Random intervals and coverage probabilities

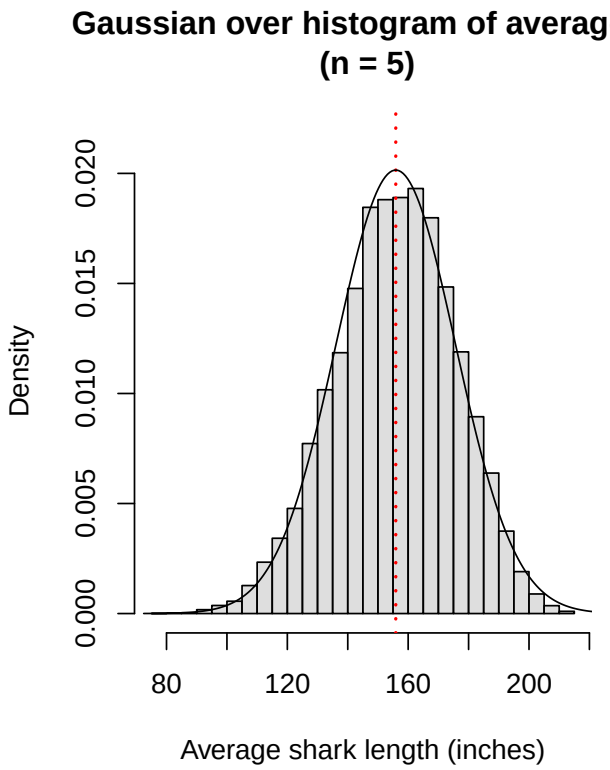
Contents

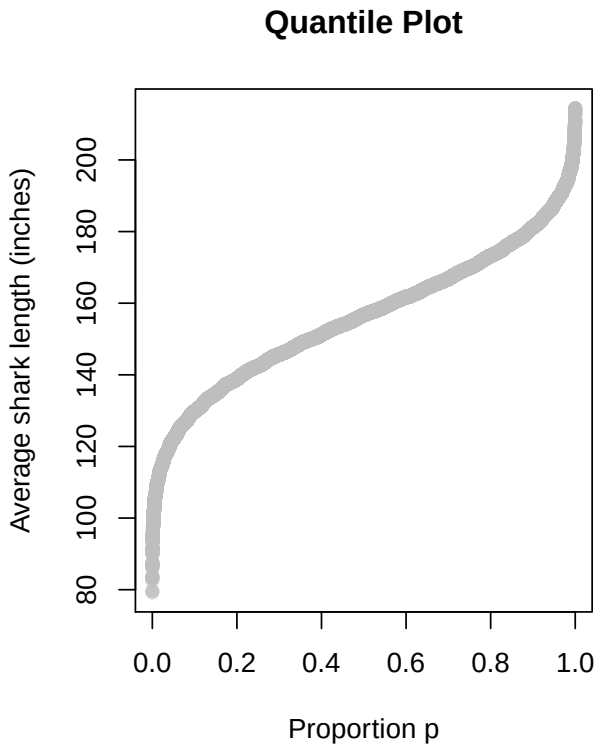
4.1 Random intervals and coverage probabilities

- Consider the Australian shark encounter population $N = 28$, where we looked at all possible samples of size $n = 5$.
  - We calculated $a(\mathcal{S})$ for each sample and the constructed a histogram.

```
sharks <- read.csv("../Data/Sharks/sharks.csv")
popSharks <- rownames(sharks)
popSharksAustralia <- popSharks[sharks$Australia == 1]

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


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

sdN = function(x){sqrt(var(x)*(length(x)-1)/length(x))}

tmpAve <- mean(avesSamp)
tmpSD <- sdN(avesSamp)

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

## The plot
par(mfrow=c(1,2))

hist(avesSamp, col=adjustcolor("grey", alpha = 0.5),
     freq = FALSE,
     main="Gaussian over histogram of averages \n (n = 5)",
```

```
     xlab="Average shark length (inches)",
     ylim=c(0, 0.022),
     breaks=25
     )
### Mark the population attribute in red
abline(v=avePop, col="red", lty=3, lwd=2)

### Add a Gaussian density
tmpAve <- mean(avesSamp)
tmpSD <- sdN(avesSamp)
tmpX <- extendrange(avesSamp)
tmpX <- seq(tmpX[1], tmpX[2], length.out = 200)
lines(tmpX, dnorm(tmpX, mean = tmpAve, sd = tmpSD))


p = (1:length(avesSamp) )/length(avesSamp)

plot(p, sort(avesSamp), pch = 19, col=adjustcolor("grey", alpha = 0.5),
     xlim=c(0,1),
     xlab = "Proportion p",
     ylab = "Average shark length (inches)",
     main = "Quantile Plot")
```





- The histogram is fairly symmetric and centered about the population average $a(\mathcal{P}) = 155.89$
  - From the 2.5 and 97.5 quantiles we have that 95% of the sample averages are contained in (115.4, 192.8).
  - We have 0.95 probabiltiy of getting a sample average within 40.5 of the population average.

