Bootstrap

2024-05-22

Bootstrapping is a powerful statistical technique used to estimate the sampling distribution of a statistic by repeatedly resampling with replacement from the original data. This method allows for the assessment of the variability and accuracy of the statistic without relying on strong parametric assumptions. By generating numerous resampled sets of data, bootstrapping creates a distribution of the statistic of interest, which can be used to derive confidence intervals and other measures of statistical uncertainty. This technique is particularly useful when the theoretical distribution of the statistic is complex or unknown.

In our analysis, bootstrapping is applied to estimate the efficacy of the BNT162b2 vaccine. By generating multiple resampled datasets from the original data, we can repeatedly calculate the vaccine efficacy and construct a distribution of these efficacy estimates. This distribution provides valuable insights into the range, variability, and confidence intervals of the vaccine's effectiveness. Using bootstrapping, we can assess the reliability of the observed efficacy, ensuring that the conclusions drawn from the data are well-supported by a rigorous statistical framework.

```r
data <- read_csv("data.csv", show_col_types = F)

colnames(data)[colnames(data) == 'No_COVID'] <- 'n'
```

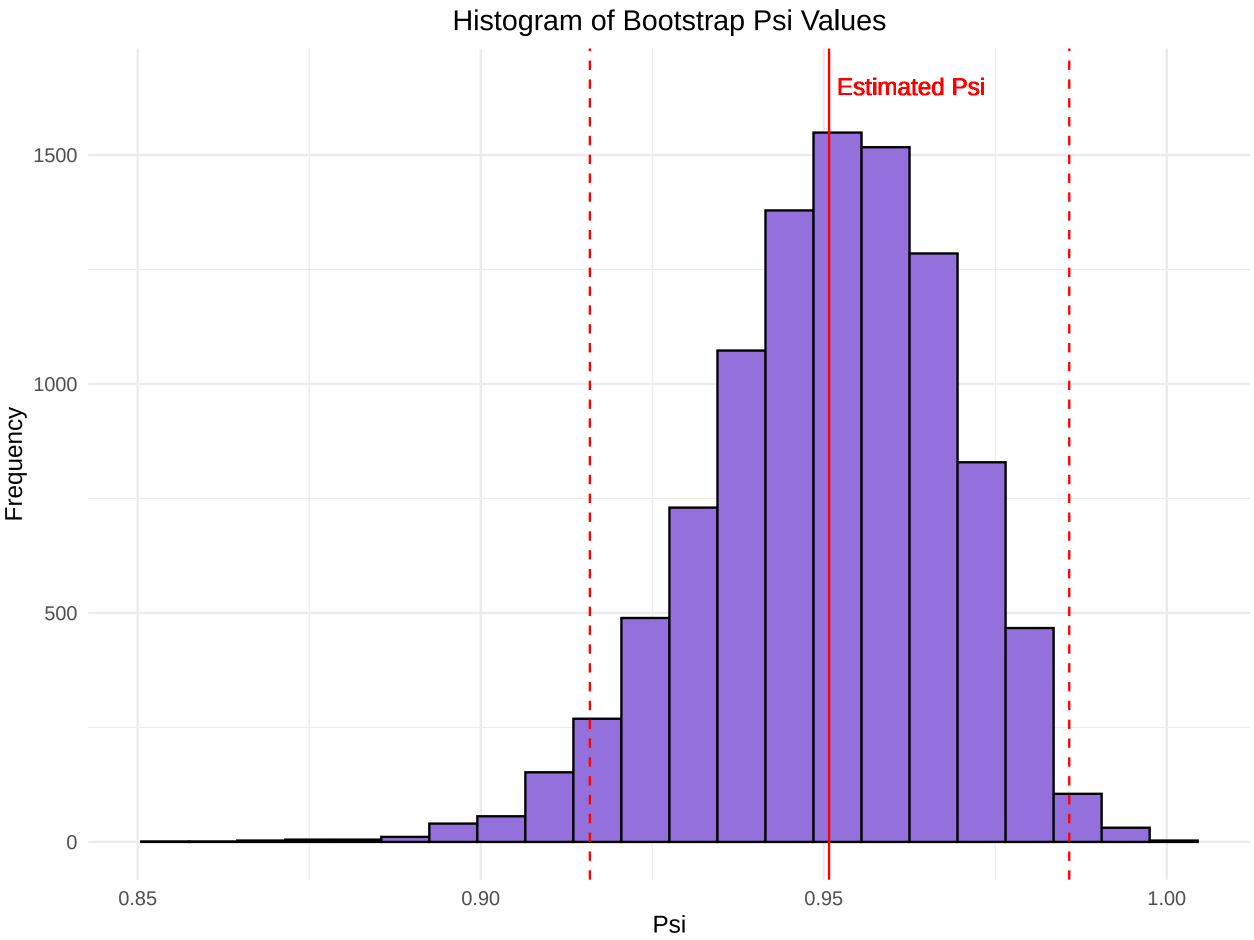
```r
vaccine <- data %>%
  filter(Test == "Vaccine")

placebo <- data %>%
  filter(Test == "Placebo")

prop_vaccine <- vaccine$COVID / vaccine$n
prop_placebo <- placebo$COVID / placebo$n

observed_pi <- prop_vaccine/(prop_vaccine + prop_placebo)

observed_psi <- (1 - 2*observed_pi)/(1 - observed_pi)
```

```r
n_bootstrap <- 10000
bootstrap_psis <- numeric(n_bootstrap)
set.seed(123)

for (i in 1:n_bootstrap) {
  vaccine_sample <- sample(c(0, 1), size = vaccine$n, replace = TRUE, prob = c(1 - prop_vaccine, prop_va
  placebo_sample <- sample(c(0, 1), size = placebo$n, replace = TRUE, prob = c(1 - prop_placebo, prop_pl

  prop_vaccine_boot <- mean(vaccine_sample)
  prop_placebo_boot <- mean(placebo_sample)

  bootstrap_pi <- prop_vaccine_boot / (prop_vaccine_boot + prop_placebo_boot)

  bootstrap_psis[i] <- (1 - 2 * bootstrap_pi) / (1 - bootstrap_pi)
}
```

```r
bootstrap_df <- data.frame(psi = bootstrap_psis)

bootstrap_summary <- bootstrap_df %>%
  summarise(n = n(),
            mean = mean(psi),
            sd = sd(psi),
            lower_critical_value = mean - qnorm(0.975)*sd,
            upper_critical_value =  mean + qnorm(0.975)*sd)

lower_critical_value <- bootstrap_summary$lower_critical_value
upper_critical_value <- bootstrap_summary$upper_critical_value

bootstrap_summary
```

