**Inferences for one proportion**



This section investigates inferences about a population proportion! A probability of success could also be thought of as a proportion too. Thus, we are interested in making inferences about π from a binomial distribution.

To learn more about statistical analyses involved with categorical outcomes beyond these notes, please see my textbook *Analysis of Categorical Data with R*.

Let Y be a random variable representing the number of successes out of n trials. Thus, Y has a binomial distribution with μ = nπ and variance σ2 = nπ(1-π). The statistic  = y/n is the observed proportion of successes that estimates π. The random variable version of the statistic is .

Wald CI for π:

The normal probability distribution can be used to approximate probabilities from a binomial probability distribution. Thus,



has an approximate standard normal distribution for a large n.

Working with this quantity, we can re-write it as



and this also has an approximate standard normal random variable for large n. Thus,



Rearranging some terms in the probability produces,



⇒ ⇒ 

⇒ 

Notice that  is in the lower and upper bounds above. By replacing it with , we obtain the (1-α)100% confidence interval for π:



However, this replacement of π by  causes problems in terms of the confidence interval’s “coverage” level. Thus, we will not use this interval in our class.

Agresti-Coull CI for π:

Let . A (1-α)100% confidence interval for π is given by:



Notes:

* The Agresti-Coull CI is the same as the Wald CI except an adjusted estimate is used for π and the sample size is adjusted.
* Why adjust the estimate for π? It helps to move the estimate away from 0 or 1. In the end, the confidence interval has better “coverage”. This means the true percentage of times the Agresti-Coull CI will contain π will be closer to the (1-α)100% level than the true percentage of times the Wald CI will contain π.
* Suppose α = 0.05 so that a 95% CI is being calculated. Notice that zα/2 = 1.96 ≈ 2. The estimator for π becomes:





Thus, two successes and two failures are added to the observed data!

Wilson CI for π:

We will see shortly that the test statistic for the test
Ho:π = π0 vs. Ha:π ≠ π0 is



which has an approximate normal distribution. This means that P(-zα/2 < Z < zα/2) ≈ 1 – α. We can solve for all possible values of π0 that satisfy

-zα/2 < z < zα/2

to obtain a set of values that form a confidence interval. Using the quadratic formula, one can derive the (1-α)100% confidence interval for π to be:



Note: This interval is sometimes called the “score” CI to reflect the statistical procedure being used by Wilson to derive the interval.

Which interval should you use?

* Do not use the Wald CI
* Overall, the Wilson CI is better than the Agresti-Coull CI in most situations except perhaps when π is very close to 0 or 1. However, Agresti-Coull is a little more convenient to calculate and its performance is not much worse than Wilson, especially for larger sample sizes. “Performance” here is with respect to “coverage” (percent of time an interval truly contains the parameter of interest).
* The Wilson interval always has limits between 0 and 1. The Wald and Agresti-Coull intervals do not!
* The performance of the Wilson interval can be improved by changing the lower limit to

–log(1 – α)/n

when y = 1 and changing the upper interval limit to

1 + log(1 – α)/n

when y = n – 1. Brown et al. (2001, p. 112) provide justification.

* There are many other confidence intervals for π!

Example: HCV (HCV.R)

Blood donations need to be screened for infectious diseases to prevent the spread from donor to recipient. Organizations that collected these donations monitor the infection rates among their donors.

Liu et al. (*Transfusion*, 1997) reported results on 1,875 blood donors screened for Hepatitis C (HCV) at the Blood Transfusion Service in Xuzhou City, China. There were 42 positive blood donors found.

Find the estimate for π and the 95% CI for the population proportion of blood donors in Xuzhou City, China that are positive.

The estimate for π is .

Let

.

The 95% Agresti-Coull confidence interval for π is:



⇔

⇔ 0.0165 < π < 0.0302

I am 95% confident that the population proportion of blood donors in Xuzhou City, China that are HCV positive is between 0.0165 and 0.0302.

The 95% Wilson confidence interval for π is given by:

⇔

⇔ 0.0166 < π < 0.0301

I am 95% confident that the population proportion of blood donors in Xuzhou City, China that are HCV positive is between 0.0166 and 0.0301.