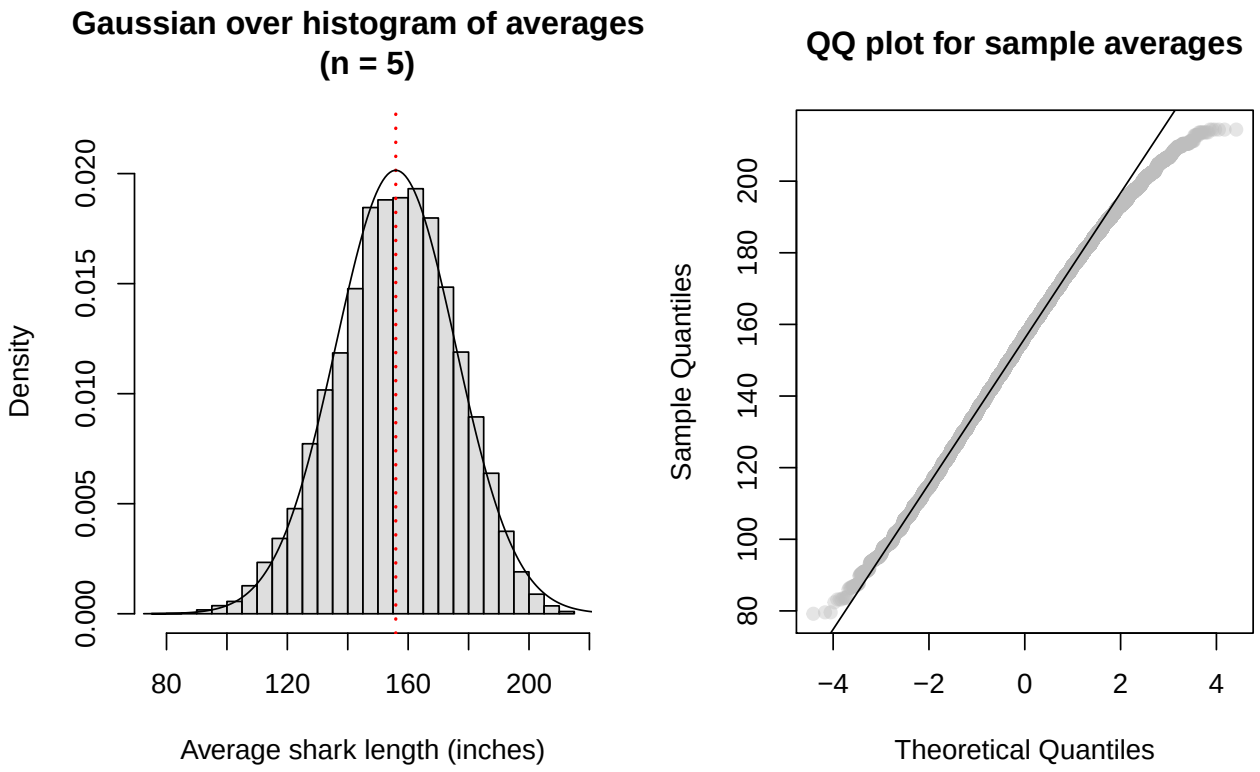


Figure 1: All possible samples: great white encounters in Australia

- Using the average, `mean(avesSamp) =` 155.89 and standard deviation, `sd(avesSamp) =` 19.8, from all possible $a(\mathcal{S})$,
  - we could try to approximate this distribution/histogram using a Gaussian distribution.

```
par(mfrow=c(1,2))

hist(avesSamp, col=adjustcolor("grey", alpha = 0.5), freq = FALSE,
     main="Gaussian over histogram of averages \n (n = 5)",
     xlab="Average shark length (inches)",
     ylim=c(0, 0.022),
     breaks=25
     )
### Mark the population attribute in red
abline(v=avePop, col="red", lty=3, lwd=2)

### Add a Gaussian density
tmpAve <- mean(avesSamp)
tmpSD <- sdN(avesSamp)
tmpX <- extendrange(avesSamp)
tmpX <- seq(tmpX[1], tmpX[2], length.out = 200)
lines(tmpX, dnorm(tmpX, mean = tmpAve, sd = tmpSD))


###Drawing the qqplot
qqnorm(avesSamp, main='QQ plot for sample averages', col=adjustcolor("Grey", 0.4),pch=19)
qqline(avesSamp)
```

- The Gaussian approximation is not bad, except for the tails.
  - You can see this both in the histogram and qq-plot.

Random Interval

- The Gaussian density provides a model for the histogram. We can use it to represent the population of average and simplify calculations.
  - Specially we will use the Gaussian approximation to construct a confidence interval for population average.

