Samples

Samples

- It may not be possible to calculate an attribute for the population $\mathcal{P}$. For example,
  - *we might not have access to the entire population,*
  - *the population is too large,*
  - *or the attribute too complex.*
- If only a **sample** or subset $\mathcal{S}$ of $n << N$ is available.
  - *Then the attribute $a(\mathcal{S})$ calculated based on this sample is an **estimate** of its population counterpart $a(\mathcal{P})$.*

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

a(S) is an estimate of a(P) => a(S) = â(P)

S is the estimate of P

=> a(S) = a(P̂)

  - *The second equality emphasizes that $\mathcal{S}$ as an estimate of $\mathcal{P}$.*
- Two things we might consider are
  - *sample error, and*
  - *Fisher consistency.*

as n -> N

we expect a(S) -> a(P)

Sample error

- Sample error

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

  - *Any difference between the actual values of the estimate $a(\mathcal{S})$ and the thing being estimated (the **estimand**) $a(\mathcal{P})$ is an **error**.*
  - *The error will depend on the sample and the attribute.*

- Quantifying error; how good is the sample / what attribute we are looking at

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

Sample error Examples - Agriculture Data

- Differences among some attributes

index

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

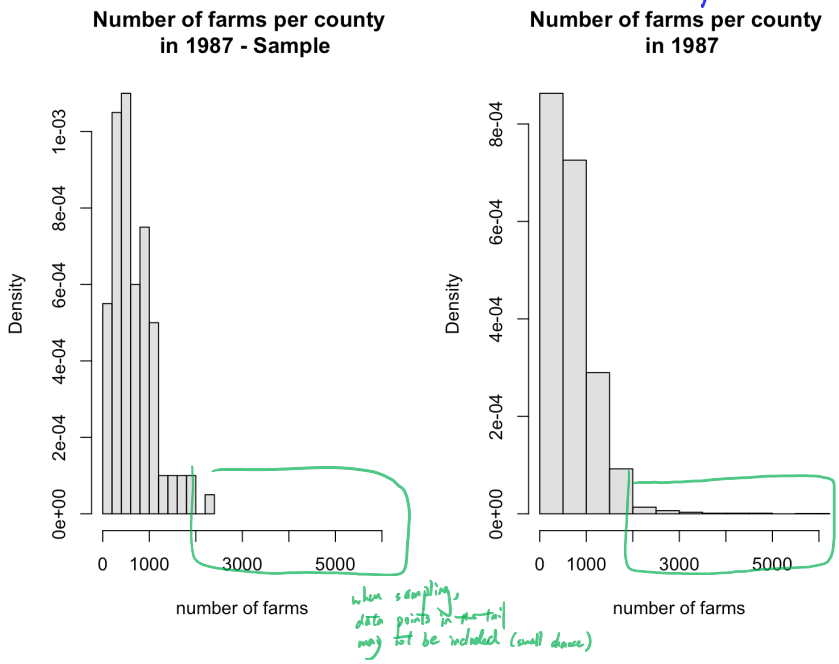
center

spread

⇒ under-estimate (not always the case)

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

Be aware of the unit. e.g. for variance not the same unit as the sample data



when sampling, data points in the tail may not be included (small chance)

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. $\alpha(\mathcal{P}) = \alpha(\mathcal{S})$.
  - *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 $\mathcal{S}$ could be considered a population itself and might even sensibly be called a "sample population". /psuedo population
  - *Such nomenclature, while arguably legitimate, does unfortunately fly in the face of traditional statistical language and common English usage – it is to be avoided therefore and will not be used here.*
  - *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.*

Sample error usually decreases as sample size ↑

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.
- 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.*

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

Shark Data

- 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 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
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

```
### 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"
```

- 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.

$\binom{28}{5}$

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

- For the Australia Shark data

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

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

First five samples and the last sample of 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
9	9	9	9	9	59
10	11	14	16	18	61

Sample Error

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

$$a(S) - a(\mathcal{P}_{Australia}) = \frac{1}{n}\sum_{u \in S} y_u - \frac{1}{28}\sum_{u \in \mathcal{P}_{Australia}} y_u.$$

- The sample error for all possible samples.

