Simple and Effective Confidence Intervals for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures

Authors: Jueyao Liu, Vanessa Yang, Zhening Zhang

Abstract:

The *plus four* method is well-known as a good adjustment to the *Wald interval* which has poor performance as the coverage probability is lower than expected. Agresti's study showed that for a 95% confidence interval, the average coverage probability of Wald interval is less than 93%, whereas the coverage probability for plus four interval tends to be above 95%. We further investigated the plus four interval using one-sample and two-sample proportions in large samples and unbalanced samples, and it still performs better than the Wald interval. For other common levels of confidence interval, the plus four interval at 90% confidence interval performs slightly worse than *Wilson score interval*, but still better than Wald interval, while at the 99%, it performs even better than the score interval.

Keywords: Wald interval, Wilson score interval, Plus four interval, Binomial distribution, Large sample size

Introduction:

Many introductory statistics classes teach the Wald interval when calculating 95% of confidence intervals, even though studies have shown that it does not have good coverage. According to Alan Agresti and Brian Caffo (2000), elementary statistics courses almost always teach confidence intervals derived from inverting the large sample Wald's test, even though it underperforms with actual coverage probability being much lower. They investigated a new method of estimating the confidence interval, which was later well-known as "the plus four method." In the study, Agresti tested 95% plus four confidence intervals and whether the method could be used to teach elementary statistics courses.

The plus four method works by adding two successes and two failures to the sample. Agresti performed simulations to show that the 95% plus four interval has better coverage than Wald interval and consistently performed well compared to other good intervals, including the score interval for one-sample tests, and the *Newcombe hybrid interval* (Newcombe, 1998) for two-sample tests. In addition, the plus four method is easy to understand for elementary courses. The score test has good accuracy but is difficult to understand, as the formula is complex and requires calculus and substantial statistical knowledge. The plus four method, in contrast, is easy to remember without a strong statistical background. In light of the assigned paper, we delved into the plus four method and tested if the conclusion of Agresti's study holds for large sample size and different confidence levels.

Description:

The Wald interval is based on the asymptotic normality of the sample proportion, which evaluates the standard errors at the maximum likelihood estimates.

(1) $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where $\hat{p} = \frac{X}{n}$

(2) $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$, where $\hat{p}_i = \frac{X_i}{n_i}$

For one-sample tests, (1) is the 100(1-α)% confidence interval for *p*. For two-sample tests, (2) is the 100(1-α)% confidence interval for $p_1 - p_2$, the difference between the two proportions.

It has been proven that this method for finding confidence intervals behaves poorly (Agresti, 2000). So an alternative solution was proposed, by inverting the test with standard error evaluated at the null hypothesis. This confidence interval (3) is the Wilson score interval for one-sample tests, which also claims to give a 100(1-α)% confidence interval for *p*.

$$(3) \quad \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + \frac{z^2_{\alpha/2}}{n}}, \text{ where } \hat{p} = \frac{X}{n}$$

The midpoint shrinks the sample proportion towards 0.5, and the shrinking becomes less severe as n increases.

However, the score interval is hard to understand at the introductory level, because it requires the quadratic equation to derive. Agresti and Coull (1998) proposed the "Add two successes and two failures and then use the Wald interval," or the "plus four interval". The midpoint of this interval, $p = (X + 2)/(n + 4)$, is nearly identical to the midpoint of the 95% Score interval. This is because the score interval adds $z^2_{\alpha/2}$ to the original sample size, and $z^2_{\alpha/2}$ at 95% confidence interval is $1.96^2 \approx 4$. Therefore, for the 95% confidence level, it has the form

$$(4) \quad \tilde{p} \pm z_{0.025}\sqrt{\tilde{p}(1 - \tilde{p})/\tilde{n}}, \text{ where } \tilde{n} = n + 4 \text{ and } \tilde{p} = (X + 2)/(n + 4)$$

This plus four interval behaves surprisingly well, even for small sample sizes. In addition, compared to the score interval, it does not have spikes with seriously low coverage where p is close to 0 or 1.