- Recall, the sample average estimator under random sampling without replacement has the properties

$$\overline{Y} = \mu \qquad \text{and} \qquad SD\left(\overline{Y}\right) = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

and

$$\sigma^2 = \frac{\sum_{u \in \mathcal{P}} (y_i - \mu)^2}{N}$$

- If the Gaussian model holds then we view the average shark length for a sample of size $n = 5$ as a random variate

$$\overline{Y} \sim G\left(\mu, \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}\right)$$

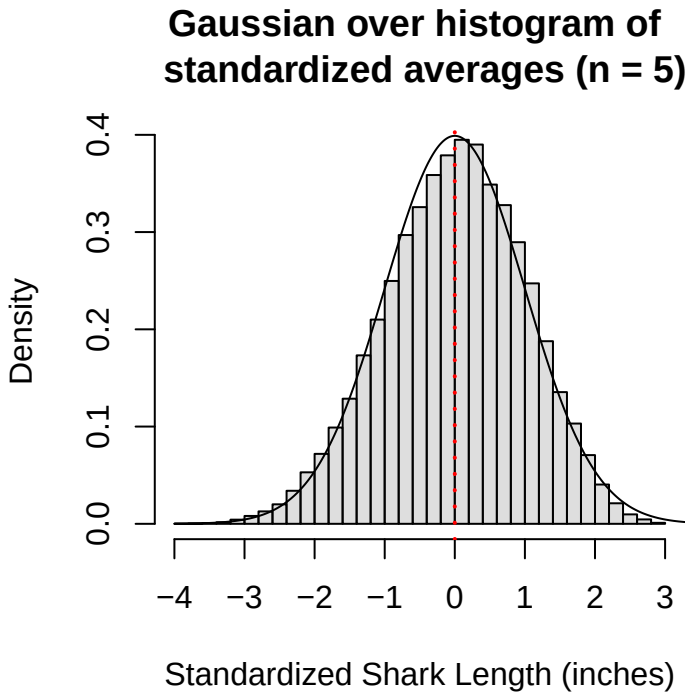
  - where (in this case) $n = 5$, $N = 28$,
  - $\mu = a(\mathcal{P}) = 155.89$,
  - $\sigma = 49.74$ is the population's standard deviation and
  - $SD(\overline{Y}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = 19.8$.

- Standardizing this random variable yields

$$\frac{\overline{Y} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} = Z \sim G(0,1)$$

- We can calculate this quantity for all possible sample and overlay a $G(0,1)$

```
Z = (avesSamp - mean(avesSamp))/sdN(avesSamp)

hist(Z, col=adjustcolor("grey", alpha = 0.5),
     freq = FALSE,
     main="Gaussian over histogram of \n standardized averages (n = 5)",
     xlab="Standardized Shark Length (inches)",
     breaks=25 )
```




```
### Mark the population attribute in red
abline(v=0, col="red", lty=3, lwd=2)

### Add a Gaussian density
x = seq(-4, 4, length.out=1000)
lines( x, dnorm(x) )
```

- Using the standardized random variable and specified $p \in (0,1)$ we can find a constant $c > 0$ such that

$$\begin{aligned} 1-p &= Pr\left(-c \leq Z \leq c\right) \\ 1-p &= Pr\left(-c \leq \frac{\overline{Y}-\mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} \leq c\right) \\ &= Pr\left(\left[\overline{Y} - c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}, \overline{Y} + c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right] \ni \mu\right). \end{aligned}$$

- Rearranging this statement yields a **random interval**

$$\left[\overline{Y} - c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\ ,\ \overline{Y} + c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right]$$

  which contains μ with probability $1-p$.

- Note that, sometimes, when $n << N$, the finite population correction (fpc) $= \frac{N-n}{N-1} \approx 1$ is dropped from the calculations above.

Properties of a random interval

- The **random interval** $\left[\overline{Y} - c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\ ,\ \overline{Y} + c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right]$ contains μ with probability $1-p$.
- Intervals generated according to this mechanism will contain μ, $100(1-p)\%$ of the time.
  - $1-p$ is therefore called the **coverage probability**.
  - These intervals all have the same width, just different (and random) centres.

- The Gaussian distribution is symmetric about its mean μ, so p and c are related through

$$(1-p) + p/2 = 1 - p/2 = Pr(Z \leq c)$$

  where $Z \sim G(0,1)$ is a standard Gaussian random variate.
  - therefore, c can be determined for any values of p.
- Given p the value of c can determine from the quantile function of a standard Gaussian random variate in that

$$c = Q_Z\left(1 - \frac{p}{2}\right)$$

  which in `R` is calculated as `qnorm(1 - p/2)`. e.g. $c \approx 1.96$ when $1-p = 0.95$ for a standard Gaussian random variate.

A single sample of size n

- In practice, we will have only one sample, and
  - a single numerical average $\overline{y}$ for that sample and
  - we have only one of these randomly generated intervals,

$$\left[\overline{Y} - c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\ ,\ \overline{Y} + c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right].$$

- According to the Gaussian model, $100(1-p)\%$ of the intervals generated will contain μ,
  - we then have some **confidence** that this particular interval will cover μ as well, but we will never know if it will.