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

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

```
sampleErrors[1:15]
```

```
##  [1] -13.2928571  -9.2928571 -26.0928571 -13.6928571 -13.6928571
##  [6]   5.9071429  -1.8928571   2.1071429   0.7071429 -16.4928571
## [11]  -1.8928571  -9.6928571   0.3071429  -5.8928571   2.7071429
```

Average Sample Error

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

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

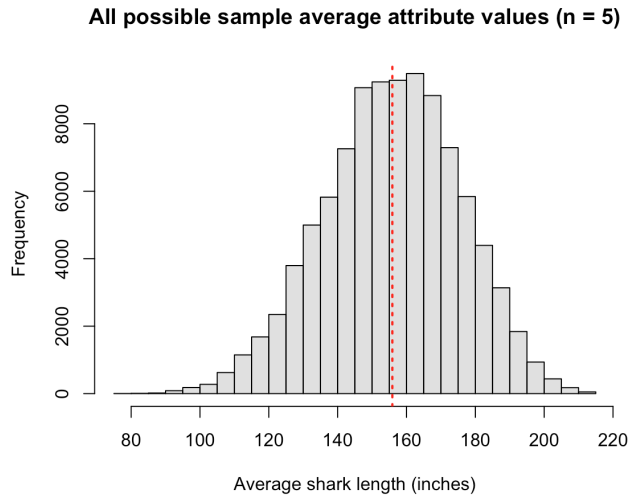
where N_s (= 98,280 here) is the number of possible samples $\mathcal{S}_i$.

- *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.* just in this case

Exercise: Prove that the average sample error must be zero when $a(\mathcal{P})$ is the arithmetic average. (Hint: How many times does each y_u appear in a sample?)

Histogram of the Sample Attribute

- The red dotted line is the value of the attribute on the population, 156.
  - *The sample error ranges from -77 to 59 inches.*



No probability here
Cannot apply CLT
still in population
N does not go to infinity
case n = N = 28

All possible samples: great white encounters in Australia

n = 5
N = 28
⇒ $\binom{28}{5}$ = 98,280

Comments

- There are a few samples that produce a value far from the population value.
- Concentration near the population value.
- *nearly* symmetrically about the population value.
- A numerical summary of the sample averages is

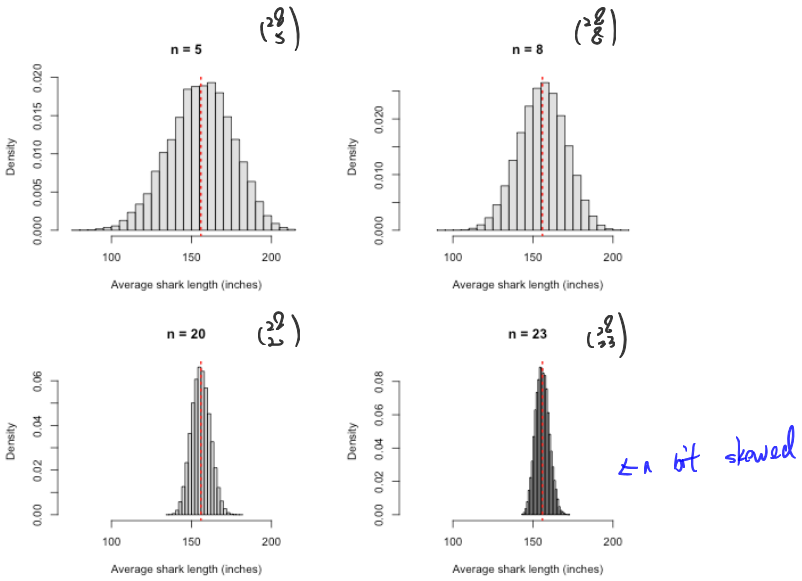
skew to left

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

50%

Effect of increasing sample size



← a bit skewed

All possible samples for different sample sizes

$\frac{\sum_{u \in s} y_u}{n} = \bar{y}_s$ As sample size ↑ get closer to true value