Our assigned paper further discussed the advantages of the plus four in the two-sampled test, with formula (5).

(5) $(\tilde{p}_1 - \tilde{p}_2) \pm z_{0.025}\sqrt{\tilde{p}_1(1 - \tilde{p}_1)/\tilde{n}_1 + \tilde{p}_2(1 - \tilde{p}_2)/\tilde{n}_2}$, where $\tilde{n}_i = n_i + 2$ and $\tilde{p}_i = (X_i + 1)/(n_i + 2)$

Agresti first tested the performance for adding different numbers of pseudo observations. It turned out that when adding four pseudo observations, one success and one failure to each of the two samples, the adjusted confidence interval has a coverage rate closest to 95%. To illustrate, let p_1 and p_2 be the proportions for the two samples, and fix n_1 and n_2 equal to 20, Agresti calculated the coverage of 95% confidence interval of p_1 when $p_2$is in the range of 0.1 to 0.5. The result shows that the coverage of 95% Wald interval tends to be smaller than 93%, whereas the plus four interval consistently has around 95% coverage probability. Also, as p_2 moves toward 0 or 1, the coverage probability of 95% Wald interval tends becomes very small and converges to 0. Then, Agresti fixed p_2 and changed the values of n_1 and n_2, and the result had the same conclusion. Even when the total sample size $n_1 + n_2$ increases, the coverage probability of Wald interval is still less than 93%, whereas the adjusted Wald interval is equal or a little bit greater than 95%. Therefore, for a two-sampled test, the coverage probability of 95% plus four interval performs better than the 95% Wald interval.

In addition, Agresti compared the plus four method to other intervals, for example, a hybrid interval from one-sample Wilson score interval for p_1 and p_2, proposed by Newcombe (1998). The hybrid score interval for a 100(1-α)% confidence interval for $p_1 - p_2$ is

$$(6)\ (\hat{p}_1 - \hat{p}_2) - \sqrt{(\hat{p}_1 - L_1)^2 + (U_2 - \hat{p}_2)^2},$$

$$(\hat{p}_1 - \hat{p}_2) + \sqrt{(U_1 - \hat{p}_1)^2 + (\hat{p}_2 - L_2)^2},$$ where $\hat{p}_i = \frac{X_i}{n_i}$ and (L_i, U_i) are the lower and upper bounds for the Wilson score interval (3) for p_i

One difference of the hybrid interval from the Wald interval is that it cannot produce overshoot, which means the interval for $p_1 - p_2$ does not extend below -1 or above + 1.

Compared to the Wald interval and other good intervals, the plus four interval is easy to understand in elementary statistical classes. For introductory level, the Wald interval is confusing in certain situations when p =0 or 1. If the number of successes is 0 in a sample, the 95% Wald interval is [0,0], which does not make any sense for most students. However, if we applied the plus four interval, we can get a plausible 95% confidence interval (an interval with non-zero range), since we add two successes into the sample. This illustrates why Agresti highly supports the plus four method to be taught in elementary statistical classes.

Example:

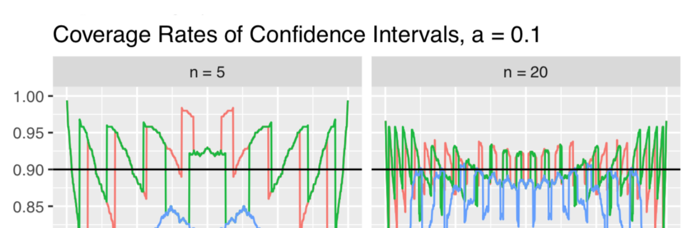
It is well-documented in the assigned paper and in other literature (Agresti & Coull, 1998) that the 95% confidence intervals for one- and two-sample tests derived from Wald tests have poor performance in terms of coverage rates, with the actual coverage probability being much lower than intended.