- The larger is $1-p$, the more confident we are that the interval will contain μ.
  - The probability statement is attached to the **method used to generate the intervals** and not to the particular interval in hand.
  - The one in hand is therefore called a $100(1-p)\%$ **confidence interval** and not a probability interval.

A 100 Samples (each of size n)

- Given that $\mu = a(\mathcal{P}) = 155.89$ and $SD(\overline{Y})$ =`sdN(avesSamp)` = 19.8.
  - We randomly select 100 samples of size n, and
  - and form the interval for any **level of confidence** measured by the coverage probability $(1-p) = 0.95$ then
  - approximately $100(1-p)\% = 95\%$ of the intervals should contain (or cover) $\mu = 155.89$.

```
## R code to simulate the GAUSSIAN approximation

ylim <- c(0, 0.022)

### Add a Gaussian density
tmpAve <- mean(avesSamp)
tmpSD <- sdN(avesSamp)
confidence <- 0.95
p <- 1 - confidence
numIntervals <- 100
cValue <- qnorm(1 - p/2)  # or qnorm((confidence + 1)/2)
set.seed(34781453) # comment out this line to get different samples every time you run the code.
ybarSampled <- sample(avesSamp, numIntervals)
heights <- seq(diff(ylim)/numIntervals, max(ylim), length.out = numIntervals)
xlim <- extendrange(avesSamp + cValue * c(-tmpSD, tmpSD))
hist(avesSamp, col=adjustcolor("grey", alpha = 0.5), freq = FALSE,
     main=paste0(numIntervals, " individual ", round(100 * confidence), "% confidence intervals"),
     xlab="Average shark length (inches)",
     ylim=ylim, xlim = xlim,
     breaks=25
     )
### Mark the population attribute in red
abline(v=avePop, col="red", lty=3, lwd=2)
numIntervalsMissed <- 0
for(i in 1:numIntervals) {
lines(ybarSampled[i] + cValue * c(-tmpSD, tmpSD), rep(heights[i],2),
      col = "steelblue")
  if (tmpAve > ybarSampled[i] + cValue*tmpSD) {
    points(ybarSampled[i] - cValue*tmpSD, heights[i], pch=8, cex=1.2, col="red")
    points(ybarSampled[i], heights[i], pch=18, cex=1.2, col="red")
    lines(rep(ybarSampled[i], 2), c(0, heights[i]), col = "red")
  } else if (tmpAve < ybarSampled[i] - cValue*tmpSD) {
    points(ybarSampled[i] + cValue*tmpSD, heights[i], pch=8, cex=1.2, col="red")
    points(ybarSampled[i], heights[i], pch=18, cex=1.2, col="red")
    lines(rep(ybarSampled[i], 2), c(0, heights[i]), col = "red")
  }  else numIntervalsMissed <- numIntervalsMissed + 1

}

tmpX <- extendrange(avesSamp)
tmpX <- seq(tmpX[1], tmpX[2], length.out = 200)
lines(tmpX, dnorm(tmpX, mean = tmpAve, sd = tmpSD))
```

- 93 of these 100 intervals cover the value μ.
  - Those which do *not* cover μ are marked with a star.

