Bootstrap Sampling: Confidence Intervals and a One-Sample Example

BIOS 6611

CU Anschutz

Week 6

Confidence Intervals

Bootstrap Sampling for a Single Population

Given a sample of size n from a population,

1. Draw a resample of size n with replacement from the sample. Compute a statistic that describes the sample, such as the sample mean.
2. Repeat this resampling process many times, say 10,000.
3. Construct the bootstrap distribution of the statistic. Inspect its spread, bias, and shape.

From the bootstrap distribution, we can also estimate the variability of the statistic by calculating a confidence interval. We will cover two: *normal percentile* and *bootstrap percentile*.

Normal Percentile Confidence Intervals

Since we are calculating the mean of our statistic from our bootstrap distribution, we could potentially leverage the CLT to calculate a **normal percentile confidence interval**:

$$\bar{X} \pm Z_{1-\frac{\alpha}{2}} SE(\bar{X})$$

However, what if our bootstrap distribution deviates from normality? Then the approximation via the CLT would be inaccurate.

To check the performance of the CLT we can evaluate the *coverage* in each tail. Since the normal distribution is symmetric, for a 95% CI we would expect 2.5% of the bootstrap distribution to be in each tail. If our coverage is not approximately 2.5% we should use consider methods to estimate our CI.

Bootstrap Percentile Confidence Intervals

If we are concerned about assuming normality, as with the normal percentile CIs, we can take a purely empirical estimate for our confidence interval.

The interval between 2.5 and 97.5 percentiles of the bootstrap distribution of a statistic is a 95% **bootstrap percentile confidence interval** for the corresponding parameter.

Rule of thumb: If the ratio of bias/SE exceeds ± 0.10, then it could have a substantial effect on the *accuracy* of the bootstrap confidence intervals, and more accurate interval formulations should be used.

This is easy to compute, but may not always be optimal. See the "Bootstrap" paper by Tim Hesterberg (2011) to review some proposed modifications.

One-Sample Example

Arsenic in Bangladesh Groundwater

Arsenic is a naturally occurring element in the groundwater of Bangladesh. However, much of the groundwater is used for drinking water by rural populations, so arsenic poisoning is a serious health concern.

In this example we will

1. Describe the distribution of arsenic in groundwater along with the sample mean and standard deviation
2. Obtain and describe a bootstrap sample for the mean
3. Obtain a confidence interval for the bootstrap distribution using normal percentiles
4. Obtain a confidence interval for the bootstrap distribution using bootstrap percentiles

1. Describe the Distribution

We will load the data within `R` using the `resampledata` package:

```
library(resampledata)
set.seed(53)
Arsenic <- resampledata::Bangladesh$Arsenic
```

From our sample of 271 wells,

```
mean(Arsenic)
```

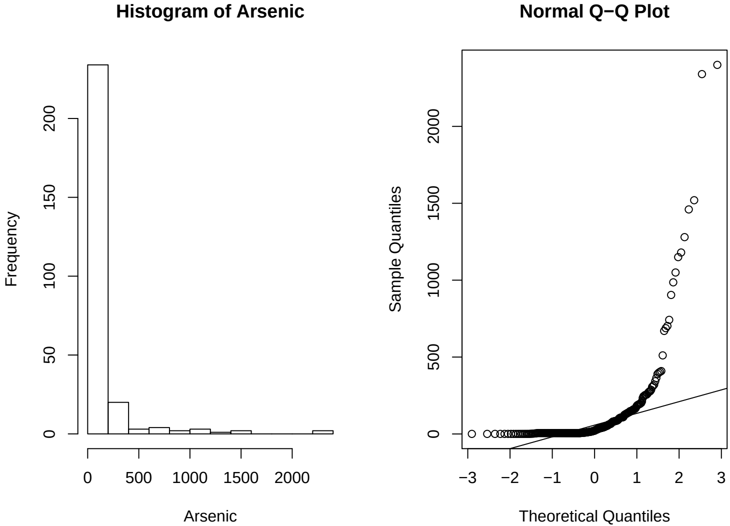
```
## [1] 125.3199
```

```
sd(Arsenic)
```

```
## [1] 297.9755
```