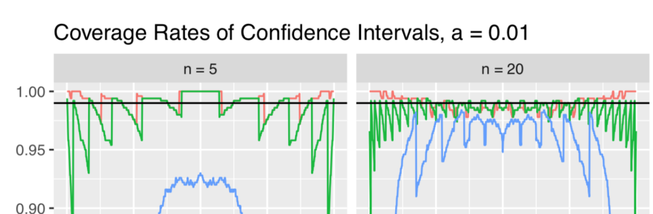
We want to investigate the performance of the plus four intervals at other common levels of significance, namely $\alpha = 0.1$(90% CI) and $\alpha = 0.01$(99% CI), in comparison with other methods of confidence intervals noted in the paper.

Testing One-Sample Confidence Intervals at Different α *Levels*

We simulated 500 datasets from a binomial distribution at different *n* values and calculated the coverage rates as a function of the true value of *p*. We are testing the Wald and the Wilson score intervals along with the plus four. First we test at $\alpha = 0.05$to verify the results of the assigned paper. The results are as expected and match with Agresti and Caffo's findings. The score and plus four intervals work quite well, and the Wald interval underperforms for small *n* values, but all three have asymptotically similar results as n increases. However, note that for *p* values close to 0 or 1, the Wald drastically underperforms even for large *n*.

Now we use the same 500 datasets and calculate the coverage rates at$\alpha = 0.1$and $\alpha = 0.01$.



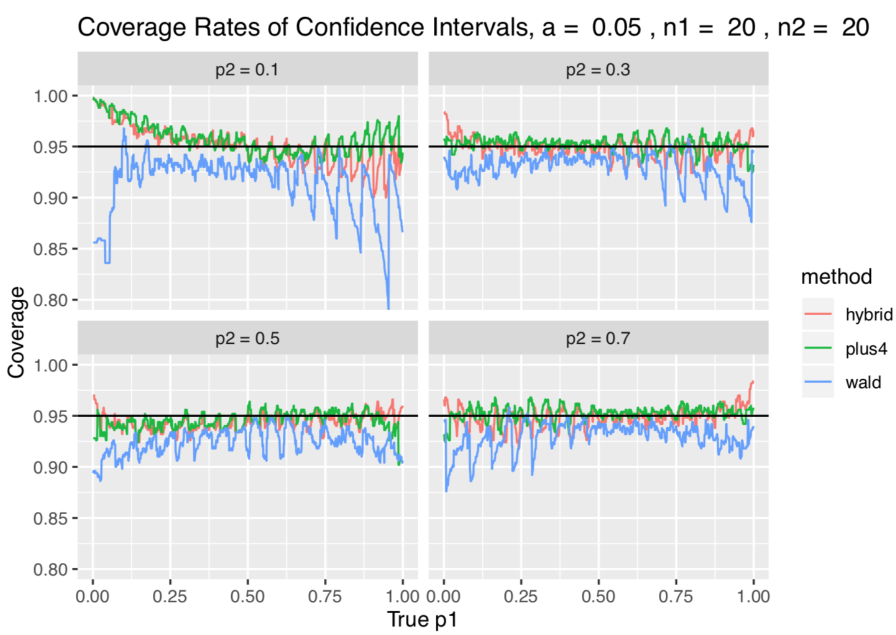


At the 90% confidence interval (Upper-left figure), the plus four now performs worse than the score interval for small *n*, though still better than the Wald interval. However, for small *n*, the plus four interval performs noticeably better than the Score interval at the 99% confidence level (Upper-right figure).