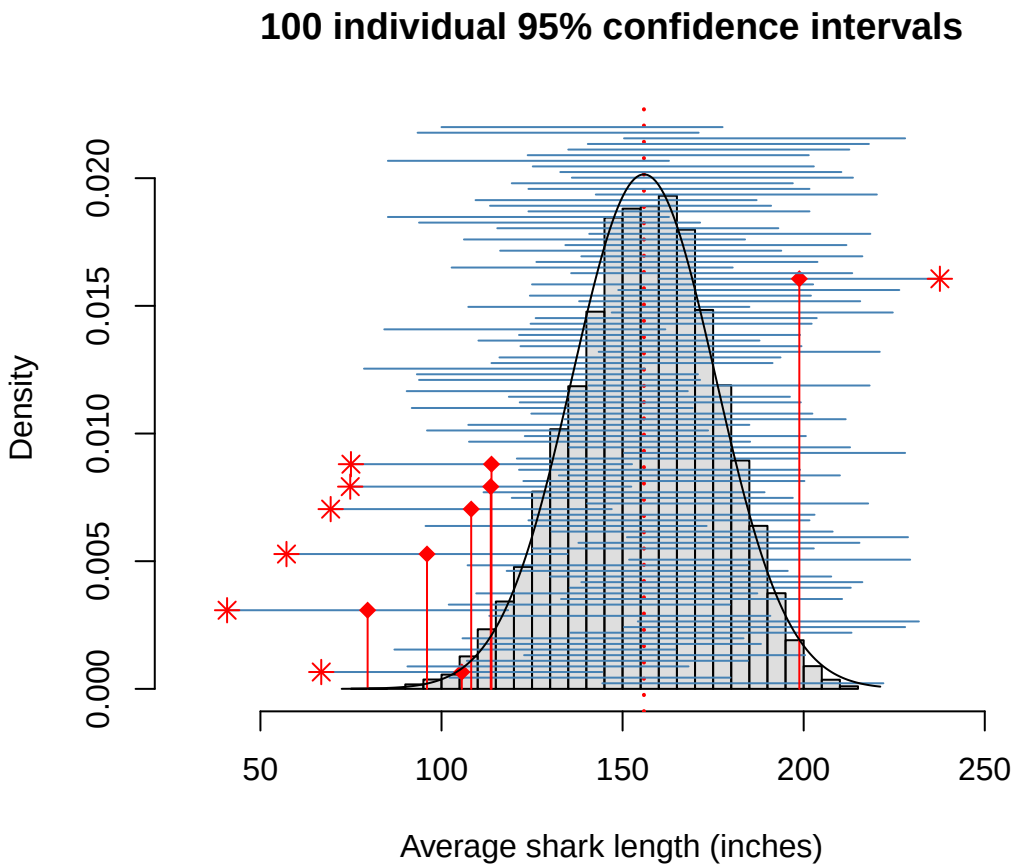


Figure 2: 100 confidence intervals

Confidence interval and test of hypothesis

Suppose that for some scientific purpose, we would like to test some hypothesis. Confidence intervals can be used to construct an appropriate test. The reasoning is as follows.

- A proportion, p, of the $100(1-p)\%$ random intervals will *not* contain the true value of μ. If our $100(1-p)\%$ confidence interval does *not* contain a, then we have reason to suspect that the hypothesis does not hold. Somehow $(1-p)$, or equivalently p, measures this suspicion. Whatever the value of a, there will be some confidence level $(1-p)$ such that a is an end point of the confidence interval.
- The value of p defining this confidence interval is called, unimaginatively, the **p-value** of this test. The smaller is p, the larger $(1-p)$ had to be for the $100(1-p)\%$ confidence interval to contain the hypothesized value $\mu = a$. The smaller p is, the *greater* is the evidence *against* the hypothesis $H_0 : \mu = a$.
- The *p-value is an observed level of significance* SL and this is a test of significance. (To reinforce that this is a significance test, we only use SL for the observed significance and do not use the term "p-value" for this concept; "p-value" is common usage by others.)

The Standard Error

- The confidence intervals we calculated used $SD(\overline{Y})$.
  - It is pretty rare to have this value in general.

- However, for many sample attributes $a(\mathcal{S})$ (e.g. Horvitz-Thompson estimators), we can estimate of the standard deviation $SD(\widetilde{a}(\mathcal{S}))$.
  - **The standard error** is an estimate of the standard deviation of the corresponding estimator, i.e. standard error = $\widehat{SD}(\widetilde{a}(\mathcal{S}))$.

- Using an estimate $\widehat{SD}$ for the SD will increase the variability of the random intervals.

4.4.1 Student-t based intervals

- If we have to estimate $\widehat{SD}$ as well.

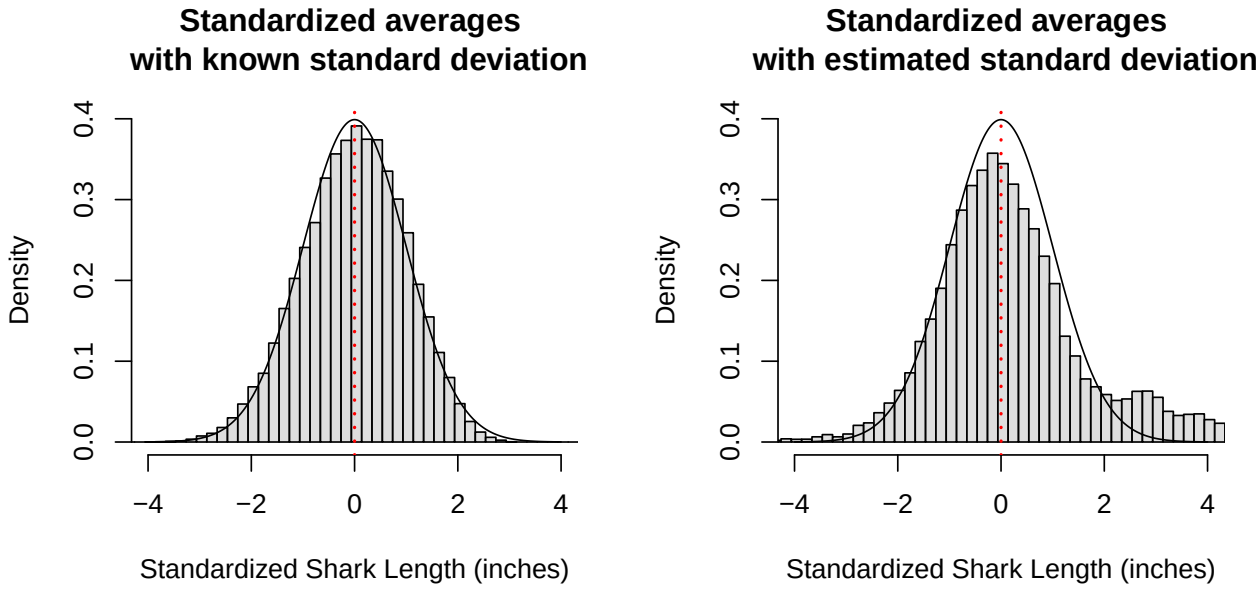
  - our we are considering the estimator
$$\frac{\widetilde{a}(\mathcal{S}) - a(\mathcal{P})}{\widehat{SD}(\widetilde{a}(\mathcal{S}))}$$
  - we might expect this quantity to have larger variation (why?)

