Samples

Contents

3 Samples

Untill now we've assumed that when calculating attributes, we could do so using all of the data (i.e., the *entire* population).

However, it **may not be** possible to calculate an attribute for the population if, for example:

we have the whole pop but...

- the population is too large, or
- the attribute is too complex, or
- we just don't have access to the entire population

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 $a(\mathcal{P})$.

$$\underline{a(\mathcal{S}) = \widehat{a(\mathcal{P})} = a(\widehat{\mathcal{P}})}$$

- The second equality emphasizes that $\mathcal{S}$ as an estimate of $\mathcal{P}$.

When using a sample instead of the entire population, we might consider

- 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**.

$\widehat{a(P)}$

$$\text{sample error} \quad = \quad \underline{a(\mathcal{S}) - a(\mathcal{P})}$$ = difference between estimate and true value

- The nature of this 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 though still conceptually applicable.

Example - Agriculture Data

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

```
agpop <- read.csv("/Users/nstevens/Dropbox/Teaching/STAT_341/Lectures/Data/agpop_data.csv", header=TRUE

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

n=100

Since we have the population we can calculate the differencs between some attributes.

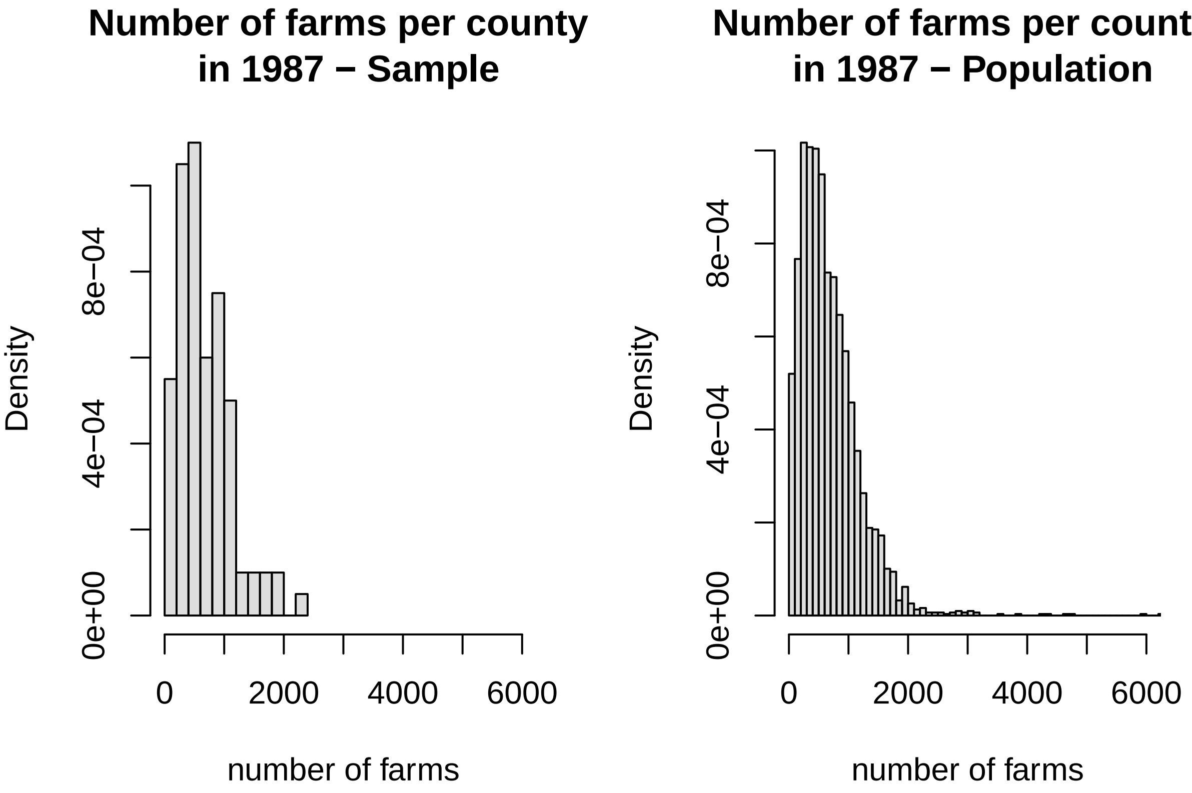
```
c(mean(agpop$farms87[s]) - mean(agpop$farms87),
median(agpop$farms87[s]) - median(agpop$farms87),
sd(agpop$farms87[s]) - sd(agpop$farms87),
IQR(agpop$farms87[s]) - IQR(agpop$farms87) )
```

sample error for the mean
" " " median
" " " standard dev
" " " IQR

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

We can also compare the difference between histograms.

```
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 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. $a(\mathcal{P}) = a(\mathcal{S})$. What happens as n → N ? We hope sample error → 0

- 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

Example - Agriculture Data

Consider what happens to the sample errors of the mean, median, standard deviation and interquartile range as $n \to N$.

$n = 1000$:

```
set.seed(341)
s <- sample(length(agpop$farms87), 1000)
c(mean(agpop$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]   1.513724   6.500000 -55.111237   0.000000
```

$n = 2000$:

```
set.seed(341)
s <- sample(length(agpop$farms87), 2000)
c(mean(agpop$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]   0.5947245   4.0000000 -21.0935978 -10.0000000
```

$n = 3000$:

```
set.seed(341)
s <- sample(length(agpop$farms87), 3000)
c(mean(agpop$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]  2.1473912  1.0000000  0.6149544 -2.5000000
```

$n = N = 3078$:

```
set.seed(341)
s <- sample(length(agpop$farms87), 3078)
c(mean(agpop$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] 0 0 0 0
```

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".

- That said, we avoid this nomenclature because it flies in the face of traditional statistical language and common English usage.
  - However, in some applications (e.g. the bootstrap, which we 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.

Example - Shark Data

- Consider the population $\mathcal{P}$ of great white shark encounters reported from 1999 to 2014.
- This example will be woven througout this section

```
sharks <- read.csv("/Users/nstevens/Dropbox/Teaching/STAT_341/Lectures/Data/sharks.csv", header = TRUE)
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

- There are $N = 65$ such encounters in our population.
  - The table below shows the number of possible samples of a given size n

$\binom{65}{n} =$

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

- 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`).
  - This sub-population contains just $N = 28$ units. There are now only $98,280$ $\binom{28}{5}$ possible samples of size $n = 5$ from this population. This is still a large number, but it's much more manageable.
  - Here are the rows from the full population now constituting our Australian sub-population:

```
### 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 Possible 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)
```

LETTERS "A", "B", "C", "D", "E", ...

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

- Generating all samples ($n = 5$) of the Australia Shark Data
  - the table below shows which units are to be included in the first 5 and last samples.

```
samples <- combn(popSharksAustralia, 5)
N_s <- ncol(samples)
kable(data.frame(first = samples[,1], second = samples[,2],
                 third = samples[,3], fourth = samples[,4],
                 fifth = samples[,5], last = samples[,N_s]))
```

first	second	third	fourth	fifth ···	last
1	1	1	1	1	54
6	6	6	6	6	55
7	7	7	7	7	58
9	9	9	9	9	59
10	11	14	16	18	61

= 98 280th

```
print(N_s)
```

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

A Population of Attributes

- For every sample we can calculate any attribute, e.g. the average shark length.
  - We can calculate the attribute (mean) on all possible samples as follows:

```
### 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"])})
```

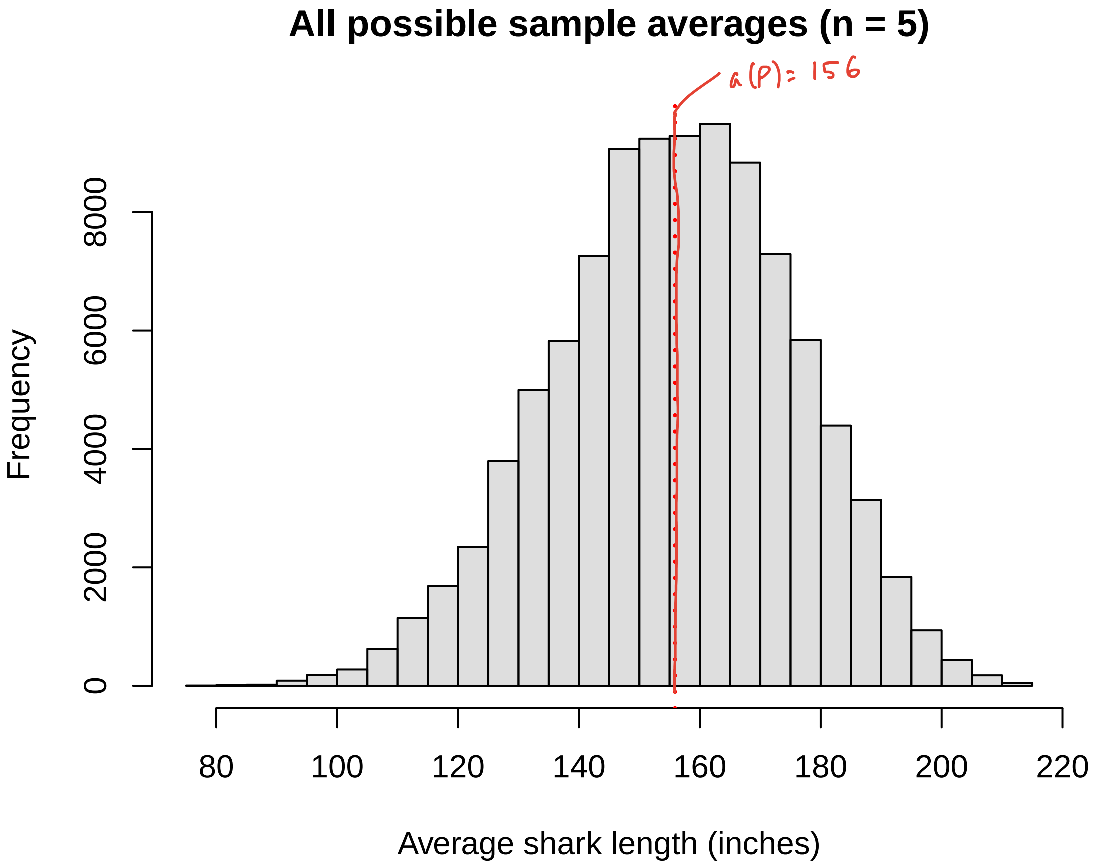
- The average in the first 5 and last samples is:

```
avesSamp[c(1:5, N_s)]
```

```
## [1] 142.6 146.6 129.8 142.2 142.2 196.8
```

- We now have a population of attributes (in this case a population of sample means).
    - We can calculate attributes on this population to summarize it.

- For example, the histogram of the sample attributes:



- Comments:
    - 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.
    - Sample averages are concentrated around the true population average
    - Dispersion about the true population is bell-shaped and symmetric.

- We can also calculate a numerical summary of the sample attributes:

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

- Comments:
    - Half of the samples will produce an average shark length between 142 and 170 inches.

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

Sample Error

- The sample error for a sample $\mathcal{S}$ of size n is

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

- In general, we can calculate the sample error for all possible samples
- Assume again we are interested in the average length (in inches) of great white sharks encounters in Australian waters. Here are the first 5 and last sample errors.

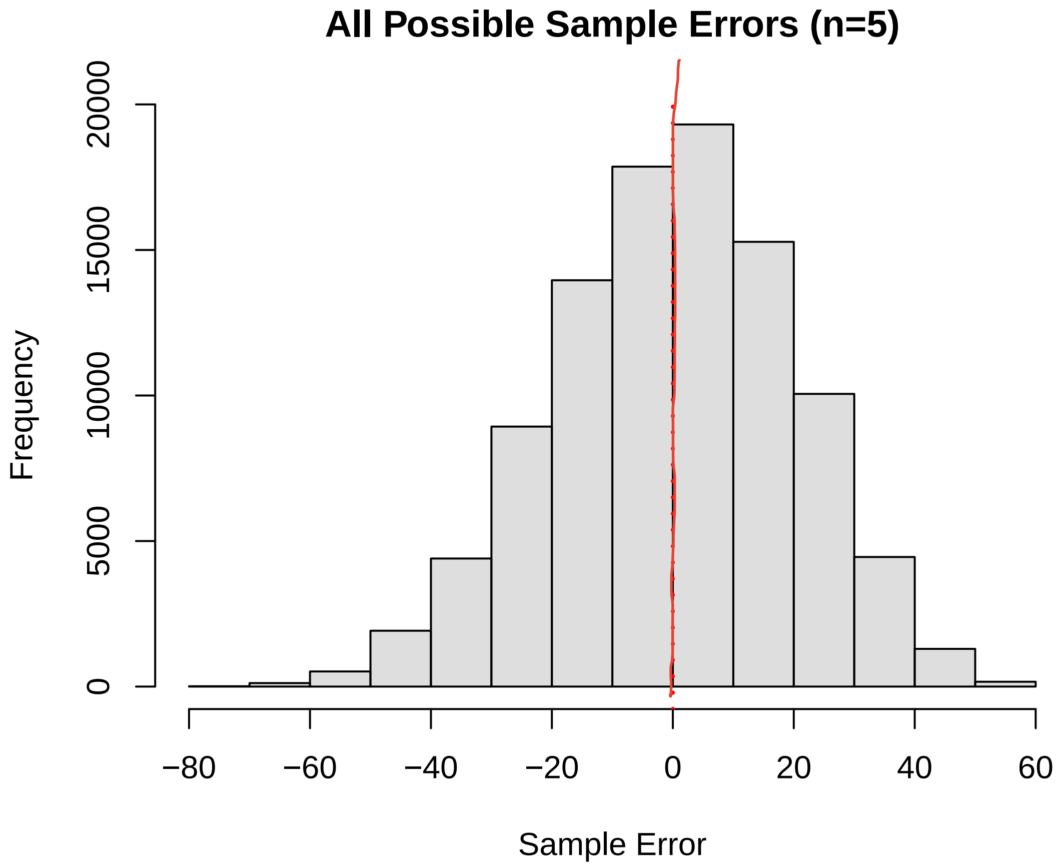
```
sampleErrors <- avesSamp - avePop
sampleErrors[c(1:5,N_s)]
```