Testing Two-Sample Confidence Intervals at Different α *Levels*

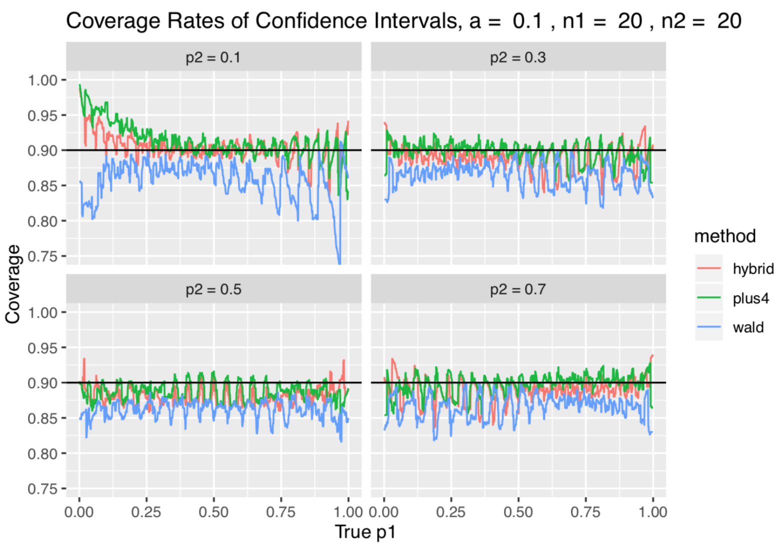
For two sample confidence intervals, we simulated two groups of 500 datasets from binomial(n_1, p_1) and binomial(n_1, p_2). The confidence interval attempts to estimate the difference $p_1 - p_2$. We hold p_2 fixed and calculate the coverage rate as a function of p_1. We also set $n_1 = n_2 = 20$, as the differences between intervals are more pronounced for smaller *n* values. We are testing the Wald and the hybrid score (Newcome Score) intervals along with the plus four.

First we test at the $\alpha = 0.05$to level as we did for the one sample tests.



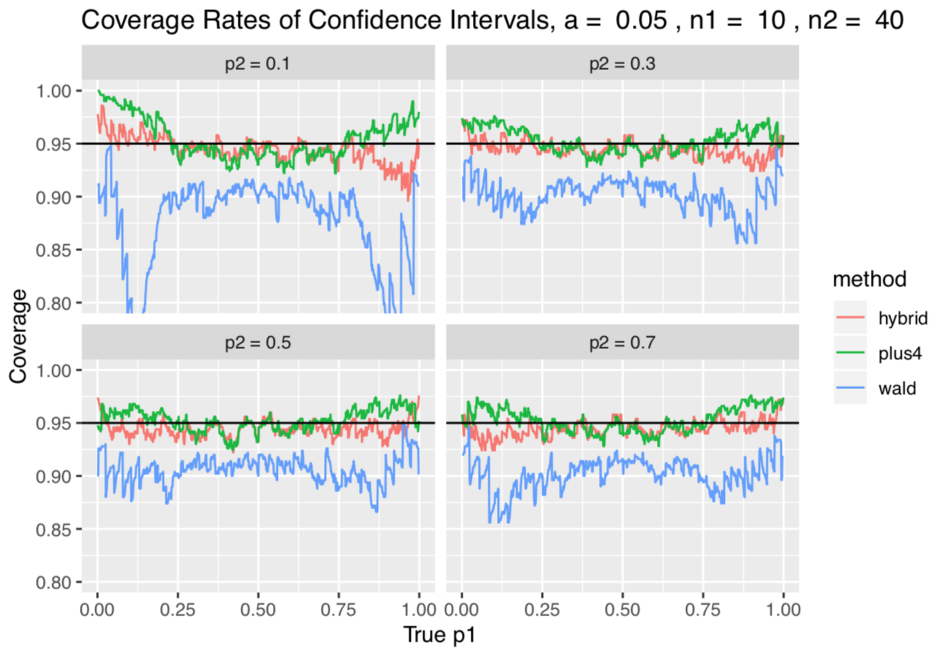
As expected from the reading, the score and plus four intervals work quite well for α = 0.05 and the Wald underperforms for p_2 close to 0 or 1, though all three intervals perform worse for those extreme p_2.