- Using Australia sharks lengths on every possible sample we can calculate
$$\frac{a(\mathcal{S}) - a(\mathcal{P})}{SD(\widetilde{a}(\mathcal{S}))} \quad \text{and} \quad \frac{a(\mathcal{S}) - a(\mathcal{P})}{\widehat{SD}(\widetilde{a}(\mathcal{S}))}$$

```
par(mfrow=c(1,2))

Z = (avesSamp - mean(avesSamp))/sdN(avesSamp)

n = 5
N = 28
se = sdsSamp/sqrt(n)*sqrt((N-n)/(N-1))
t = (avesSamp - mean(avesSamp))/se

## So the two histograms have the same bins
delta = 0.2
brk = seq(min(t)-delta, max(t)+delta, delta)


hist(Z, col=adjustcolor("grey", alpha = 0.5),
     freq = FALSE,
     main="Standardized averages \n with known standard deviation",
     xlab="Standardized Shark Length (inches)",
     breaks=brk, xlim=c(-4,4), ylim=c(0,0.4) )

### Mark the population attribute in red
abline(v=0, col="red", lty=3, lwd=2)

### Add a Gaussian density
x = seq(-4, 4, length.out=1000)
lines( x, dnorm(x) )


hist(t, col=adjustcolor("grey", alpha = 0.5),
     freq = FALSE,
     main="Standardized averages \n with estimated standard deviation",
     xlab="Standardized Shark Length (inches)",
     breaks=brk, xlim=c(-4,4), ylim=c(0,0.4) )

### Mark the population attribute in red
abline(v=0, col="red", lty=3, lwd=2)

### Add a Gaussian density
x = seq(-4, 4, length.out=1000)
lines( x, dnorm(x) )
```

```
#lines( x, dt(x, n-1),col=2 )
```

- Black line is the density for $G(0,1)$



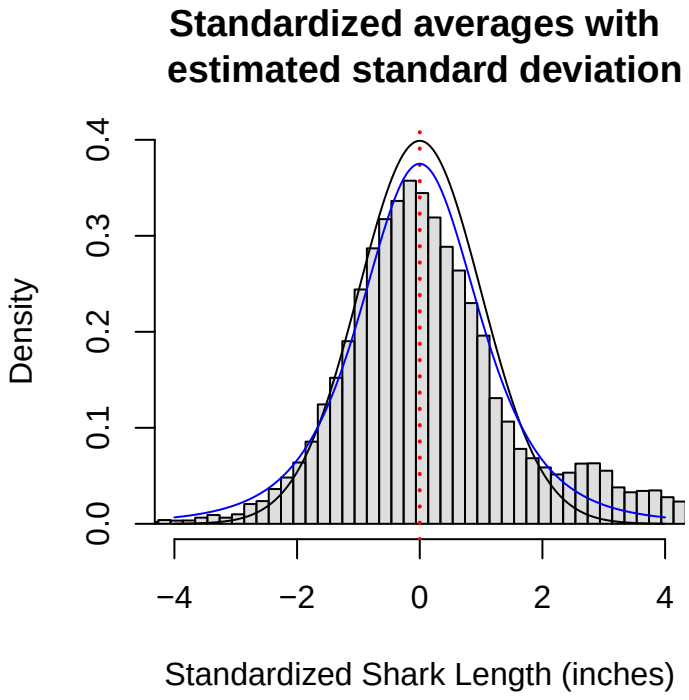
  - the Gaussian does not seem be a good approximation when we have to estimate the standard deviation.

- Under the Gaussian model we know that the following is a pivotal quantity

$$\frac{\overline{Y}-\mu}{SD(\overline{Y})} = \frac{\overline{Y}-\mu}{\widetilde{\sigma}/\sqrt{n}} \sim t_{n-1}$$

- This quantity is called a **pivotal** statistic in that
  - it is a function of the unknown parameter μ and the sample values Y_u (for $u \in \mathcal{S}$) whose **sampling distribution is completely known**.
  - Maybe try this approximation?

```
n = 5
N = 28
se = sdsSamp/sqrt(n)*sqrt((N-n)/(N-1))
t = (avesSamp - mean(avesSamp))/se

## So the two histograms have the same bins
delta = 0.2
brk = seq(min(t)-delta, max(t)+delta, delta)

hist(t, col=adjustcolor("grey", alpha = 0.5),
     freq = FALSE, ylim = c(0,.4),
     main="Standardized averages with \n estimated standard deviation",
     xlab="Standardized Shark Length (inches)",
     breaks=brk, xlim=c(-4,4) )

### Mark the population attribute in red
abline(v=0, col="red", lty=3, lwd=2)

### Add a Gaussian density
x = seq(-4, 4, length.out=1000)
```




```
lines( x, dnorm(x) )
lines( x, dt(x, n-1), col=4 )
```

- The t-distribution in blue and Gaussian in black.
  - The t-distribution with $n-1=4$ is an okay approximation, but not great.