```
## [1] -13.292857  -9.292857 -26.092857 -13.692857 -13.692857  40.907143
```

- The sample errors range from -77 to 59 inches.
- Here is numerical summary of the sample errors:

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

- And here is a histogram of them:



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

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

  where N_s (= 98,280 here) is the number of possible samples.

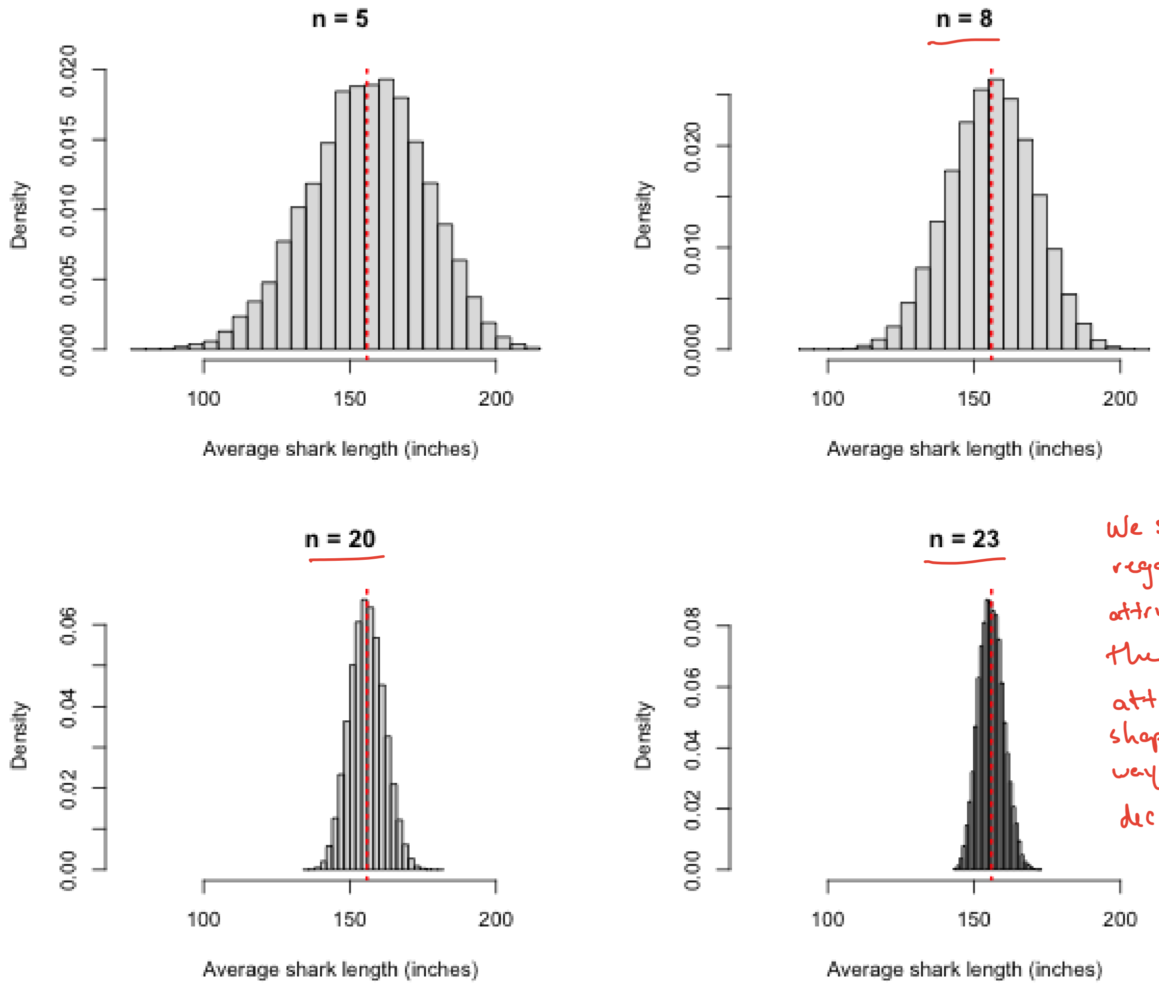
  - For the average shark length, the average sample error was actually `round(mean(avesSamp) - avePop,5)` = 0.
  - At least for this attribute, the sample error is zero on average.

3.1.1 Consistency and the Effect of Sample Size

The nature of sample error depends largely on the sample size.

- As the sample size increases, the sample approaches the population
- Attribute values will concentrate even more around the population value

- This is illustrated below in the case of the average shark length attribute



$N = 28$

We still find that regardless of n, sample attributes cluster around the true population attribute in a bell-shaped and symmetric way, but the dispersion decreases as n increases

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 this particular attribute.

- To quantify this concentration we could look at

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

absolute difference between sample attribute and population attribute

for some $c > 0$

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

- Consider a population $\mathcal{P}$ of size $N < \infty$.

  - For each n, we can construct the set of all possible samples.