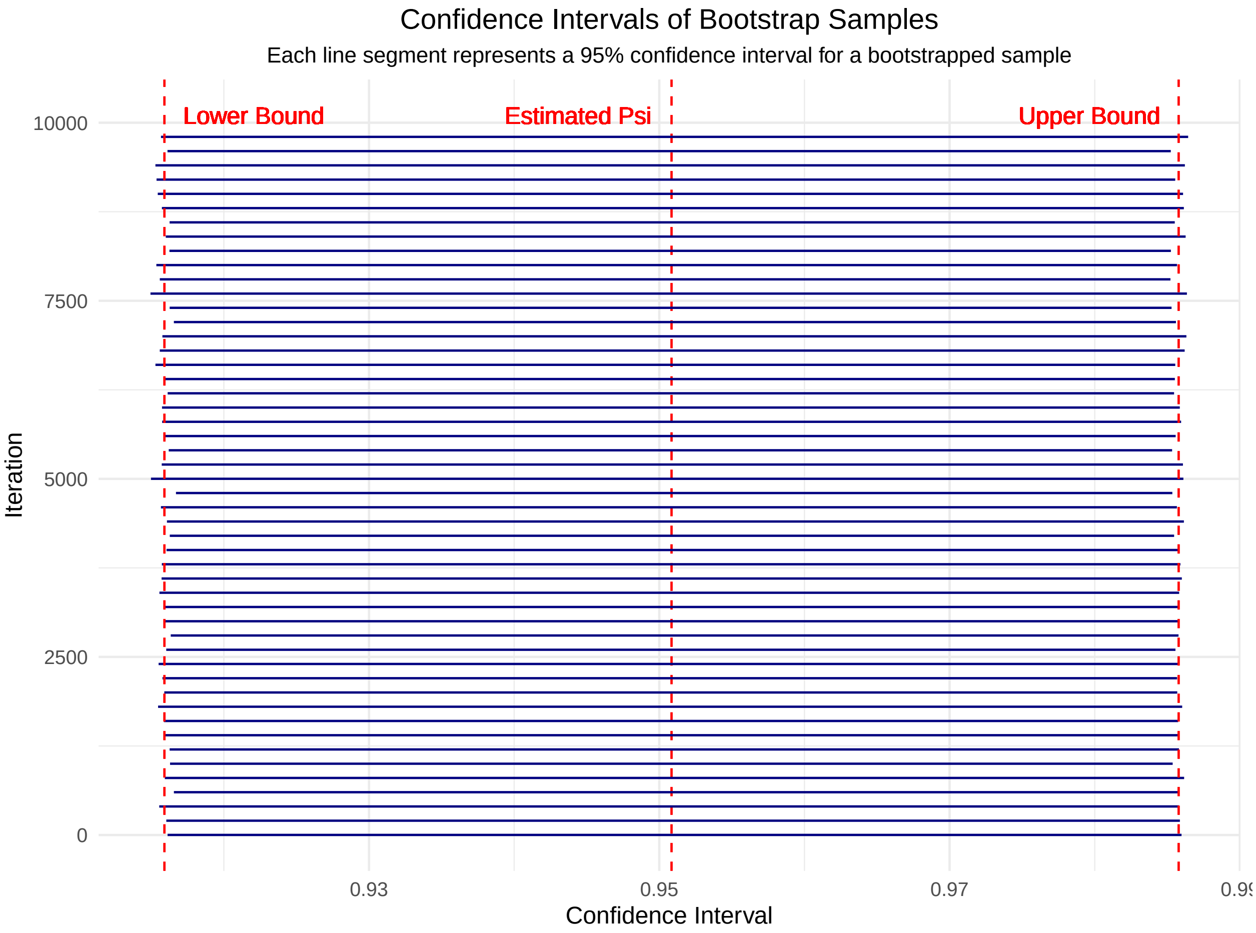
```
##       n     mean         sd lower_critical_value upper_critical_value
## 1 10000 0.950843 0.01782605            0.9159046            0.9857814
```

```r
ggplot(bootstrap_df, aes(x = psi)) +
  geom_histogram(binwidth = 0.007, fill = "mediumpurple", color = "black") +
  geom_text(aes(x = bootstrap_summary$mean, y = 1650, label = "Estimated Psi"), color = "red", hjust =
  labs(title = "Histogram of Bootstrap Psi Values", x = "Psi", y = "Frequency") +
  theme_minimal() +
  geom_vline(xintercept = observed_psi, linetype = "solid", color = "red") +
  geom_vline(xintercept = bootstrap_summary$upper_critical_value, linetype = "dashed", color = "red") +
  geom_vline(xintercept = bootstrap_summary$lower_critical_value, linetype = "dashed", color = "red") +
  theme(plot.title = element_text(hjust = 0.5))
```



```r
ci_data <- data.frame(
  Iteration = 1:n_bootstrap,
  Lower = numeric(n_bootstrap),
  Upper = numeric(n_bootstrap)
)

for (i in 1:n_bootstrap) {
  sample_psis <- sample(bootstrap_psis, n_bootstrap, replace = TRUE)
  ci_data$Lower[i] <- mean(sample_psis) - qnorm(0.975)*sd(sample_psis)
  ci_data$Upper[i] <- mean(sample_psis) + qnorm(0.975)*sd(sample_psis)
}
```

```r
plot_data <- ci_data[seq(1, n_bootstrap, by = 200), ]

ggplot(plot_data, aes(y = Iteration)) +
  geom_vline(xintercept = bootstrap_summary$mean, linetype = "dashed", color = "red") +
  geom_text(aes(x = bootstrap_summary$mean - 0.012, y = max(Iteration) + 300, label = "Estimated Psi"),
  geom_segment(aes(yend = Iteration, x = Lower, xend = Upper), color = "navy") +
  geom_vline(xintercept = lower_critical_value, linetype = "dashed", color = "red") +
  geom_vline(xintercept = upper_critical_value, linetype = "dashed", color = "red") +
  geom_text(aes(x = lower_critical_value + 0.012, y = max(Iteration) + 300, label = "Lower Bound"), col
  geom_text(aes(x = upper_critical_value - 0.012, y = max(Iteration) + 300, label = "Upper Bound"), col
  labs(title = "Confidence Intervals of Bootstrap Samples",
       y = "Iteration",
       x = "Confidence Interval",
       subtitle = "Each line segment represents a 95% confidence interval for a bootstrapped sample") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5, size = 10),
        axis.text.y = element_text(hjust = 1))
```



```
pnorm(0.3, mean = bootstrap_summary$mean, sd = bootstrap_summary$sd, lower.tail = T)
```

```
## [1] 3.74005e-292
```