Now we use the same datasets and calculate the coverage rates at α = 0.1 and α = 0.01.



Notice that while there is no significant difference between the coverage rates of the hybrid score and the plus four at the 90% confidence level (Upper-left figure), the plus four now underperforms on average even for p_2 near 0.5. When α = 0.01 (Upper-right figure), we can see that not only the plus four performs as well as the hybrid score and, and now the plus four does not perform worse for extreme p_2 and actually does much better than the Hybrid for those values.

In addition, Fargerland et al. (2015) mentioned the Wald interval does not cope well with unequal sample sizes, so we compared the different methods with n_1= 10, and n_2= 40. We found out that the hybrid interval and the plus four interval works better for unbalanced samples



Generally, from the simulations we observed that the adjusted plus four interval performs well for a confidence level of 95%, underperforms as the confidence level lowers to 90%, and performs better when the confidence level grows to 99%.

Discussion:

Given that Agresti only performed simulation of a small sample size in his study, our group was curious about how the 95% plus four interval performs when the sample size is large. Using large sample sizes (100 and 500) for Wald interval, score interval, and plus four interval, the result shows that even though 95% Wald interval's coverage probability becomes higher and closer to 95% when the sample size is increasing, 95% plus four interval and score interval still have better coverage. Therefore, it supports Agresti's conclusion that, for both small and larger sample sizes, we can always use the plus four method as a better substitution of the Wald interval. In addition, the hybrid and score interval cope well with unequal sample sizes, whereas the Wald interval has lower coverage with unbalanced samples.

We also tested the plus four method for a different confidence level. It shows that when α is smaller, that is, the confidence level is larger, the plus four interval converge rate becomes higher than other methods. Thus, the plus four interval works best for a 99% confidence interval for our test, which even has a higher coverage than the score interval. We suggest that future research could investigate the plus four interval for other confidence levels. Finally, in Agresti & Coull's (1998) study (Figure 10 in the Appendix, page 17), although 95% plus four interval consistently has higher coverage probability than 95% Wald interval, the width of 95% Wald interval is narrower than 95% plus four interval when the sample size is small. Statistically, confidence interval should be as narrow as possible to be helpful, while still having the intended coverage rate on average. Therefore, we also suggest future research to investigate the widths of different confidence intervals.

References

Agresti, A., & Coull, B. (1998). Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician, 52*(2), 119-126.

Agresti, A., & Caffo, B. (2000). Simple and Effective Confidence Intervals for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures. *The American Statistician, 54*(4), 280-288.

Newcombe R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in medicine*, *17*(8), 857–872.

Fagerland, M., Lydersen, S., & Laake, P. (2015). Recommended confidence intervals for two independent binomial proportions. *24*(2), 224-254.