$Var(\bar{y}_s) = \frac{1}{n} Var(y_p)$ as $n \uparrow$, var ↓ => more concentrate

Question: Which one is better?
The proportion of sample which has less error

Consistency

- This concentration indicates some kind of **consistency** for the particular attribute here (viz. the arithmetic average).
  - *To quantify this concentration we could we*

absolute error

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

max amount of error

  for some $c > 0$
  - *Then we could count the proportion of samples that satisfy this.*

There should be a trend. i.e. as sample size ↑, the concentration should get closer to true value

Consistency

- For a population $\mathcal{P}$ of size $N < \infty$.
  - *Then for each n, we can construct the set of all possible sample.* a set of samples of size n

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

  - *Then for any $c > 0$,* among the set, those sets satisfy the abs. err condition

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

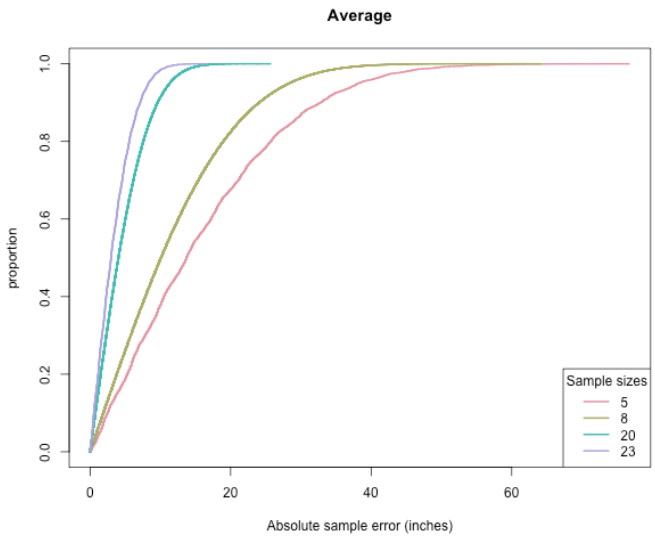
  and define the proportion attributes

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

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

- From the plot with varying sample size, we noticed that for a fixed $c > 0$, $p_a(c, n)$ increases with n.

plotting as a function of c for different n

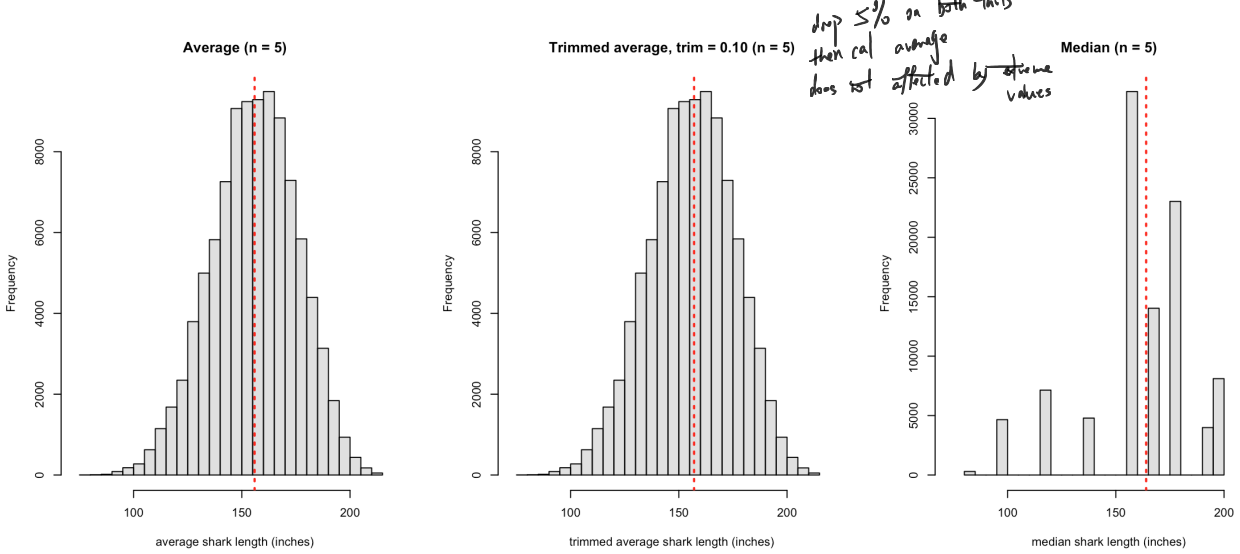


For a given sample error, as sample size ↑ more proportion falls in the range of error.