Based on the analysis conducted, the vaccine efficacy (ψ) for the BNT162b2 vaccine was estimated using a bootstrap method with 10,000 iterations, ensuring accurate and stable estimates. The observed vaccine efficacy from the original data was calculated using the given formula, resulting in an estimated ψ of approximately 95%.

The bootstrap resampling provided a distribution of ψ values, from which the mean efficacy was estimated to be 0.9508 with a standard deviation of 0.0178. The 95% confidence interval for the vaccine efficacy, derived from the bootstrap samples, ranged from approximately 91.59% to 98.58%. This confidence interval is quite narrow, indicating a high level of precision in the efficacy estimate.

The histogram of the bootstrap ψ values, along with the confidence intervals, confirms that the estimated efficacy is consistently around 95%, supporting the robustness of the estimate. Furthermore, the probability that the vaccine efficacy exceeds 30% is 1, with the calculated p-value being effectively zero (3.74×10^{-292}). This overwhelming evidence suggests that the vaccine significantly reduces the risk of infection compared to the placebo.

In summary, the results from the bootstrap analysis strongly support the conclusion that the BNT162b2 vaccine is highly effective, with an estimated efficacy of around 95%. The narrow confidence interval and extremely low p-value reinforce the reliability and significance of the vaccine's protective effect against COVID-19.