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

all possible samples of size n

- For any $c > 0$,

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

all samples for which absolute sample error < c

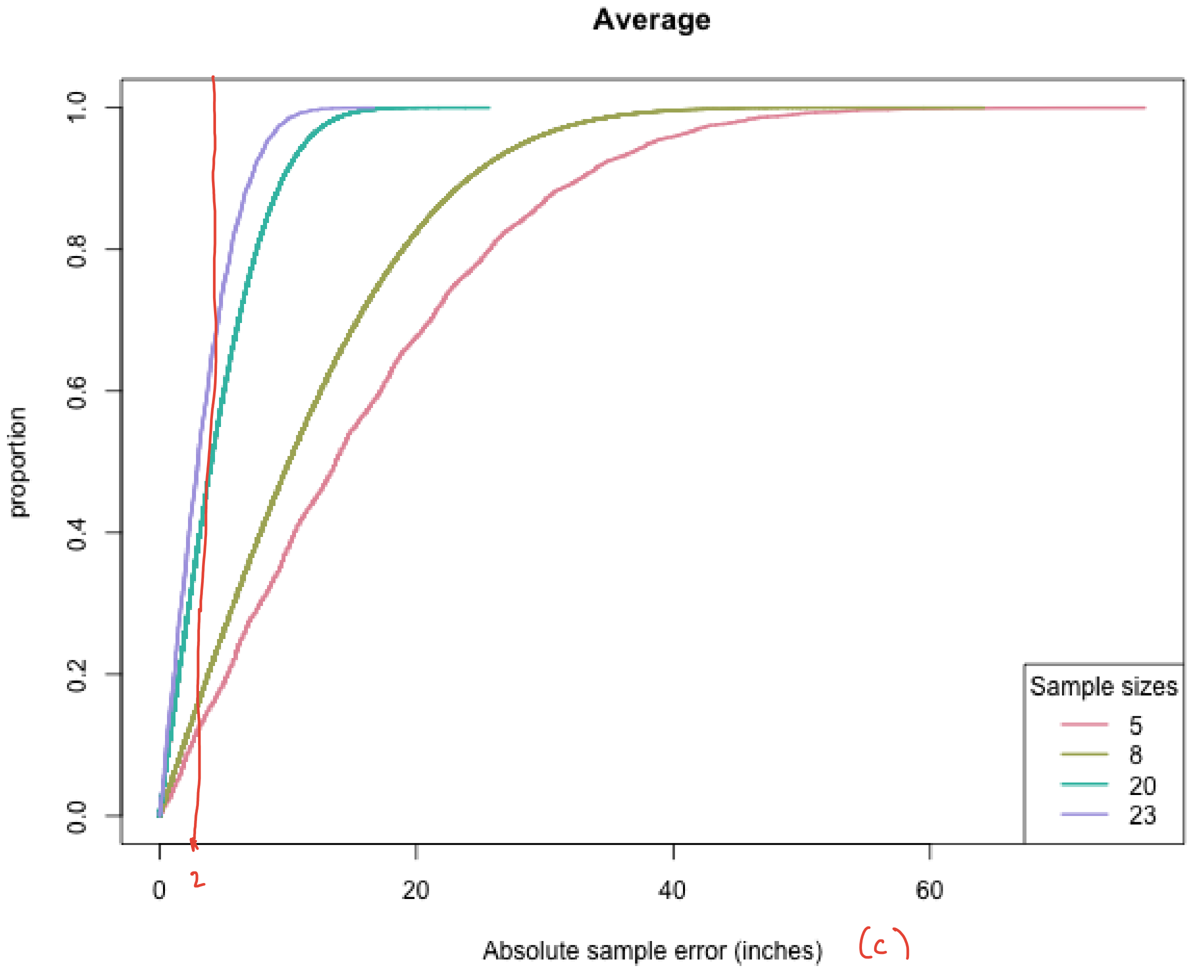
and define the proportion

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

|·| = cardinality (i.e. # of elements)

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

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



more quickly for small c

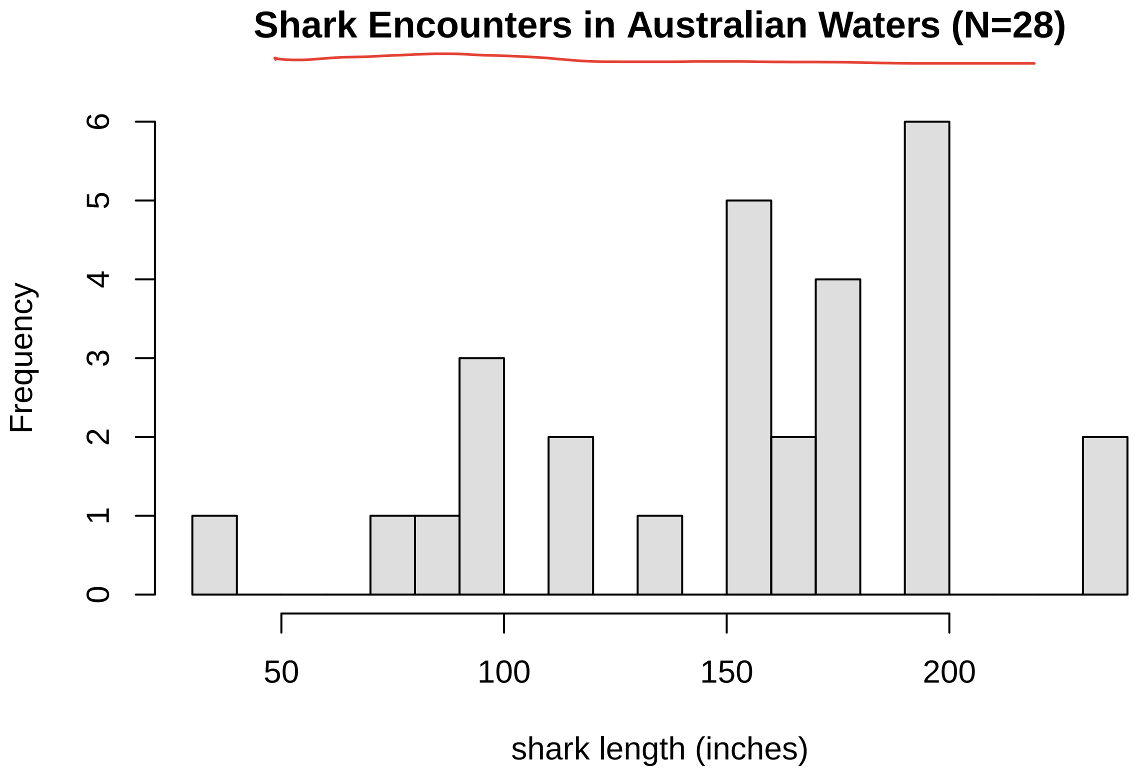
- **Note** *this* notion of **consistency** is different and separate from Fisher consistency.

As n increases the proportion $p_a(n,c) \rightarrow 1$ and so the bulk of the sample attributes cluster tightly around the true population value

3.1.2 Extensions to Other Attributes

As before, we will focus on the lengths of sharks encountered in Austrialian as our population ($N = 28$). Here is a histogram of that population.

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



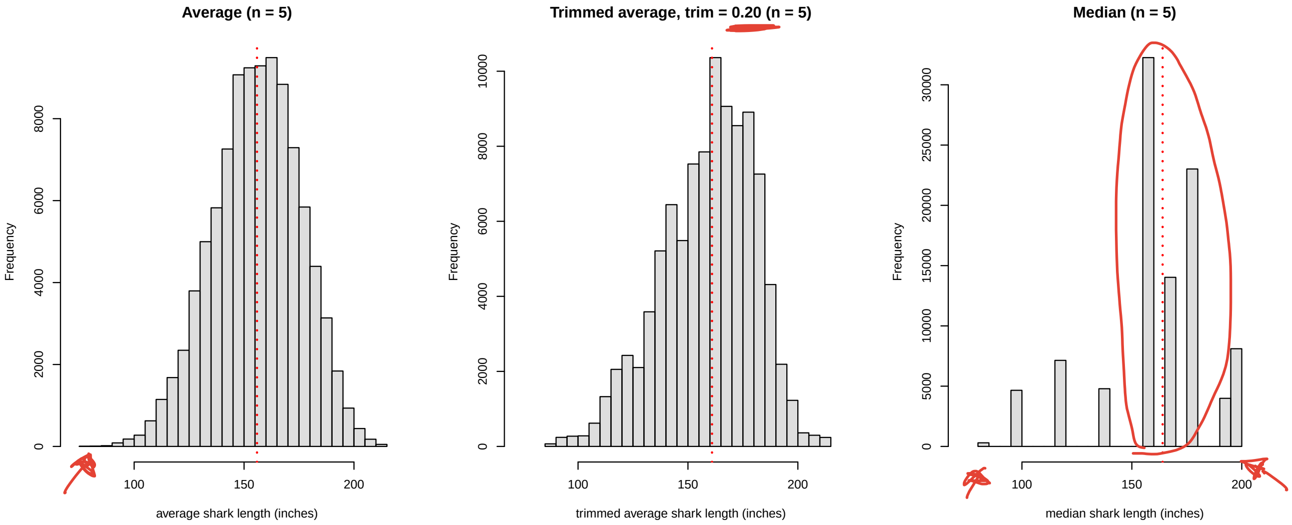
When studying consistency here, we will focus on two types of attributes:

- Location attributes (measures of center)
  - Mean
  - Trimmed Mean: The 100p% trimmed mean is the mean of the middle 100(1-2p)% of the sorted values
  - Median

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

Location attributes

- The location attributes for all possible samples of size $n = 5$.
  - Note that these are all plotted on the same scale to aid comparisons.
  - The value of the attribute calculated on the whole population is marked with red dotted line.



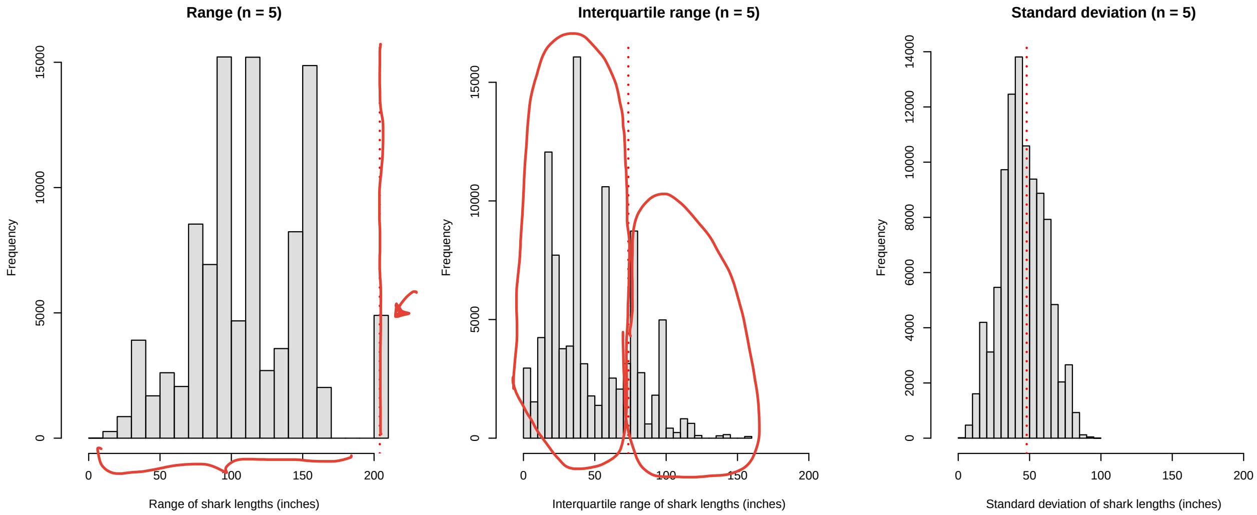