Increasing function.

Location attributes

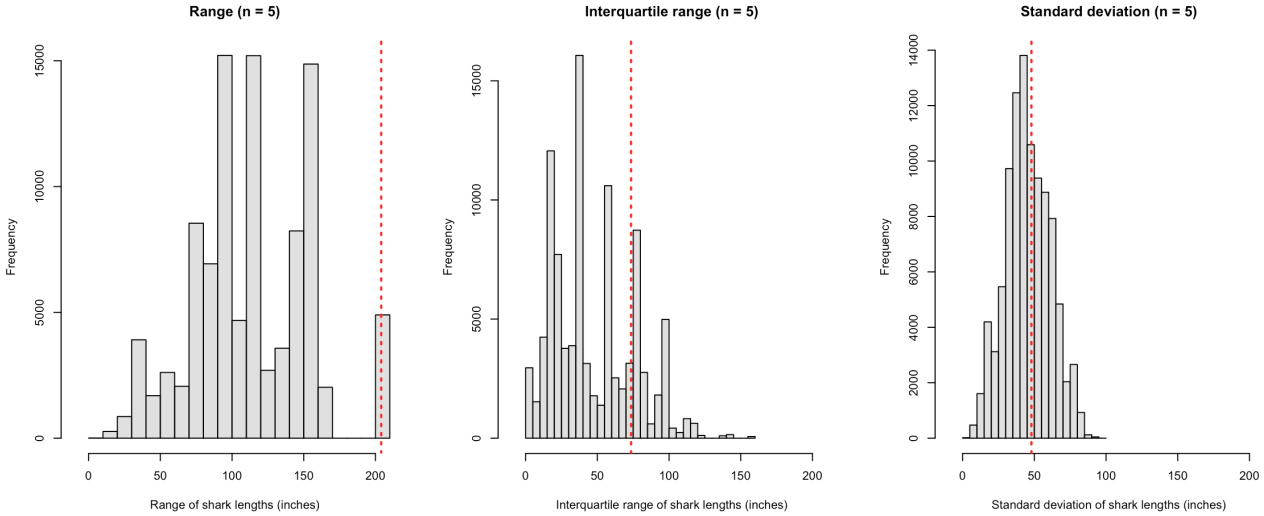
- The location attributes for samples of size $n = 5$.
  - *Note that these are all plotted on the same scale to aid the comparisons.*
  - *The value of the attribute calculated on the whole population is marked.*