Appendix

1) The following contains the code for one-sample simulations.

```
# checks if a given p is contained in an interval
contained <- function(p, interval) {
  if (p >= interval[1] && p <= interval[2]) {
    return(TRUE)
  } else {
    return(FALSE)
  }
}
```

```
# calculates the coverage rate of a given interval at a given alpha level by
# simulating datasets from binom(n, p)
coverage_one <- function(interval, datasets, n, p, alpha) {
  set.seed(1234)
  rsamples <- rbinom(n = datasets, size = n, prob = p)
  int <- sapply(rsamples, FUN = interval, n = n, alpha = alpha)
  int <- split(int, col(int))
  covered <- sapply(int, FUN = contained, p = p)
  coverage <-  sum(covered) / datasets
  return(coverage)
}
```

```
# calculates the wald interval for one sample of size n from a binomial dist
wald_int_one <- function (sample, n, alpha) {
  mle <- sample/n
  z <- qnorm(1 - alpha/2)
  lower <- mle - z * sqrt((mle)*(1-mle)/n)
  upper <- mle + z * sqrt((mle)*(1-mle)/n)
  c(lower, upper)
}
```

```
# calculates the score interval for a given sample of size n from a binomial dist
score_int_one <- function (sample, n, alpha) {
  mle <- sample/n
  z <- qnorm(1 - alpha/2)
  lower <- (mle+((z^2)/(2*n))-z*sqrt(((mle*(1-mle))/n)+((z^2)/(4*n^2))))/(1+((z^2)/n))
  upper <- (mle+((z^2)/(2*n))+z*sqrt(((mle*(1-mle))/n)+((z^2)/(4*n^2))))/(1+((z^2)/n))
  c(lower, upper)
}
```

```r
# calculates the plus-4 interval for a given sample of size n from a binomial dist
plus4_int_one <- function (sample, n, alpha) {
  pi <- (sample + 2)/(n + 4)
  z <- qnorm(1 - alpha/2)
  lower <- pi - z*sqrt((pi*(1-pi))/(n+4))
  upper <- pi + z*sqrt((pi*(1-pi))/(n+4))
  c(lower, upper)
}
```

```r
# functions that graphs simulated coverage rates at a given alpha level
true_p <- seq(0.001, 0.999, 0.001)
simulate_one <- function(datasets, n, alpha) {
  df <- data.frame(
    true_p,
    wald = sapply(true_p,
                  FUN = coverage_one,
                  interval = wald_int_one,
                  datasets = datasets,
                  n = n,
                  alpha = alpha),
    score = sapply(true_p,
                   FUN = coverage_one,
                   interval = score_int_one,
                   datasets = datasets,
                   n = n,
                   alpha = alpha),
    plus4 = sapply(true_p,
                   FUN = coverage_one,
                   interval = plus4_int_one,
                   datasets = datasets,
                   n = n,
                   alpha = alpha))
  df <- gather(df, key = method, value = coverage, -true_p) %>%
        mutate(n = paste("n =", toString(n)))
  return(df)
}

plot_sims_one <- function(data, alpha, ylim) {
  ggplot(data = data,
       aes(x = true_p, y = coverage, group = method, color = method)) +
  geom_line() +
  geom_hline(yintercept = 1 - alpha) +
  coord_cartesian(ylim = c(ylim, 1)) +
  labs(
    title = paste("Coverage Rates of Confidence Intervals, a =", toString(alpha)),
    x = "True p",
    y = "Coverage"
  ) +
  facet_wrap(~n)
}
```

```r
# graphs simulated coverage rates at different alpha = 0.05
alpha <- 0.05
sims <- 500
one_samp_coverage <- rbind(simulate_one(sims, 5, alpha),
                           simulate_one(sims, 20, alpha),
                           simulate_one(sims, 100, alpha),
                           simulate_one(sims, 500, alpha))
one_samp_coverage$n <- factor(one_samp_coverage$n,
                              levels = c("n = 5","n = 20", "n = 100", "n = 500"))
plot_sims_one(one_samp_coverage, alpha, 0.8)
```

```r
# graphs simulated coverage rates at different alpha = 0.1
alpha <- 0.1
one_samp_coverage <- rbind(simulate_one(sims, 5, alpha),
                           simulate_one(sims, 20, alpha),
                           simulate_one(sims, 100, alpha),
                           simulate_one(sims, 500, alpha))
one_samp_coverage$n <- factor(one_samp_coverage$n,
                              levels = c("n = 5","n = 20", "n = 100", "n = 500"))
plot_sims_one(one_samp_coverage, alpha, 0.75)
```

```r
# graphs simulated coverage rates at different alpha = 0.01
alpha <- 0.01
one_samp_coverage <- rbind(simulate_one(sims, 5, alpha),
                           simulate_one(sims, 20, alpha),
                           simulate_one(sims, 100, alpha),
                           simulate_one(sims, 500, alpha))
one_samp_coverage$n <- factor(one_samp_coverage$n,
                              levels = c("n = 5","n = 20", "n = 100", "n = 500"))
plot_sims_one(one_samp_coverage, alpha, 0.85)
```