- **Mean:** as before.

- **Trimmed Mean:** behaves similarly to the mean, though less symmetric.

- **Median:** quite different. When there are an odd number units in the sample (like $n = 5$), the median will be one of the observed sample values.
  - 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 for the means.

Scale Attributes

- The location attributes for all possible samples of size $n = 5$.
  - Note that these are all plotted on the same scale to aid comparisons.
  - The value of the attribute calculated on the whole population is marked with red dotted line.



- **Range:**
  - Sample values are quite far from the population range, considerably underestimating its value. (Why?)
  - Sample errors would be mostly negative – except ones in the rightmost bar. (Which are those?)

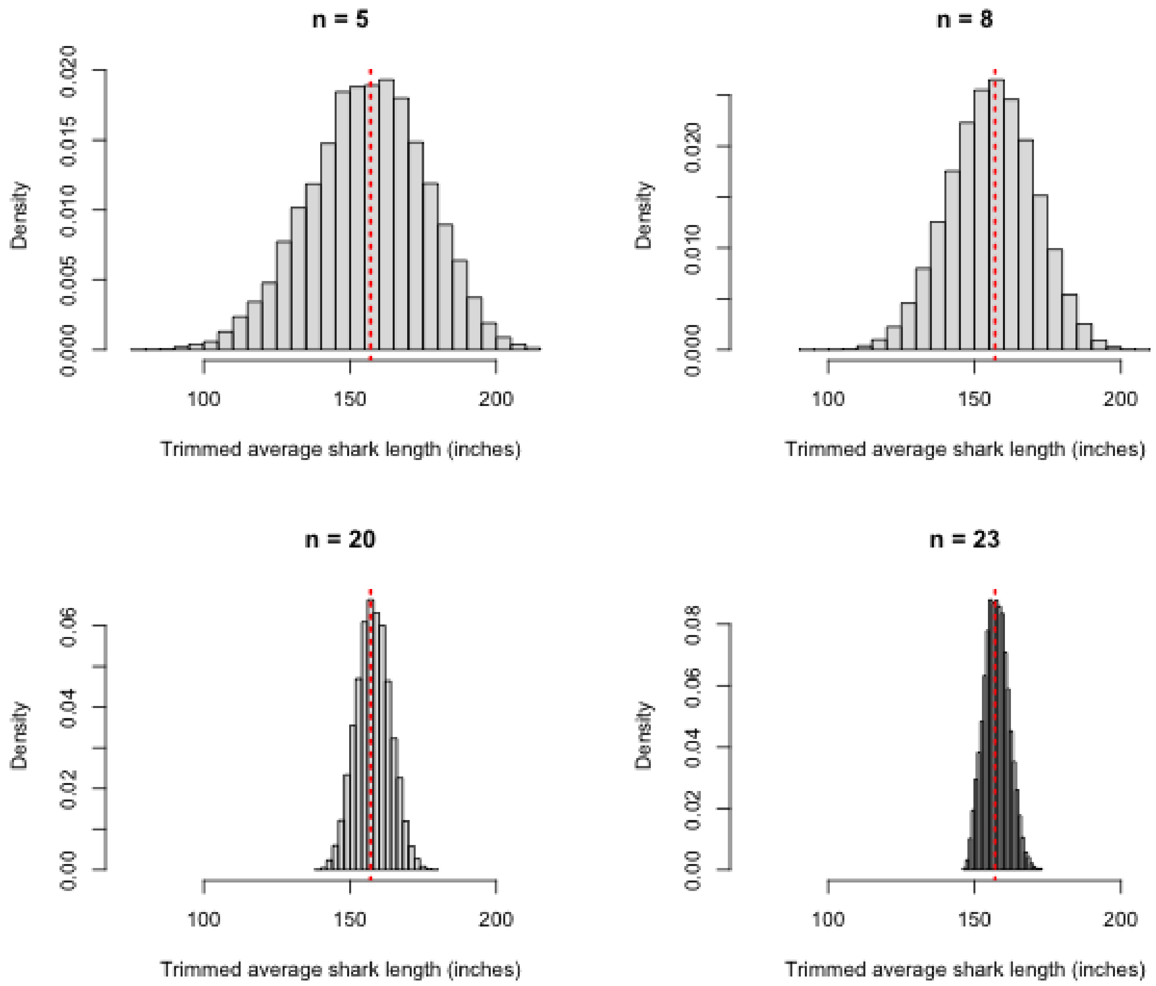
- **Interquartile range:**
  - Performs much better than the range (both positive and negative sample errors.
  - However, the population IQR appears to be far more frequently underestimated than over estimated.

- **Standard deviation:**
  - Behaves much more like an average: sample values concentrate roughly symmetrically (compared to the IQR) about the population value.

Sample Size

Let's now investigate the effect of sample size on concentration for each of these other attributes:

Trimmed Mean

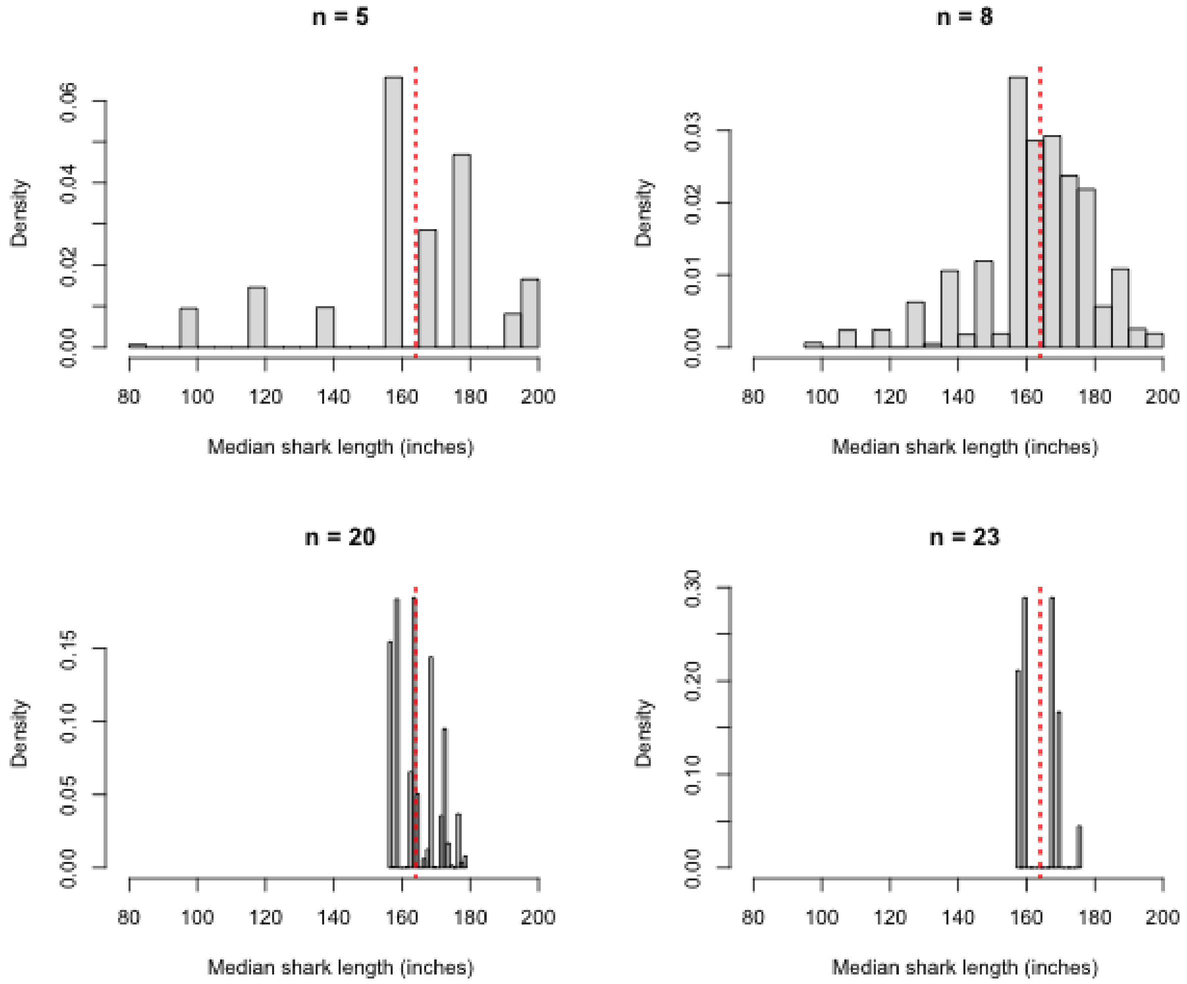


- The trimmed mean behaves much like the ordinary mean.

- Note the effect of increasing n on concentration

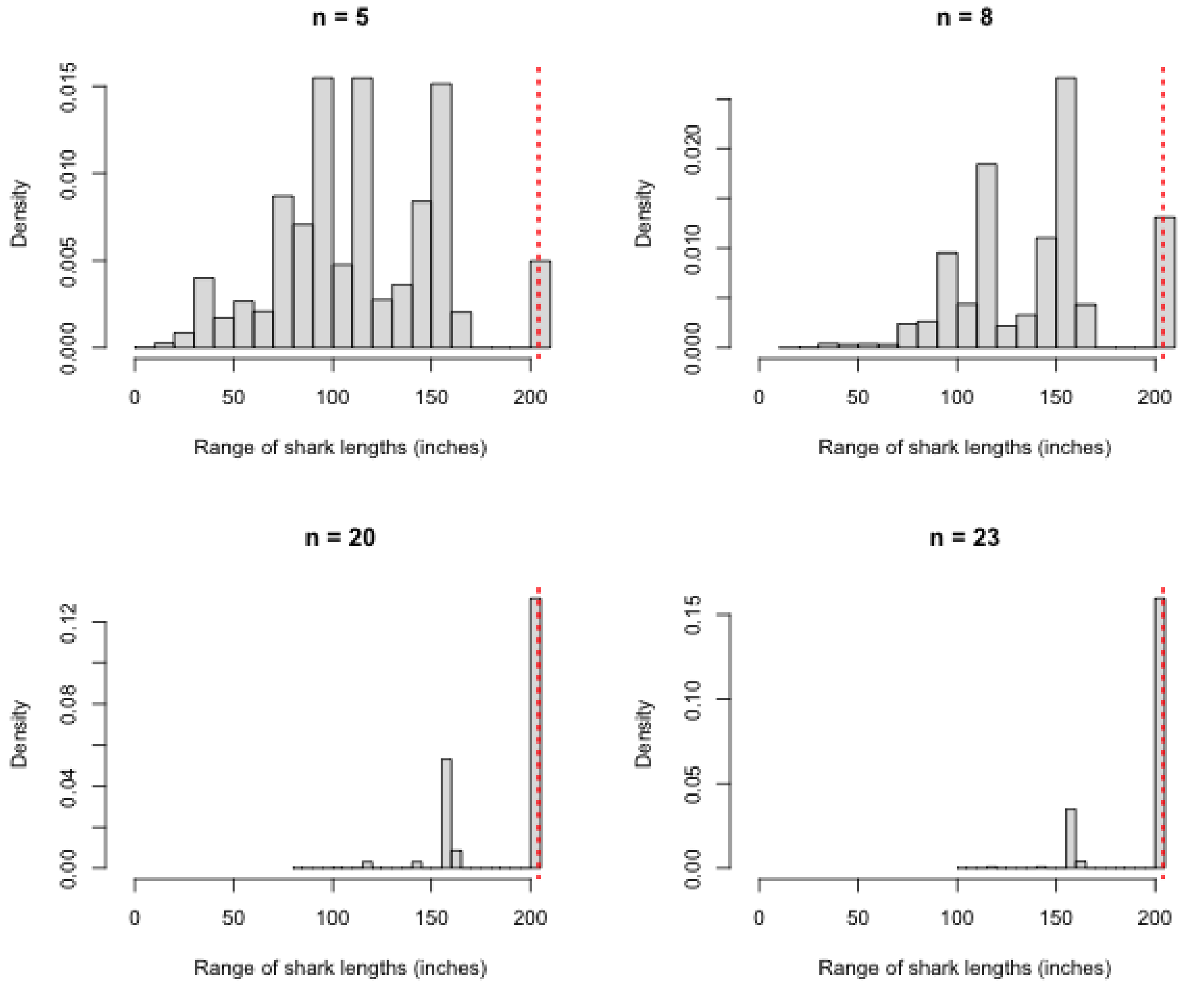
it makes the histograms less disperse

Median



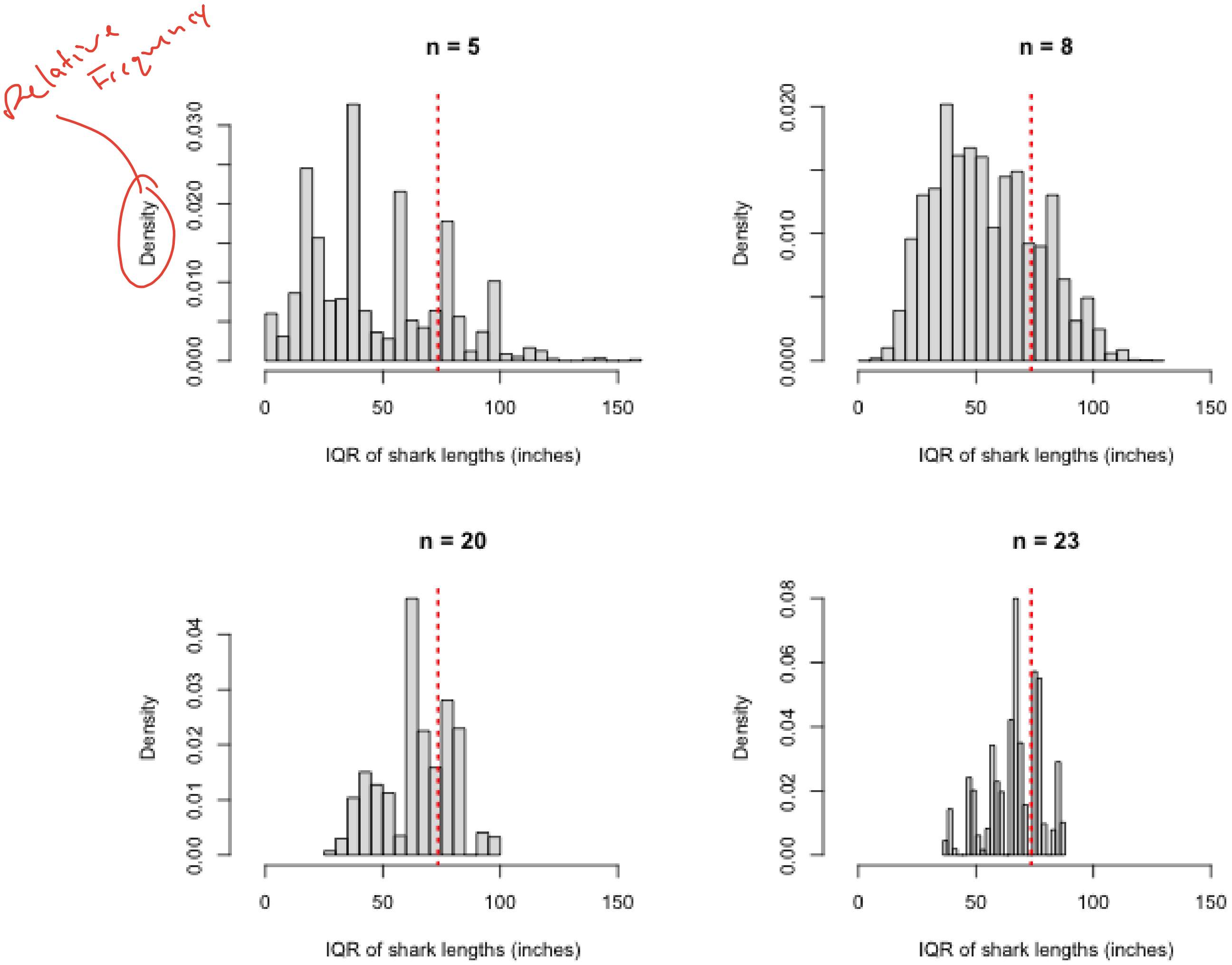
