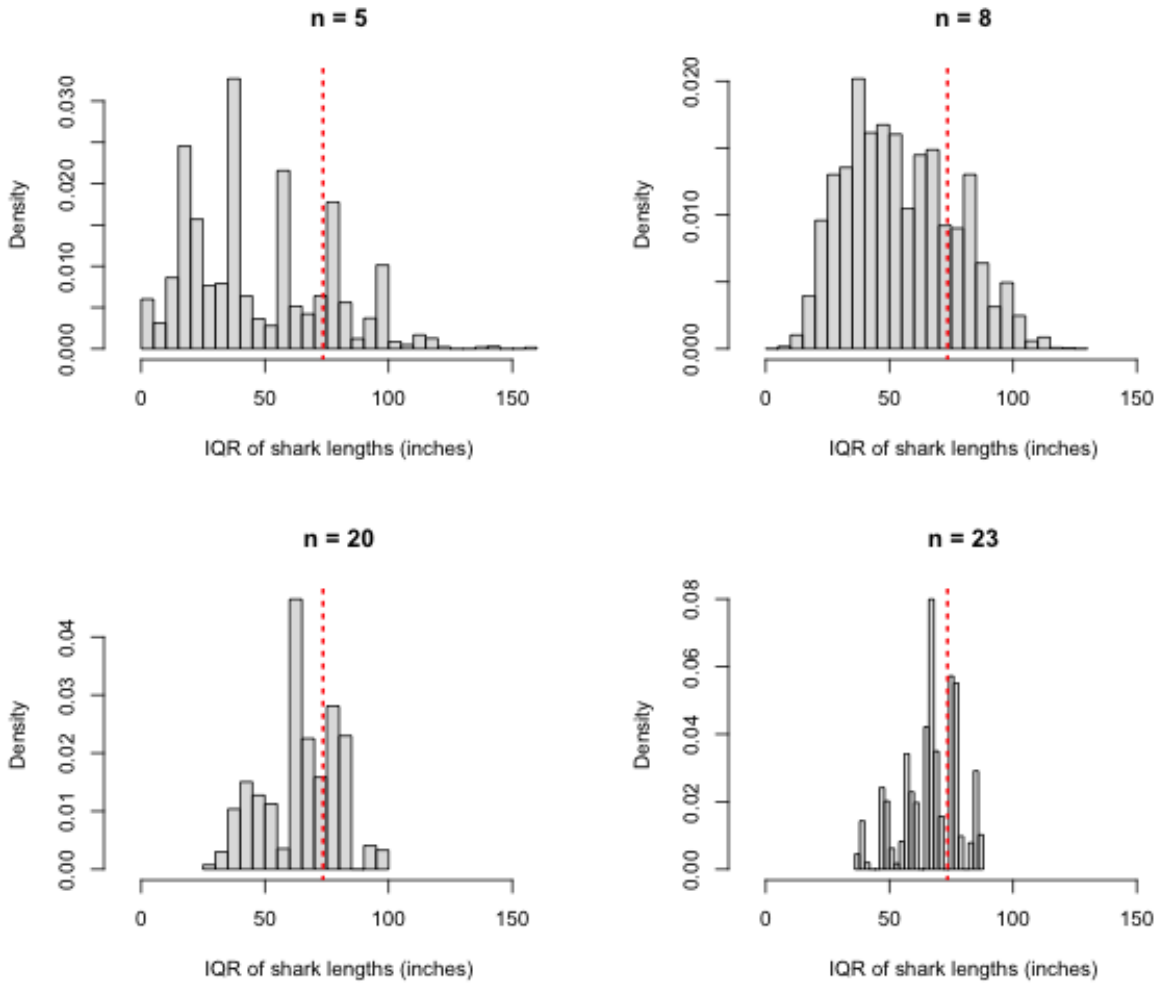


Figure 9: Interquartile ranges over all possible samples for different sample sizes