2) The following contains the code for two-sample simulations.

```r
# calculates the coverage rate of a given interval at a given alpha level by
# simulating datasets from binom(n1, p1)  and the same number of datasets from
# binom(n2, p2)
coverage_two <- function(interval, datasets, n1, n2, p1, p2, alpha) {
  set.seed(1234)
  rsamples1 <- rbinom(n = datasets, size = n1, prob = p1)
  rsamples2 <- rbinom(n = datasets, size = n2, prob = p2)
  int <- mapply(sample1 = rsamples1,
                sample2 = rsamples2,
                FUN = interval,
                MoreArgs = list(n1 = n1, n2 = n2, alpha = alpha))
  int <- split(int, col(int))
  covered <- sapply(int, FUN = contained, p = (p1 - p2))
  coverage <-  sum(covered) / datasets
  return(coverage)
}
```

```r
# calculates the wald interval for one two samples of size n1 and n2 from
# binomial distributions (p1 - p2)
wald_int_two <- function (sample1, sample2, n1, n2, alpha) {
  mle1 <- sample1/n1
  mle2 <- sample2/n2
  z <- qnorm(1 - alpha/2)
  lower <- (mle1 - mle2) - z * sqrt(((mle1*(1-mle1))/n1)+((mle2*(1-mle2))/n2))
  upper <- (mle1 - mle2) + z * sqrt(((mle1*(1-mle1))/n1)+((mle2*(1-mle2))/n2))
  c(lower, upper)
}
```

```r
# calculates the hybrid score interval for one two samples of size n1 and n2
# from binomial distributions (p1 - p2)
score_int_two <- function (sample1, sample2, n1, n2, alpha) {
  mle1 <- sample1/n1
  mle2 <- sample2/n2
  score1 <- score_int_one(sample1, n1, alpha)
  score2 <- score_int_one(sample2, n2, alpha)
  lower <- (mle1 - mle2) - sqrt(((mle1 - score1[1])^2) + ((score2[2] - mle2)^2))
  upper <- (mle1 - mle2) + sqrt(((score1[2] - mle1)^2) + ((mle2 - score2[1])^2))
  c(lower, upper)
}
```

```r
# calculates the plus-4 interval for one two samples of size n1 and n2 from
# binomial distributions (p1 - p2)
plus4_int_two <- function (sample1, sample2, n1, n2, alpha) {
  p1 <- (sample1 + 1)/(n1 + 2)
  p2 <- (sample2 + 1)/(n2 + 2)
  z <- qnorm(1 - alpha/2)
  lower <- (p1 - p2) - z * sqrt(((p1*(1-p1))/(n1+2)) + ((p2*(1-p2))/(n2+2)))
  upper <- (p1 - p2) + z * sqrt(((p1*(1-p1))/(n1+2)) + ((p2*(1-p2))/(n2+2)))
  c(lower, upper)
}
```

```r
# functions that graphs simulated coverage rates at a given alpha level
true_p1 <- seq(0.001, 0.999, 0.001)
simulate_two <- function(datasets, n1, n2, p2, alpha) {
  df <- data.frame(
    true_p1,
    wald = sapply(true_p1,
                  FUN = coverage_two,
                  interval = wald_int_two,
                  datasets = datasets,
                  p2 = p2,
                  n1 = n1,
                  n2 = n2,
                  alpha = alpha),
```

```
    hybrid = sapply(true_p1,
                    FUN = coverage_two,
                    interval = score_int_two,
                    datasets = datasets,
                    p2 = p2,
                    n1 = n1,
                    n2 = n2,
                    alpha = alpha),
    plus4 = sapply(true_p1,
                   FUN = coverage_two,
                   interval = plus4_int_two,
                   datasets = datasets,
                   p2 = p2,