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

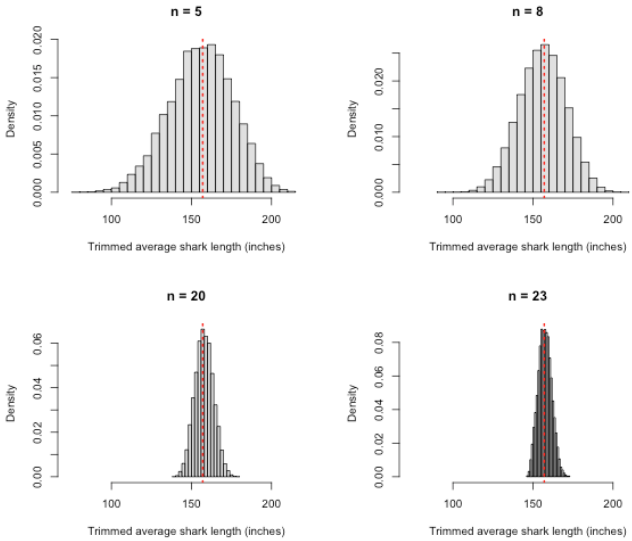
Scale Attributes

alike trimmed average



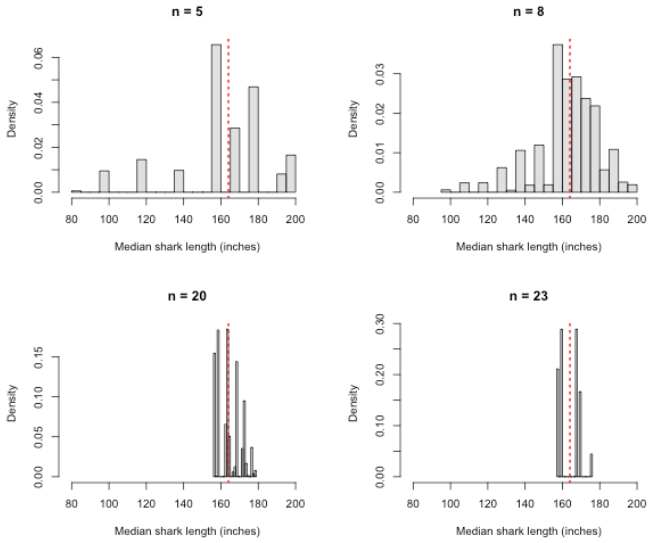
Different scale attributes: over all possible samples (n = 5)

Trimmed Average



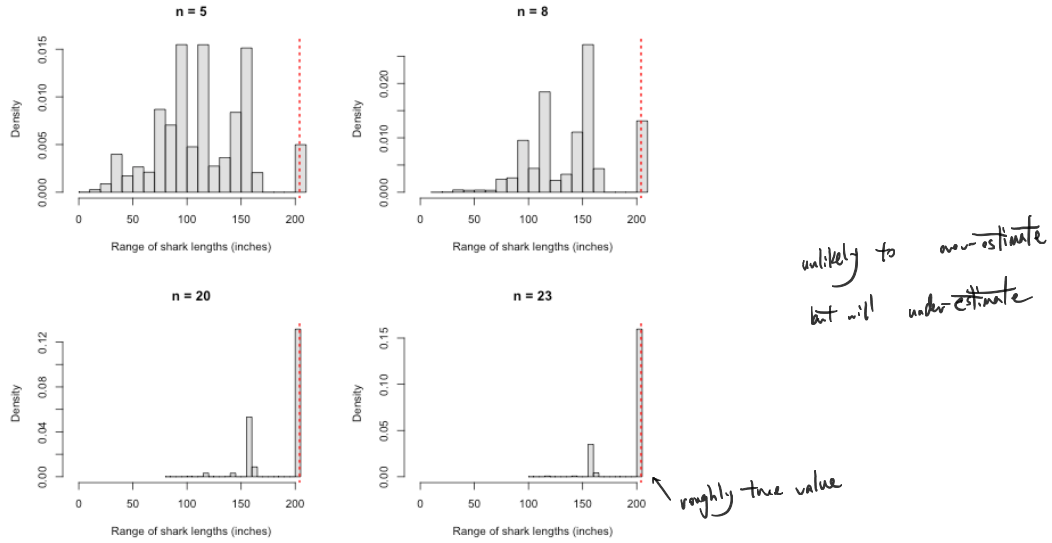
Trimmed averages (trim = 0.10) over all possible samples for different sample sizes

Median



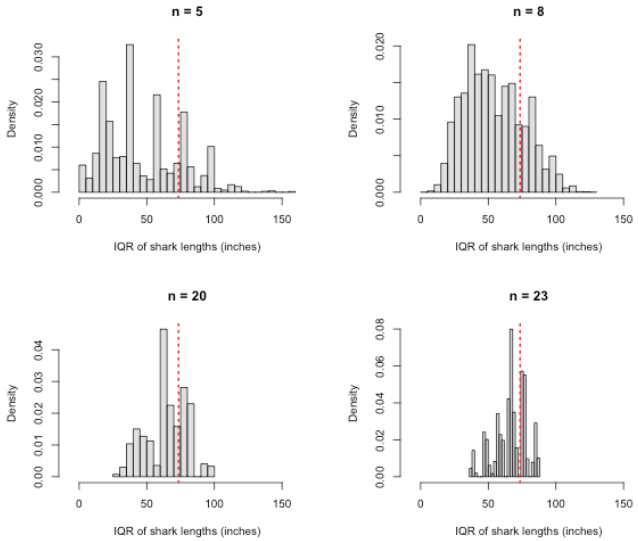
Medians over all possible samples for different sample sizes