Standard Deviation

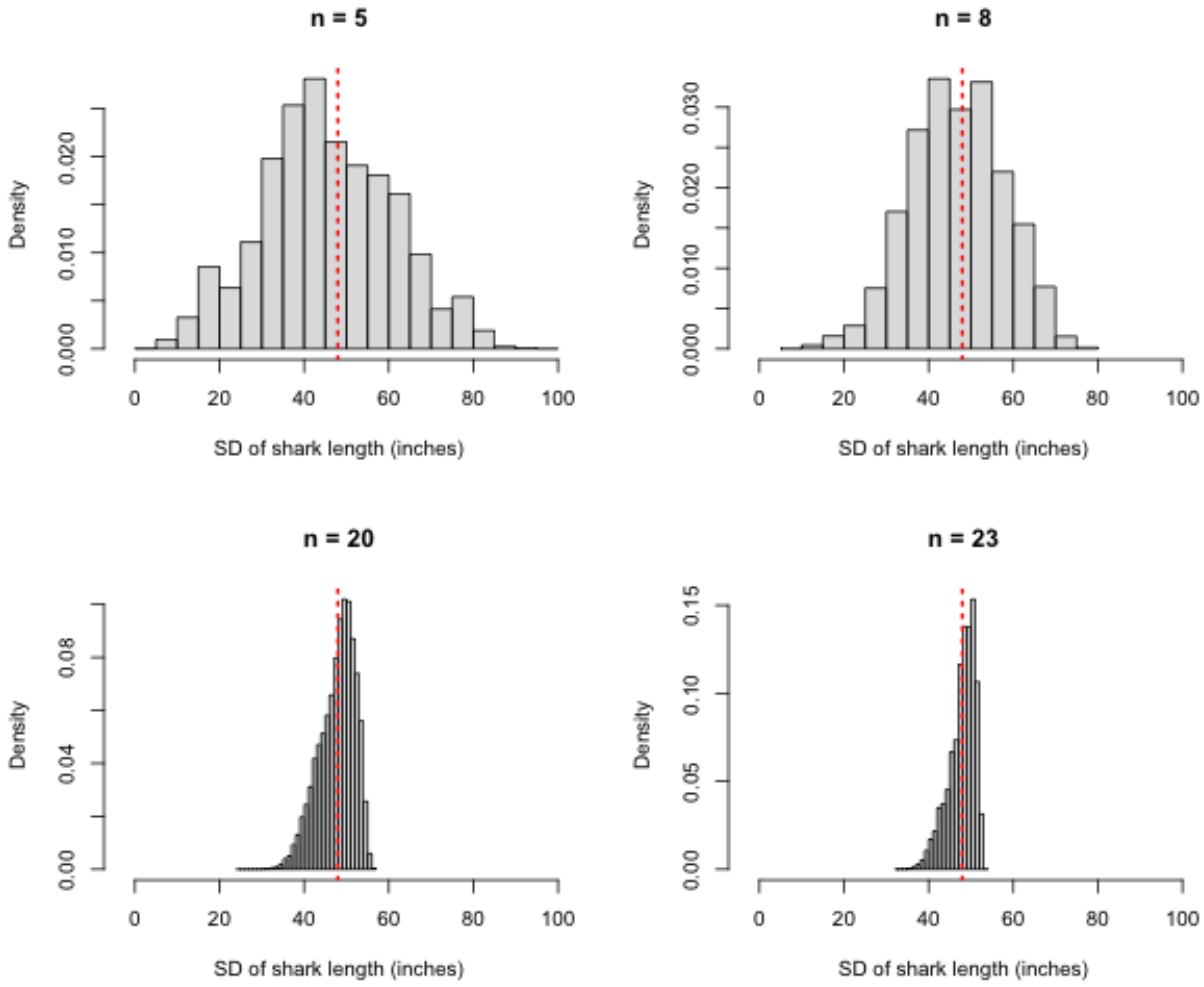


Figure 10: Standard deviations over all possible samples for different sample sizes

- The standard deviations concentrate about the population value as the sample size n increases.
- The histogram is quite skewed.

3.1.3 Comparisons across attributes

To compare different attributes, we use the **relative** absolute sample error. For any $c > 0$, let

$$\mathcal{P}_a^\star(c,n) = \left\{ \mathcal{S} \ : \ \mathcal{S} \subset \mathcal{P}_\mathcal{S}(n) \ \text{ and } \ \frac{\mid a(\mathcal{S}) - a(\mathcal{P}) \mid}{\mid a(\mathcal{P}) \mid} < c \right\}$$

and define the corresponding proportion, for all $c > 0$, and $n \leq N$

$$p_a^\star(c,n) = \frac{|\mathcal{P}_a^\star(c,n)|}{|\mathcal{P}_\mathcal{S}(n)|}$$

- $p_a^\star(c,n)$ measures the consistency of the sample attribute with respect to the *same* population attribute.
- When making comparisons between attributes, we are evaluating each attribute on how well its sample values track its population value on the *same scale*.