Below are my calculations from R:

> y <- 42

> n <- 1875

> alpha <- 0.05

> pi.hat <- y/n

> #Wald CI – will not use

> lower.wald <- pi.hat - qnorm(p = 1-alpha/2, mean = 0, sd

 = 1) \* sqrt(pi.hat\*(1-pi.hat)/n)

> upper.wald <- pi.hat + qnorm(p = 1-alpha/2, mean = 0, sd

 = 1) \* sqrt(pi.hat\*(1-pi.hat)/n)

> data.frame(lower.wald, upper.wald)

 lower.wald upper.wald

1 0.01570189 0.02909811

> #Agresti-Coull CI

> # Remember that mean = 0 and sd = 1 are the default

 argument values for qnorm().

> pi.tilde <-( y + qnorm(p = 1-alpha/2)^2 /2) / (n +

 qnorm(p = 1-alpha/2)^2)

> pi.tilde

[1] 0.0233765

> lower.AC <- pi.tilde - qnorm(p = 1-alpha/2) \*

 sqrt(pi.tilde\*(1-pi.tilde) / (n + qnorm(p = 1-

 alpha/2)^2))

> upper.AC <- pi.tilde + qnorm(p = 1-alpha/2) \*

 sqrt(pi.tilde\*(1-pi.tilde) / (n +q norm(p = 1-

 alpha/2)^2))

> data.frame(lower.AC, upper.AC)

 lower.AC upper.AC

1 0.01654436 0.03020863

> #Wilson CI

> lower.wilson <- pi.tilde - qnorm(1-alpha/2) \* sqrt(n) /

 (n + qnorm(1-alpha/2)^2) \* sqrt(pi.hat\*(1-pi.hat) +

 qnorm(1-alpha/2)^2 / (4\*n))

>upper.wilson <- pi.tilde + qnorm(1-alpha/2) \* sqrt(n) /

 (n + qnorm(1-alpha/2)^2) \* sqrt(pi.hat\*(1-pi.hat) +

 qnorm(1-alpha/2)^2 / (4\*n))

> data.frame(lower.wilson, upper.wilson)

 lower.wilson upper.wilson

1 0.01661436 0.03013863

Below are two easier ways to perform these calculations in R:

> library(binom) #Need to install the package

> binom.confint(x = y, n = n, conf.level = 1-alpha, methods

 = "agresti-coull")

 method x n mean lower upper

1 agresti-coull 42 1875 0.0224 0.01654436 0.03020863

> binom.confint(x = y, n = n, conf.level = 1-alpha, methods

 = "wilson")

 method x n mean lower upper

1 wilson 42 1875 0.0224 0.01661436 0.03013863

> #This function also calculates the Wilson interval

> prop.test(x = y, n = n, conf.level = 0.95, correct =

 FALSE)

 1-sample proportions test without continuity correction

data: y out of n, null probability 0.5

X-squared = 1710.763, df = 1, p-value < 2.2e-16

alternative hypothesis: true p is not equal to 0.5

95 percent confidence interval:

 0.01661436 0.03013863

sample estimates:

 p

0.0224

Questions:

* Suppose Xuzhou City officials hypothesize that HCV prevalence was only 0.01. What do you think about the correctness of their hypothesis?
* Suppose Xuzhou City officials hypothesize that HCV prevalence was only 0.02. What do you think about the correctness of their hypothesis?

What about hypothesis testing?

Tests can be performed using confidence intervals, test statistics, and p-values in a similar manner as we have seen before. We just now use different formulas for confidence intervals and test statistics than we did in the last chapter.

Example: HCV (HCV.R)

Suppose Xuzhou City officials say that HCV prevalence is 0.01. What do you think about the correctness of their statement? Use α = 0.05 to perform a hypothesis test in order to examine their statement.

Hypothesis test using the CI method:

1. Ho:π = 0.01
 Ha:π ≠ 0.01
2. Agresti-Coull CI: 0.0165 < π < 0.0302. If the Wilson CI was used instead: 0.0166 < π < 0.0301.
3. Because 0.01 is not in the CI, reject Ho.
4. There is sufficient evidence to conclude that the proportion of blood donors in Xuzhou City with HCV is not 0.01.

Of course, you could also do the hypothesis test using the test statistic and p-value methods.

We can write the hypotheses for a two-tail test as

Ho:π = π0
Ha:π ≠ π0

and left-tail or right-tail tests would have the usual adjustments.

Test statistic: Earlier, we saw that



What is the test statistic?



What are the critical values for two-tail, left-tail, and right-tail tests?

P-value for two-tail test: 2×P(Z > |z|)

What is the p-value for a left-tail or right-tail test?

When is the normal probability distribution APPROXIMATION for z good enough to be used?