- There is a greater variety of possible values when the sample size n is even. (Why?)
- As the sample size n increases, there is a greater concentration of the sample values about the population values.

Range



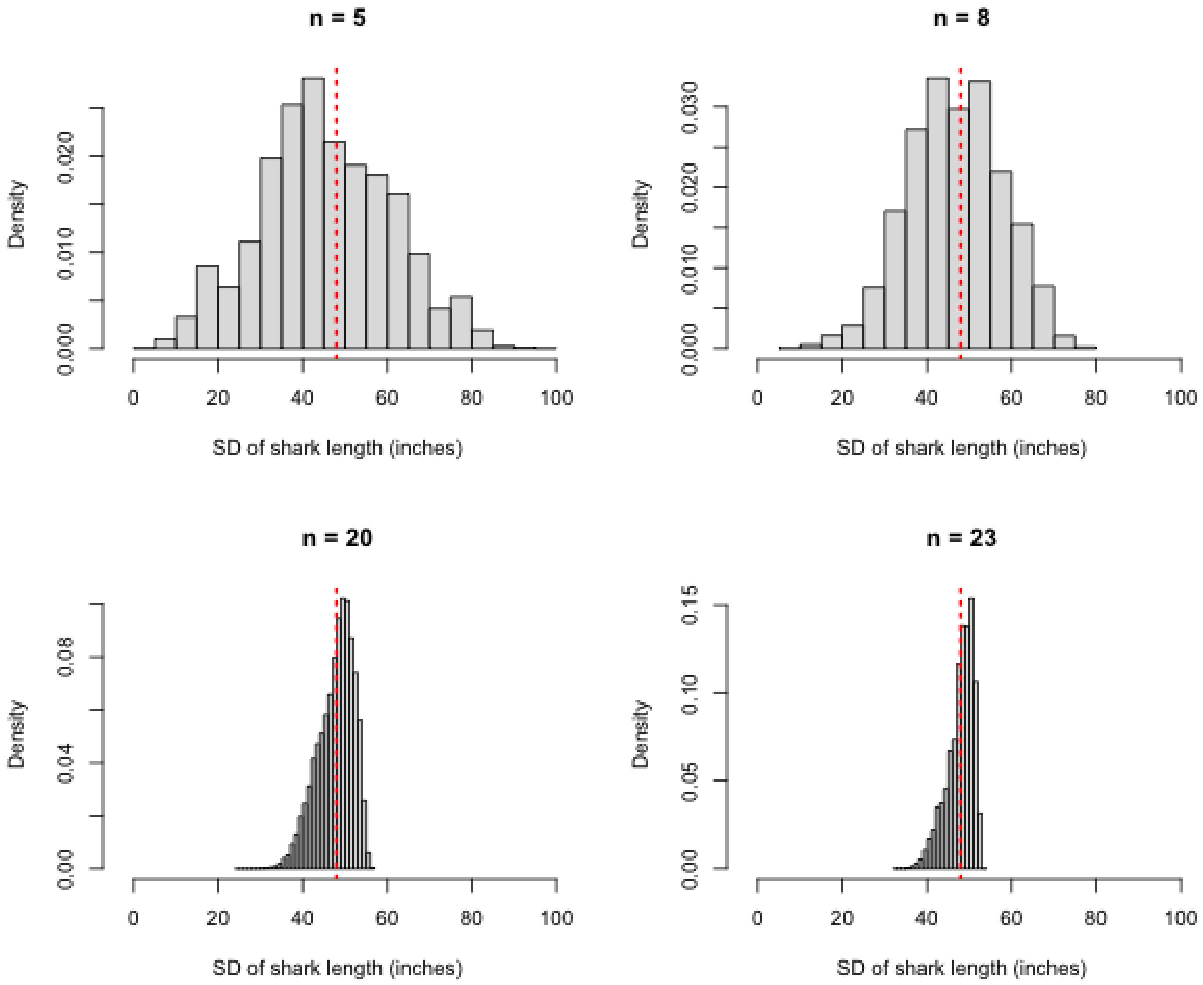
- The range shows a consistent underestimation of the population value. The average sample error will always be negative.
- As the sample size n increases, more samples will contain both $y_{(1)}$ and $y_{(N)}$ and so will match the population value of the range.

Interquartile Range



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

Standard Deviation



- The sample standard deviations concentrate about the population value as the sample size n increases.
- The histogram is quite skewed (compared to, for example, the histograms for the means).

3.1.3 Comparisons across attributes

Previously we defined consistency in terms of **absolute** sample error. This allowed us to evaluate the impact of sample size on concentration.

However, if we want to compare *different* attributtes, we use the **relative absolute** sample error. For any $c > 0$, let

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

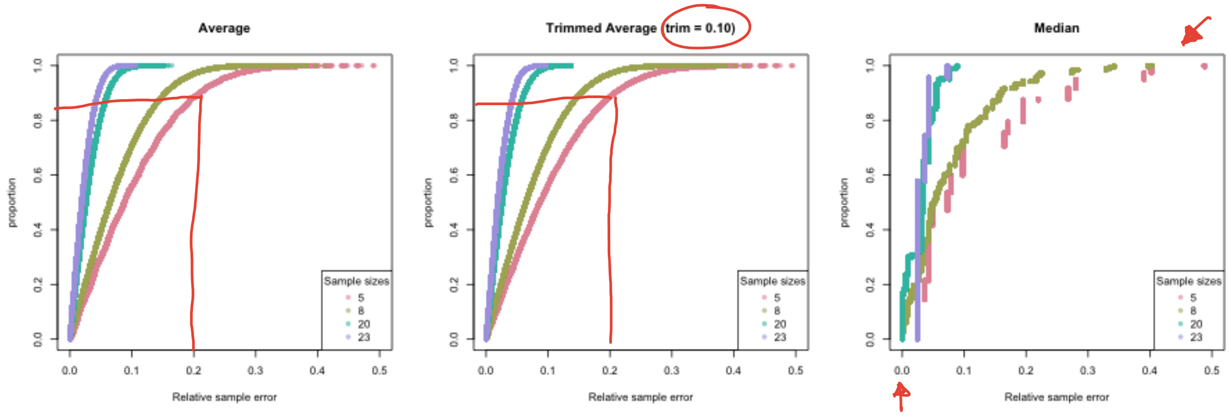
the set of samples whose relative absolute sample error < c and define the corresponding proportion, for all $c > 0$, and $n \leq N$

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

- $p_a^\star(c,n)$ 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

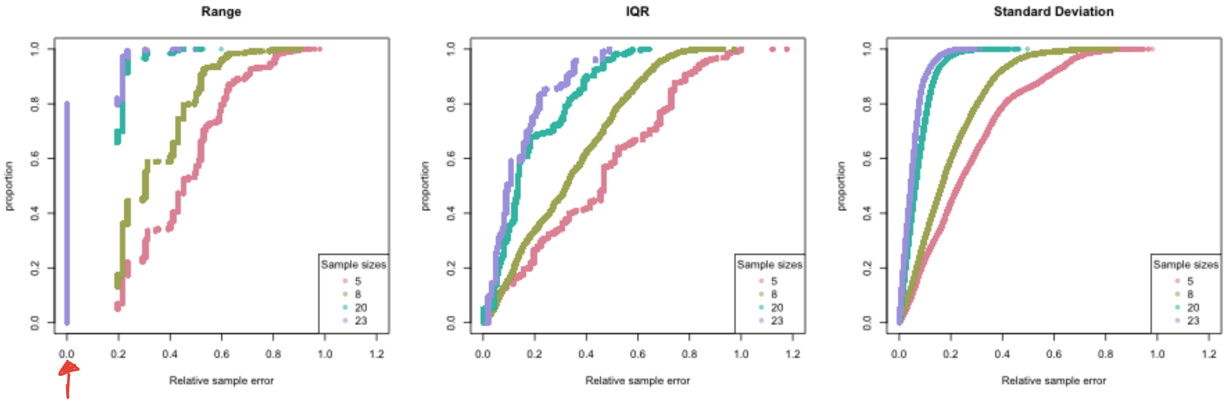
- For each of the location attributes we've been considering, we plot the proportion $p_a^\star(c,n)$ versus the relative sample error for different sample sizes.