Location Attributes

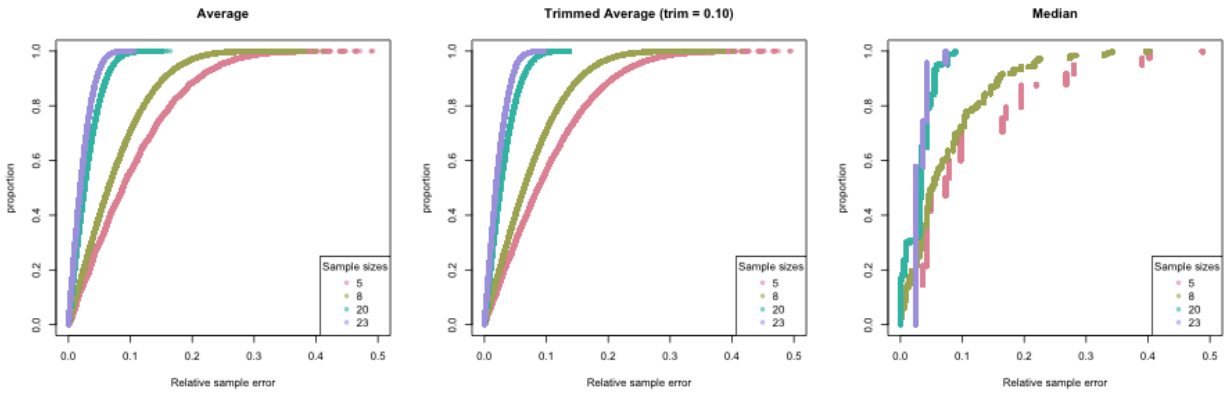


Figure 11: Location Attributes: proportion vs. sample relative error curves over all possible samples for different sample sizes

- The 10% trimmed average performs similar to the average itself for this population (why?).
- However, the trimmed average is slightly less stretched, hence "less error."
- The medians for sample sizes $n = 5$ and $n = 23$ never achieve zero relative error.
  - Why? Note that the population size N is even.
- Unless these are identical, the median for an odd sample size cannot exactly reproduce the population median.

Scale Attributes

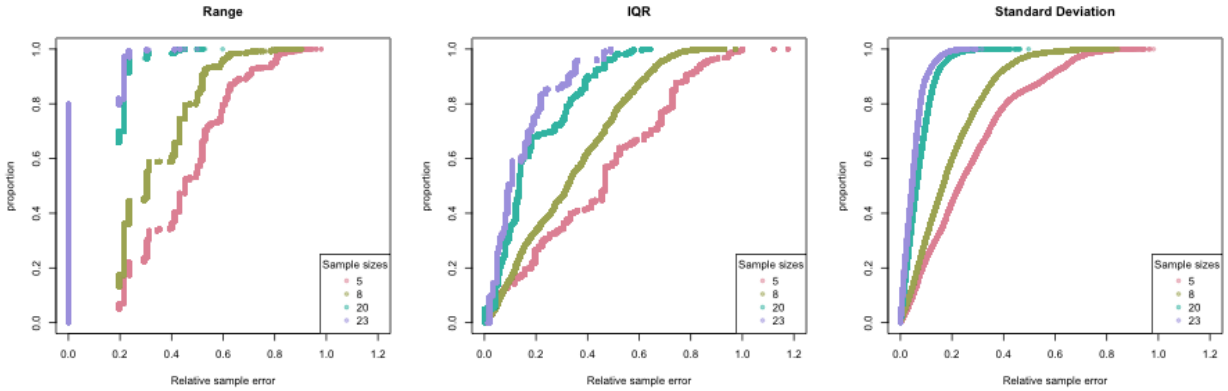


Figure 12: Scale Attributes: proportion vs. sample relative error curves over all possible samples for different sample sizes

- The range has zero sample error for any sample that includes y_{min} and y_{max} in the population.
  - Under the purple vertical line at 0 are the other ones, but shorter.
- Relative error curves for the range are consistently to the left and above those of the interquartile range, so range *outperforms* IQR.
- With the exception of the proportion of zero error samples for the range when $n = 23$, the standard deviation outperforms both the range and the interquartile range.

This particular population: Shark Data

- It is important to note that these findings hold for *this particular population*.
  - To see how things might change dramatically when the population is slightly different, we could introduce a single outlier into the population.