1. Plots of the Sample

```r
par( mfrow=c(1,2))
hist(Arsenic)
qqnorm(Arsenic); qqline(Arsenic)
```



Histogram of Arsenic

Normal Q−Q Plot

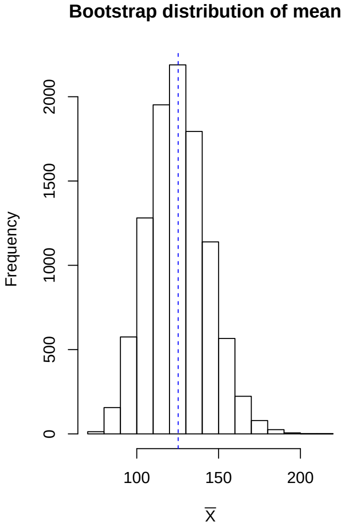
2. Bootstrap Sample of the Mean

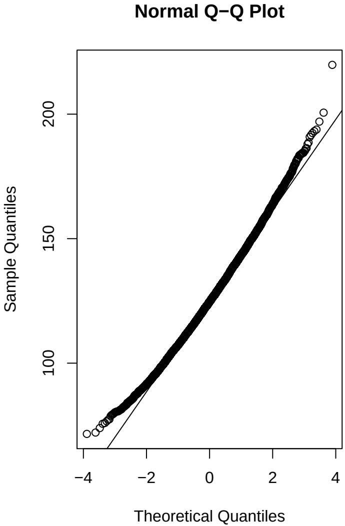
Now let's implement our bootstrap with 10,000 samples:

```r
n <- length(Arsenic)
B <- 10^4

# Initialize object to store results in
arsenic.mean <- numeric(B)

# Implement bootstrap
for (i in 1:B){
  x <- sample(Arsenic, n, replace = TRUE)
  arsenic.mean[i] <- mean(x)
}
```

2. Plots of the Bootstrap Distribution

```
hist(arsenic.mean, main = "Bootstrap distribution of means",xlab=expression
abline(v = mean(Arsenic), col = "blue", lty = 2) # observed mean
qqnorm(arsenic.mean); qqline(arsenic.mean)
```





2. Bootstrap Mean, Bias, and SE

From our bootstrap distribution we can calculate our mean, bias, and standard error:

```r
mean(arsenic.mean) # bootstrap mean
```

```
## [1] 125.3642
```

```r
mean(arsenic.mean)-mean(Arsenic) # bias
```

```
## [1] 0.04429897
```

```r
sd(arsenic.mean) # bootstrap SE
```

```
## [1] 18.05039
```

We see there is minimal bias in our bootstrap estimator for the mean arsenic concentration.

Additionally, if we applied the CLT we would expect $\frac{s}{\sqrt{n}} = \frac{297.98}{\sqrt{271}} = 18.101$, which is very similar to our bootstrap SE for mean arsenic.

3. Normal Percentile CI

Let's calculate and evaluate the 95% normal percentile CI for our bootstrap distribution, where $Z_{0.975} \approx 1.96$:

```r
#Lower limit of 95% Normal CI
LL <- mean(arsenic.mean)-1.96*sd(arsenic.mean)
LL
```

```
## [1] 89.98547
```

```r
#Upper limit of 95% Normal CI
UL <- mean(arsenic.mean)+1.96*sd(arsenic.mean)
UL
```

```
## [1] 160.743
```

Interpretation: The 95% normal percentile CI is (90.0, 160.7). We are 95% confident that the true mean lies in this interval, assuming the central limit theorem applies.

3. Coverage of the Normal Percentile CI

```
sum(arsenic.mean < LL)/B  # Coverage of CI at lower end
```

```
## [1] 0.0169
```

```
sum(arsenic.mean > UL)/B  # Coverage of CI at upper end
```

```
## [1] 0.0313
```

Interpretation: Based on our estimates of coverage, the 95% normal percentile estimates are too low for both the lower and upper bounds since the lower bound has coverage of 1.69% and the upper bound has coverage of 3.13% instead of the desired 2.5%, suggesting the CLT may be inaccurate.

4. Bootstrap Percentile CI

Let's calculate the 95% bootstrap percentile CI:

```
quantile( arsenic.mean, c(0.025,0.975))
```

```
##      2.5%     97.5%
##  92.08075 162.97526
```

Interpretation: The 95% bootstrap percentile CI is (92.1, 163.0). We are 95% confident that the true mean is in this interval.

Additionally, because (92.1, 163.0) is estimated from our data directly, 95% of the bootstrap means fall in this interval.

4. Accuracy of Bootstrap Percentile CI

Recall, we can estimate our accuracy by examining bias/SE:

```
(mean(arsenic.mean)-mean(Arsenic)) / sd(arsenic.mean)
```

```
## [1] 0.002454184
```

Interpretation: The accuracy of our bootstrap percentile can be estimated by the ratio of the bias/SE, which is 0.002. Since this does not exceed $+0.10$ we should have good accuracy.

Note that this interval is not necessarily symmetric around our estimated parameter. This allows greater flexibility than the normal percentile which enforces a symmetric confidence interval.

Conclusion

Putting it all together, our estimated mean arsenic concentration (from our sample) was 125.3 (SD=298.0) where $SE(\bar{X})$=18.10, so the 95% CI is (89.9, 161.0) if we wanted to assume underlying distribution is normal or that the CLT applies:

The mean arsenic concentration is 125.3 μg/L (95% CI: 89.9, 161.0).

Given the distribution of arsenic was not normally distributed, we may be concerned that our sample of 271 wells does not have enough accuracy to estimate the mean (even with the CLT). Indeed, our 95% normal percentile interval had potential coverage issues, so we would want to report the sample mean with its 95% bootstrap percentile interval:

The mean arsenic concentration is 125.3 μg/L (95% bootstrap CI: 92.1, 163.0).