Range



Ranges over all possible samples for different sample sizes

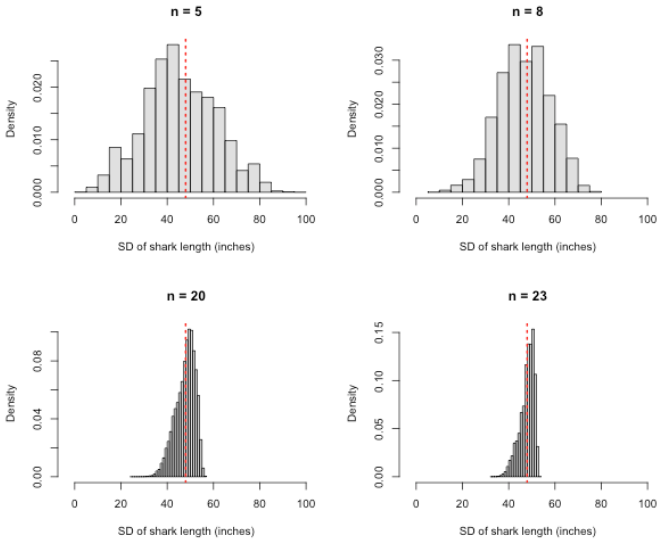
Interquartile Ranges



In this population, slightly to left
=) more likely under-estimate

Interquartile ranges over all possible samples for different sample sizes

Standard Deviation



better concentration

Standard deviations over all possible samples for different sample sizes

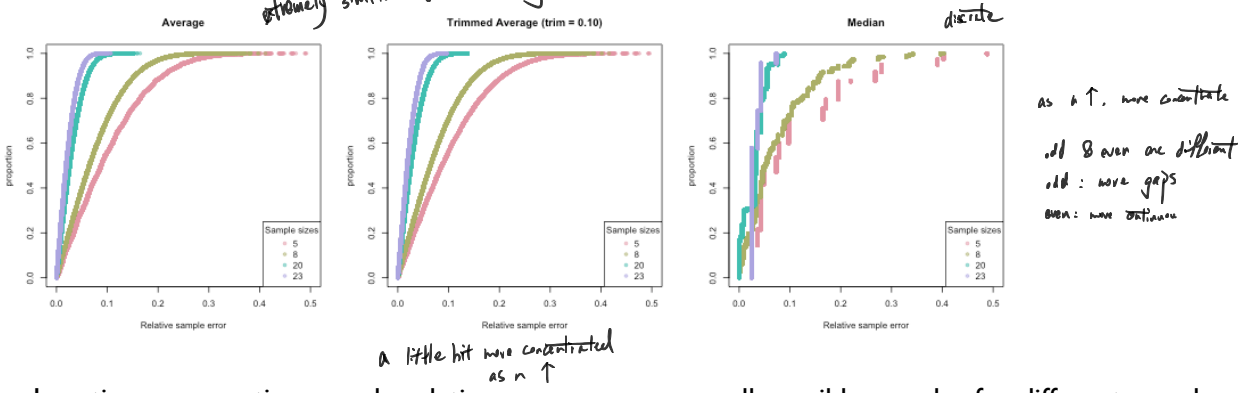
Comparing Attributes

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

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

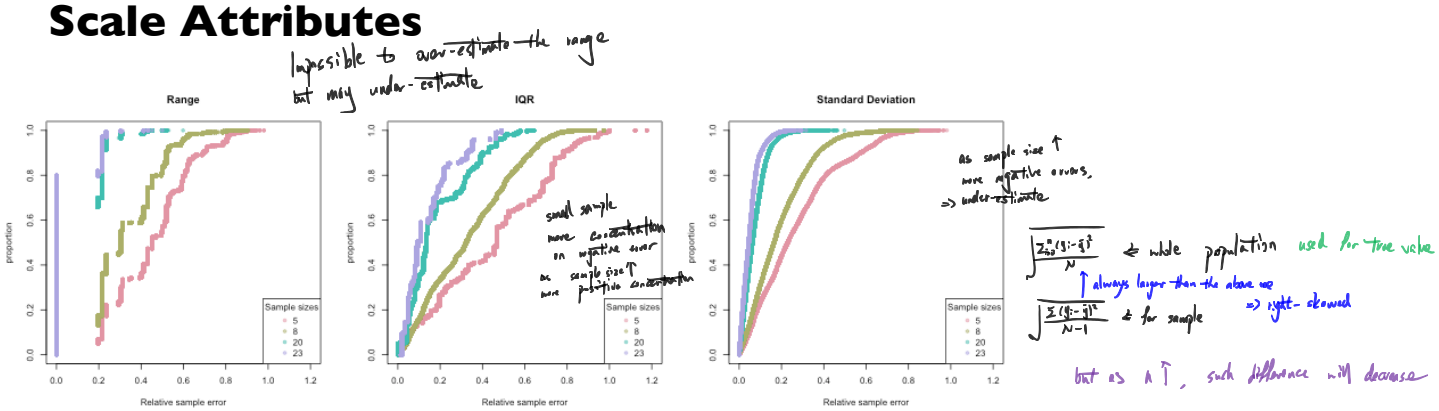
and define the corresponding proportion, for all $c > 0$, and $n \leq N$

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