- The 10% trimmed mean performs similarly to the mean for this population. (Why?)
- However, the trimmed average is slightly less stretched to the right, hence less error.
- The medians for sample sizes $n = 5$ and $n = 23$ never achieve zero relative error.
  - Note that the population size N is even.
  - The median for an odd sample size can never exactly reproduce the population median (unless the middle two values in the population are identical).

Scale Attributes

- For each of the scale attributes we've been considering, we plot the proportion $p_a^\star(c,n)$ versus the relative sample error for different sample sizes.



- The range has zero sample error for any sample that includes $y_{(1)}$ and $y_{(N)}$ from the population.
  - Note, sample sizes other than $n = 23$ have zero sample error, just far fewer and so they're hidden.
- The relative error curves for the range are consistently to the left and above those of the interquartile range, so range the *outperforms* IQR in terms of consistency.
- In terms of consistency the standard deviation outperforms both the range and the interquartile range.
  - Except when samples are large the range *can* be more consistent (as we see in the case of $n = 23$).

DISCLAIMER: This was only the 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 Australian waters.

The Shark of Darkness

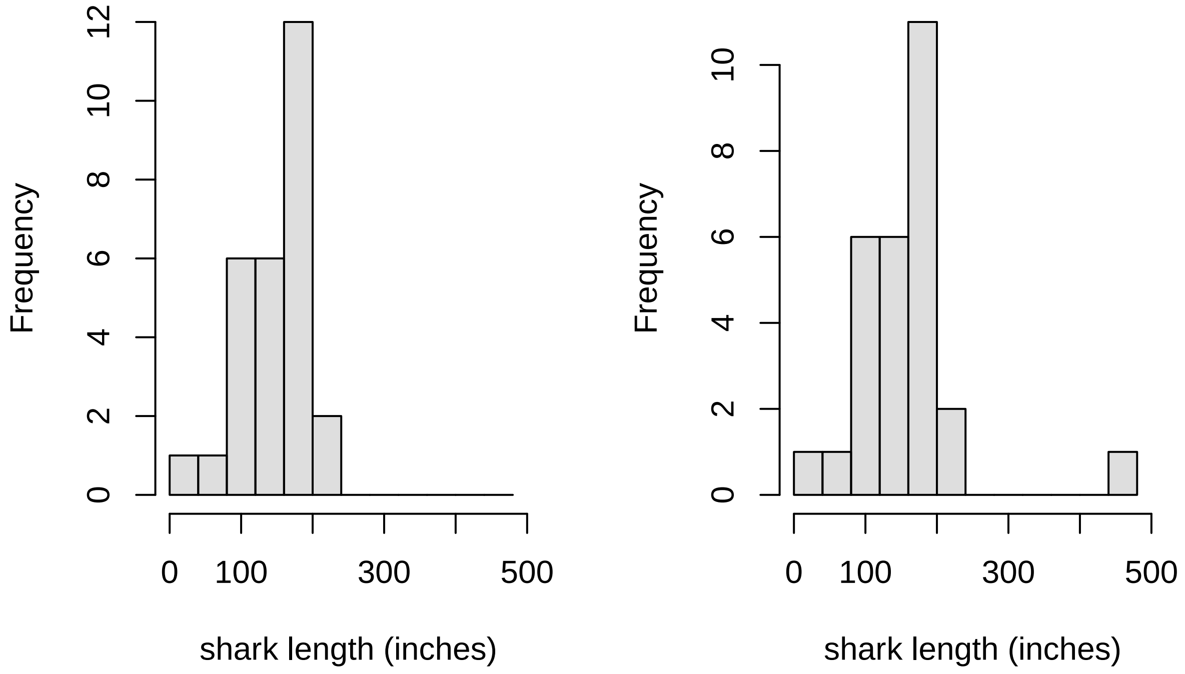
- We can examine the effect on attributes if we replace a shark with the *Shark of Darkness* in the population.

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

Histogram

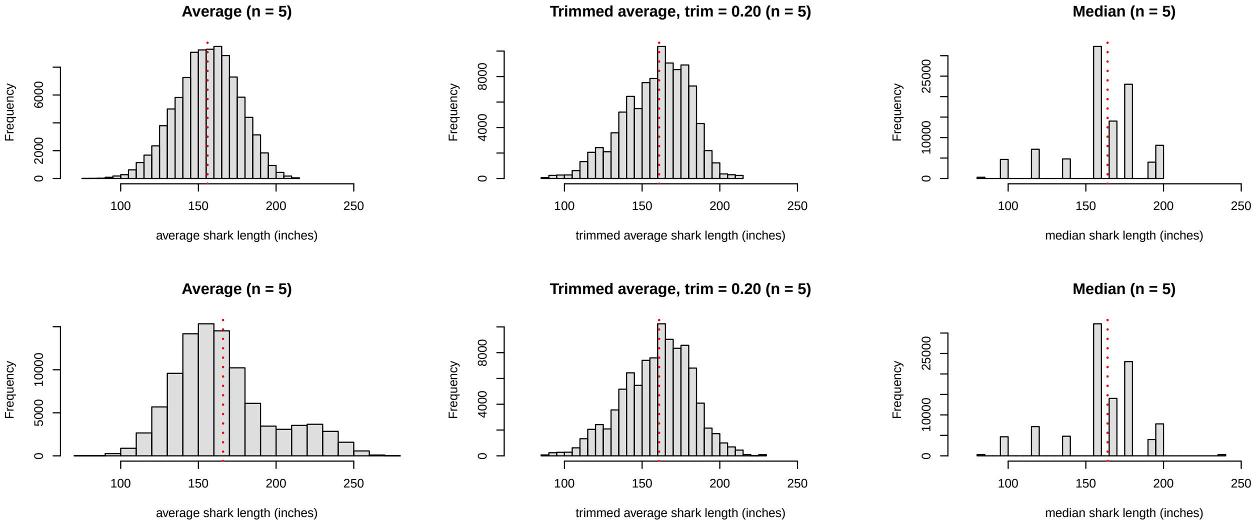
- Histograms of the population with and without the Shark of Darkness