- The "Discovery Channel" has been one of the worst offenders of demonizing sharks with its "shark week".
  - It has even produced fake documentaries to attract ratings.
  - For example, in 2014 the Discovery Channel produced the following film and, though **entirely faked**, passed it off as "documentary evidence" about a supposed 35-40 foot "cunning", "intelligent", and "stealthy" killer great white called **Submarine** Shark of Darkness – Wrath of Submarine. While fake, suppose that a great white shark the size of "submarine" was encountered in Australia waters.

The Shark of Darkness

- We can examine the effect on attributes if we replace a shark with the *Shark of Darkness* in the population.

```
sharksBigSubmarine <- sharks
set.seed(12345564)
replaceShark <- sample(length(popSharksAustralia), 1)
rownameReplaceShark <- popSharksAustralia[replaceShark]
sharksBigSubmarine[rownameReplaceShark, "Length"] <- 480
```

Histogram

- Histograms with and without Shark of Darkness

```
par(mfrow=c(1,2))
### Location estimates
###
hist(sharks[popSharksAustralia, "Length"],
     col=adjustcolor("grey", alpha = 0.5),
     main="", xlab="shark length (inches)",
     xlim = c(0,500), breaks=seq(0, 500, 40)  )
hist(sharksBigSubmarine[popSharksAustralia, "Length"],
     col=adjustcolor("grey", alpha = 0.5),
     main="", xlab="shark length (inches)",
     xlim = c(0,500), breaks=seq(0, 500, 40)  )
```

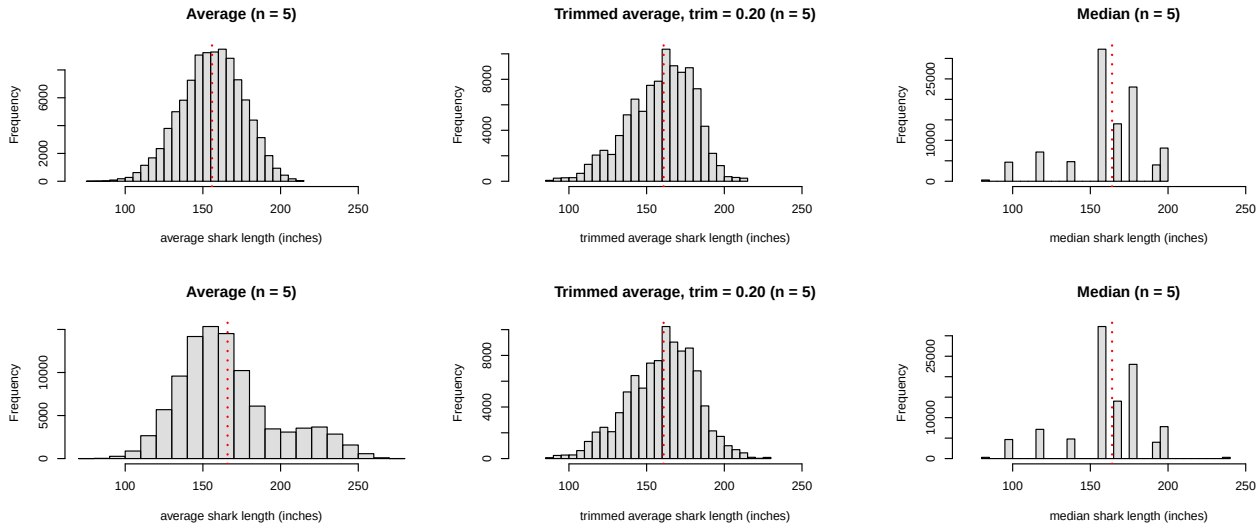
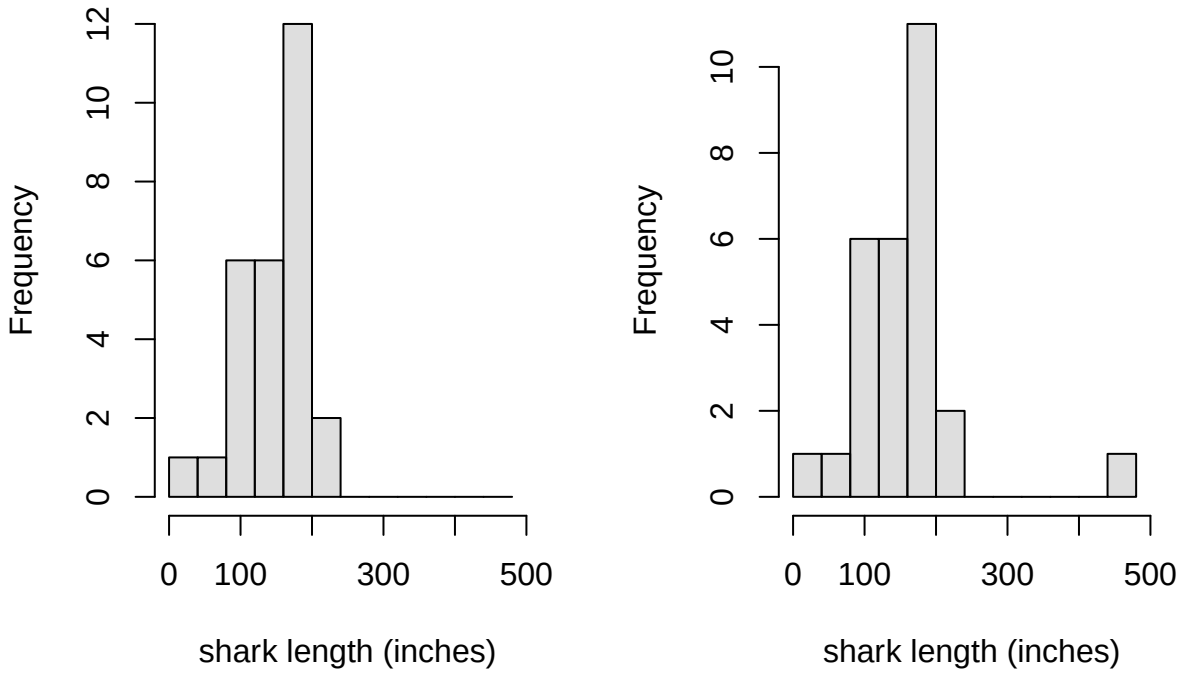