Locations: proportion sample relative error curves over all possible samples for different sample sizes

Scale Attributes



Scales: proportion sample relative error curves over all possible samples for different sample sizes

This is using the abs error
the x-axis would change
but the shape won't
The proportion is the real matter

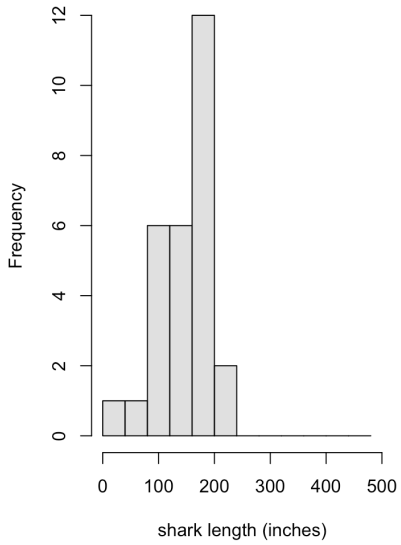
This particular population: Shark Data

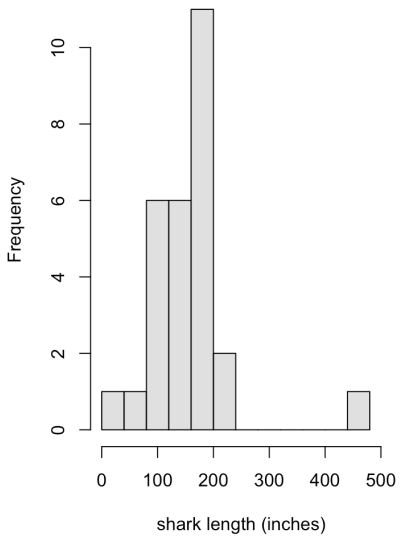
- It is important to note that these findings hold for *this particular population*.
  - *To see how things might change dramatically when the population is slightly different, we could introduce a single outlier into the population .*