- Now if we suppose that

$$\frac{\widetilde{a}(\mathcal{S})-a(\mathcal{P})}{\widetilde{SD}(\widetilde{a}(\mathcal{S}))} \sim t_{n-1}$$

- Then we choose a $p \in (0,1)$ and a corresponding $c>0$ with

$$\begin{aligned} p &= Pr\left(-c \leq \tfrac{\widetilde{a}(\mathcal{S})-a(\mathcal{P})}{\widetilde{SD}(\widetilde{a}(\mathcal{S}))} \leq c\right) \\ &= Pr\left(\left[\widetilde{a}(\mathcal{S}) - c \times \widetilde{SD}(\widetilde{a}(\mathcal{S})), \widetilde{a}(\mathcal{S}) + c \times \widetilde{SD}(\widetilde{a}(\mathcal{S}))\right] \ni \mu\right). \end{aligned}$$

- Now this random interval now has a random centre *and* random length.
- The value of c is determined using the t distribution with $n-1$ degrees of freedom.
  - In `R` use `qt((p+1)/2, df = n-1)` to get the value of c
  - check that $c \approx 2.78$ when $p=0.95$ for a t_4 random variate.

Random interval for the Population Average

- The sample average $a(\mathcal{S}) = \overline{y}$ and its corresponding estimator $\widetilde{a}(\mathcal{S}) = \overline{Y}$.
- The standard deviation of the estimator is

$$SD(\widetilde{a}(\mathcal{S})) = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

  where σ denotes the population standard deviation.
- The unknown quantity is σ and a sample estimate is

$$\widehat{\sigma} = \sqrt{\frac{\sum_{u \in \mathcal{S}} (y_u - \overline{y})^2}{n}}$$

- An estimate of the standard deviation of the estimator is **the standard error** is

$$\widehat{SD}(\widetilde{a}(\mathcal{S})) = \frac{\widehat{\sigma}}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

- The random intervals constructed from these estimators is

$$\left[\overline{Y} - c \times \frac{\widetilde{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}, \overline{Y} + c \times \frac{\widetilde{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}\right].$$

Generating 100 Samples each of size $n = 5$

- We randomly select 100 samples of size $n = 5$, and for each sample we construct

$$\left[\overline{y} - c \times \frac{\widehat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}, \overline{y} + c \times \frac{\widehat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}\right].$$

  - using $c \approx 2.78$ from t_4 random variable then
  - approximately 95% of the intervals should contain (or cover) $\mu = 155.89$.

```
## R code to simulate the t-STUDENT approximation

ylim <- c(0, 0.022)
n <- 5
N <- 28
### Add a Gaussian density
tmpAve       <- mean(avesSamp)
tmpSD        <- sdN(avesSamp)
confidence   <- 0.95
numIntervals <- 100
cValue       <- qt((confidence +1)/2, df = n-1)

set.seed(341)
```

```
tmpSamp     <- sample(1:length(avesSamp), numIntervals)
ybarSampled <- avesSamp[tmpSamp]
sdSampled   <- sdsSamp[tmpSamp]
maxSD       <- max(sdSampled)/sqrt(n)

heights <- seq(diff(ylim)/numIntervals, max(ylim), length.out = numIntervals)
xlim <- extendrange(c(ybarSampled - cValue * sqrt((N-n)/(N-1)) * sdSampled/sqrt(n),
                ybarSampled + cValue * sqrt((N-n)/(N-1)) * sdSampled/sqrt(n)))
hist(avesSamp, col=adjustcolor("grey", alpha = 0.5), freq = FALSE,
     main=paste0(numIntervals, " individual ", round(100 * confidence), "% confidence intervals"),
     xlab="Average shark length (inches)",
     ylim=ylim, xlim = xlim,
     breaks=25
)
### Mark the population attribute in red
abline(v=avePop, col="red", lty=3, lwd=2)
numIntervalsMissed <- 0
for(i in 1:numIntervals) {
  tmpSampSD <- sdSampled[i]/sqrt(n)*sqrt((N-n)/(N-1))

  lines(ybarSampled[i] + cValue * c(-tmpSampSD, tmpSampSD), rep(heights[i],2),
        col = "steelblue")
  if (tmpAve > ybarSampled[i] + cValue*tmpSampSD) {
    points(ybarSampled[i] - cValue*tmpSampSD, heights[i], pch=8, cex=1.2, col="red")
    points(ybarSampled[i], heights[i], pch=18, cex=1.2, col="red")
    lines(rep(ybarSampled[i], 2), c(0, heights[i]), col = "red")
  } else if (tmpAve < ybarSampled[i] - cValue*tmpSampSD) {
    points(ybarSampled[i] + cValue*tmpSampSD, heights[i], pch=8, cex=1.2, col="red")
    points(ybarSampled[i], heights[i], pch=18, cex=1.2, col="red")
    lines(rep(ybarSampled[i], 2), c(0, heights[i]), col = "red")
  }  else numIntervalsMissed <- numIntervalsMissed + 1
}
tmpX <- extendrange(avesSamp)
tmpX <- seq(tmpX[1], tmpX[2], length.out = 200)
lines(tmpX, dnorm(tmpX, mean = tmpAve, sd = tmpSD))
```