```r
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)  )
```



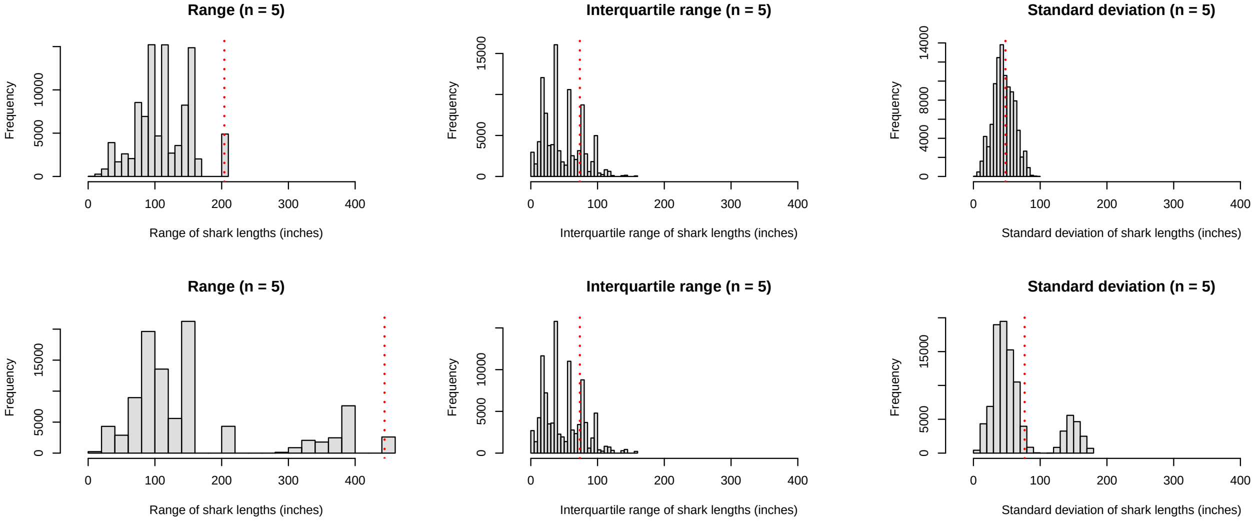
Location Attributes

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



Scale Attributes

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


Range (n = 5)
Interquartile range (n = 5)
Standard deviation (n = 5)
Frequency
Range of shark lengths (inches)
Interquartile range of shark lengths (inches)
Standard deviation of shark lengths (inches)
Range (n = 5)
Interquartile range (n = 5)
Standard deviation (n = 5)
Frequency
Range of shark lengths (inches)
Interquartile range of shark lengths (inches)
Standard deviation of shark lengths (inches)