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

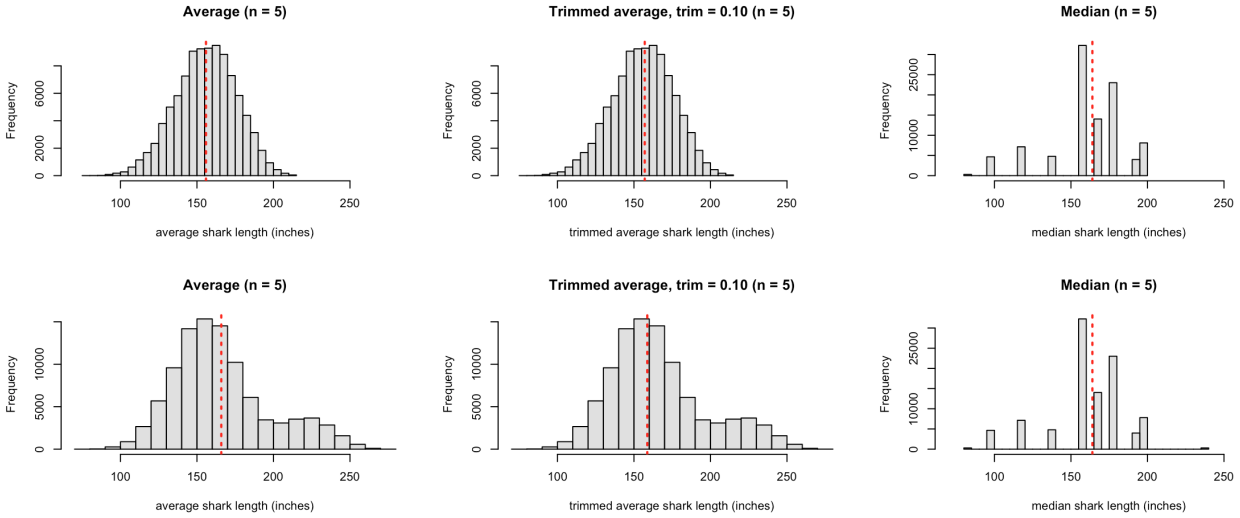
```
sharksBigSubmarine <- sharks
set.seed(12345564)
replaceShark <- sample(length(popSharksAustralia), 1)
rownameReplaceShark <- popSharksAustralia[replaceShark]
sharksBigSubmarine[rownameReplaceShark, "Length"] <- 480
```

Histogram with and without Shark of Darkness



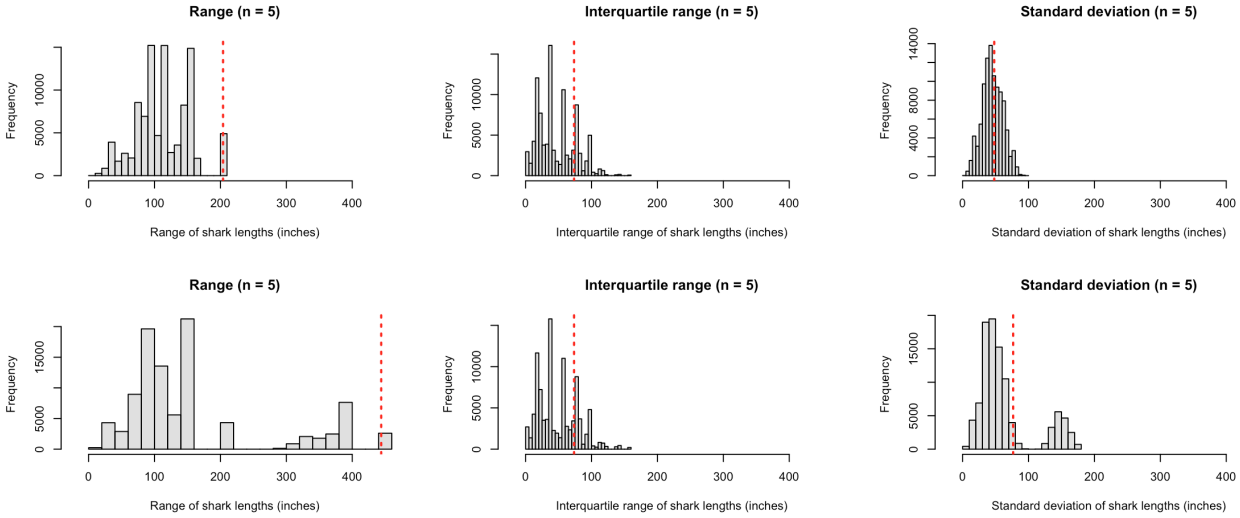


Location Attribute with and without Shark of Darkness



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

Scale Attributes with and without Shark of Darkness



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