- 96 of these 100 intervals cover the value μ.
  - Those which do *not* cover μ (the red dashed line in the centre) are marked with a star
- We can find the exact coverage probability by calculating all possible confidence intervals.

Pivotal Quantity

- Pivotal quantities are the basis for constructing random intervals.
- Many random intervals are constructed via

$$\frac{\widetilde{a}(\mathcal{S}) - a(\mathcal{P})}{\widetilde{SD}(\widetilde{a}(\mathcal{S}))}$$

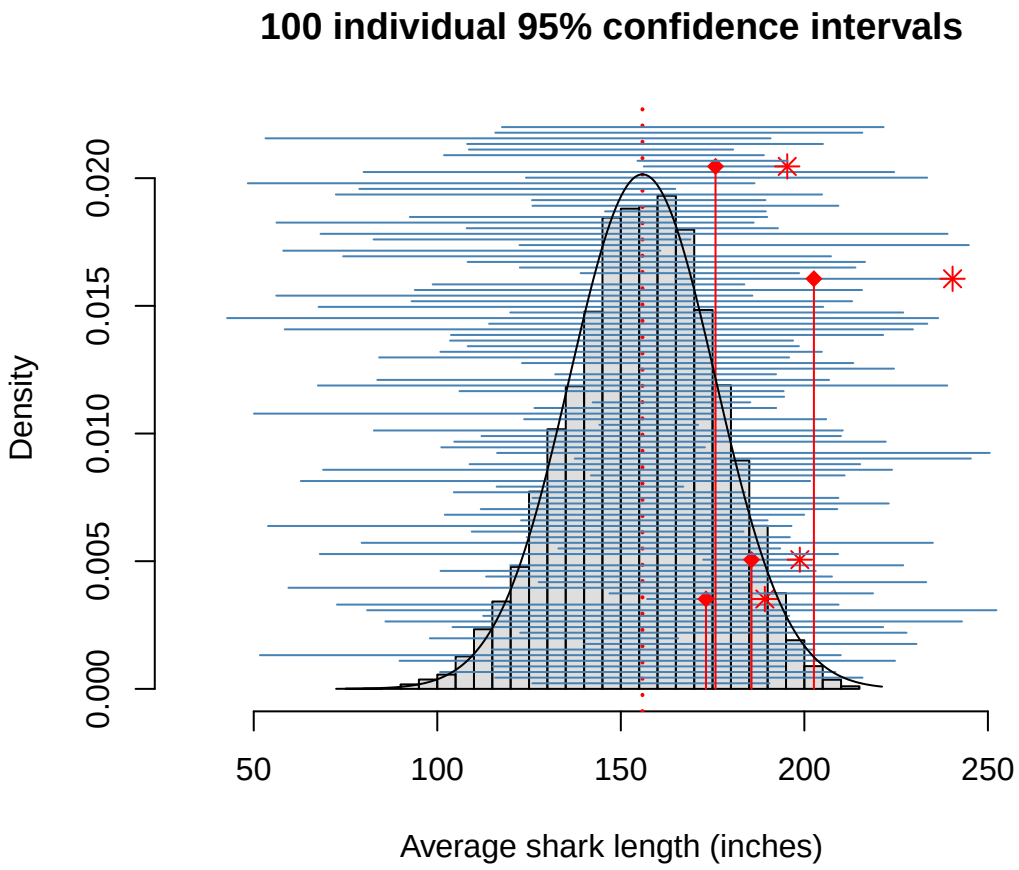


Figure 3: 100 confidence intervals

provided it is a **pivotal** function or even **approximately pivotal**.

  - Then we *pivot* and isolate $a(\mathcal{P})$ to construct a random interval.

- e.g. When conducting significance tests on comparing sub-populations,
  - we observed that distribution of t-like discrepancy measures could be approximated with a Student-t distribution.

- Other common **pivotals** or **approximately pivotal** quantities, usually used for scale attributes $s(\cdot)$, are of the form

$$\frac{\widetilde{s}(\mathcal{S})}{s(\mathcal{P})}$$