Figure 13: Different location attributes: over all possible samples (n = 5)



Location Attributes

- The upper panel without the Shark of Darkness
- The lower panel with the Shark of Darkness

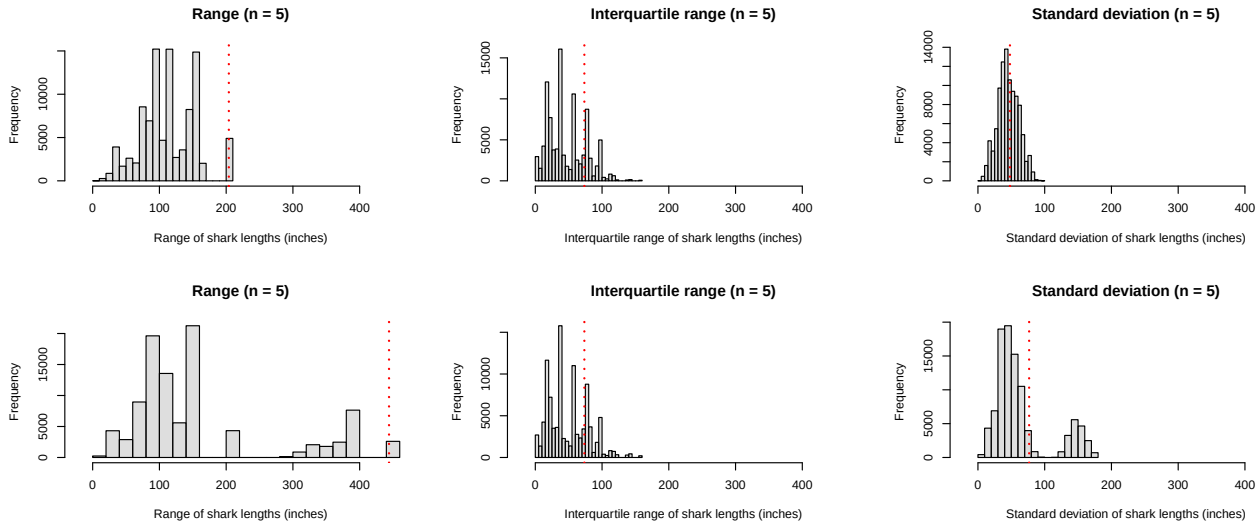


Figure 14: Different scale attributes: over all possible samples (n = 5)

Scale Attributes

- The upper panel without the Shark of Darkness
- The lower panel with the Shark of Darkness