                   n1 = n1,
                   n2 = n2,
                   alpha = alpha))
  df <- gather(df, key = method, value = coverage, -true_p1) %>%
        mutate(p2 = paste("p2 =", toString(p2)))
  return(df)
}

plot_sims_two <- function(data, n1, n2, alpha, ylim) {
  ggplot(data = data,
       aes(x = true_p1, y = coverage, group = method, color = method)) +
  geom_line() +
  geom_hline(yintercept = 1 - alpha) +
  coord_cartesian(ylim = c(ylim, 1)) +
  labs(
    title = paste("Coverage Rates of Confidence Intervals, a = ",
                  toString(alpha),
                  ", n1 = ",
                  toString(n1),
                  ", n2 = ",
                  toString(n2)),
    x = "True p1",
    y = "Coverage"
  ) +
  facet_wrap(~p2)
}
```

```
# graphs simulated coverage rates at different alpha = 0.05, n1 = n2 = 20
alpha <- 0.05
n1 <- 20
n2 <- 20
two_samp_coverage <- rbind(simulate_two(sims, n1, n2, 0.1, alpha),
                           simulate_two(sims, n1, n2, 0.3, alpha),
                           simulate_two(sims, n1, n2, 0.5, alpha),
                           simulate_two(sims, n1, n2, 0.7, alpha))
two_samp_coverage$p2 <- factor(two_samp_coverage$p2,
                              levels = c("p2 = 0.1","p2 = 0.3", "p2 = 0.5", "p2 = 0.7"))
plot_sims_two(two_samp_coverage, n1, n2, alpha, 0.80)
```

```r
# graphs simulated coverage rates at different alpha = 0.1, n1 = n2 = 20
alpha <- 0.1
two_samp_coverage <- rbind(simulate_two(sims, n1, n2, 0.1, alpha),
                           simulate_two(sims, n1, n2, 0.3, alpha),
                           simulate_two(sims, n1, n2, 0.5, alpha),
                           simulate_two(sims, n1, n2, 0.7, alpha))
two_samp_coverage$p2 <- factor(two_samp_coverage$p2,
                              levels = c("p2 = 0.1","p2 = 0.3", "p2 = 0.5", "p2 = 0.7"))
plot_sims_two(two_samp_coverage, n1, n2, alpha, 0.75)
```

```r
# graphs simulated coverage rates at different alpha = 0.01, n1 = n2 = 20
alpha <- 0.01
two_samp_coverage <- rbind(simulate_two(sims, n1, n2, 0.1, alpha),
                           simulate_two(sims, n1, n2, 0.3, alpha),
                           simulate_two(sims, n1, n2, 0.5, alpha),
                           simulate_two(sims, n1, n2, 0.7, alpha))
two_samp_coverage$p2 <- factor(two_samp_coverage$p2,
                              levels = c("p2 = 0.1","p2 = 0.3", "p2 = 0.5", "p2 = 0.7"))
plot_sims_two(two_samp_coverage, n1, n2, alpha, 0.85)
```

```r
# graphs simulated coverage rates at different alpha = 0.05, n1 = n2 = 20
alpha <- 0.05
n1 <- 10
n2 <- 40
two_samp_coverage <- rbind(simulate_two(sims, n1, n2, 0.1, alpha),
                           simulate_two(sims, n1, n2, 0.3, alpha),
                           simulate_two(sims, n1, n2, 0.5, alpha),
                           simulate_two(sims, n1, n2, 0.7, alpha))
two_samp_coverage$p2 <- factor(two_samp_coverage$p2,
                              levels = c("p2 = 0.1","p2 = 0.3", "p2 = 0.5", "p2 = 0.7"))
plot_sims_two(two_samp_coverage, n1, n2, alpha, 0.80)
```

3) Figure 10: the width of different 95% intervals