Many books say nπ0 ≥ 5 and n(1 – π0) ≥ 5. Unfortunately, it is not that simple. In fact, this recommendation has been perpetuated in books, but no one can find the original person who came up with this recommendation!

My recommendation is to look at the sample size AND π0 value. For moderate to large sample sizes, this approximation will work well as long as π0 is not very close to 0 or 1. For small sample sizes and/or π0 being close to 0 or 1, it is best to use an “exact inference” procedures. These procedures are discussed in categorical data analysis courses.

Example: HCV (HCV.R)

Suppose Xuzhou City officials say that HCV prevalence is 0.01. What do you think about the correctness of their statement? Use α = 0.05 to perform a hypothesis test in order to examine their statement.

Test statistic method:

1. Ho:π = 0.01
 Ha:π ≠ 0.01
2. Note that  

> pi0 <- 0.01

> z.stat <- (pi.hat - pi0)/sqrt(pi0\*(1-pi0)/n)

> z.crit <- qnorm(p = 1-alpha/2, mean = 0, sd = 1)

> pvalue <- 2\*(1 - pnorm(q = abs(z.stat), mean = 0, sd =

 1))

> data.frame(pi.hat, pi0, z.stat, z.crit, pvalue)

 pi.hat pi0 z.stat z.crit pvalue

1 0.0224 0.01 5.396407 1.959964 6.79885e-08

1. ±z0.025 = ±1.96
2.



Because 5.3964 > 1.96, reject Ho.

1. There is sufficient evidence to conclude that the proportion of blood donors in Xuzhou City with HCV is not 0.01.

P-value method:

1. Ho:π = 0.01
 Ha:π ≠ 0.01
2. 2×P(Z > |5.3964|) = 6.80×10-8

The p-value could be found through integration:

 = 6.80×10-8

Instead, we will use R:

> 2\*(1 - pnorm(q = abs(z.stat), mean = 0, sd = 1))

[1] 6.79885e-08

1. α = 0.05
2. Because 6.80×10-8 < 0.05, reject Ho.
3. There is sufficient evidence that the proportion of blood donors in Xuzhou City with HCV is not 0.01.

Another way to perform the calculation in R:

> prop.test(x = y, n = n, p = 0.01, conf.level = 0.95,

 correct = FALSE, alternative = "two.sided")

 1-sample proportions test without continuity correction

data: y out of n, null probability 0.01

X-squared = 29.1212, df = 1, p-value = 6.799e-08

alternative hypothesis: true p is not equal to 0.01

95 percent confidence interval:

 0.01661436 0.03013863

sample estimates:

 p

0.0224

The test statistic given is called X-squared and this can be compared to a chi-square distribution critical value (to be discussed later in the course). The square root of the statistic is our z, where the + or - sign of z is the sign of . This equivalence result would be shown in a mathematical statistics class where you would learn that a squared standard normal random variable is the same as a chi-square random variable with 1 degree of freedom.

Notes:

* The confidence interval method can give a different hypothesis test decision than the test statistic and p-value methods if the Agresti-Coull CI is used. However, the chances of obtaining a different decision are small.
* What if the data was not already summarized into the number of success out of a particular number of trials? There are a number of ways that the “counts” can be found. One way is to use the table() function in R. Below is an example in HCV.R:

> #Create 1's for HCV positive and 0's for HCV negative

> raw.data <- c(rep(x = 1, times = y), rep(x = 0, times = n

 – y))

> head(raw.data)

[1] 1 1 1 1 1 1

> tail(raw.data)

[1] 0 0 0 0 0 0

> #This may be what the data looked like coming in

> set.seed(9819)

> raw.data2 <- sample(raw.data, replace = FALSE)

> head(raw.data2)

[1] 1 0 0 0 0 0

> tail(raw.data2)

[1] 0 0 0 0 0 0

> #Find counts

> table(raw.data2)

raw.data2

 0 1

1833 42

> #This works too because 1's and 0's are used

> sum(raw.data2)

[1] 42

> y <- table(raw.data2)[2]

> n <- table(raw.data2)[1] + table(raw.data2)[2]

> prop.test(x = y, n = n, p = 0.01, conf.level = 0.95,

 correct = FALSE, alternative = "two.sided")

 1-sample proportions test without continuity correction

data: y out of n, null probability 0.01

X-squared = 29.121, df = 1, p-value = 6.799e-08

alternative hypothesis: true p is not equal to 0.01

95 percent confidence interval:

 0.01661436 0.03013863

sample estimates:

 p

0.0224