# Project 1

David Nguyen

2021-02-17

## 1. COVID tests (23 total points)

There are three types of COVID-19 tests: 1) RT-PCR, 2) antigen, and 3) antibody. Each type needs to have high sensitivity (probability test gives a positive result given individual is truly positive). The RT-PCR type is considered to have the highest sensitivity, but it also takes the longest amount of time to perform. The antigen test is used as an alternative because it takes much less time. Unfortunately, this test can be less accurate than the RT-PCR. The purpose of this problem is to examine the sensitivity of the antigen test.

The BD Veritor System is an antigen test that received an Emergency Use Authorization (EUA) from the Food and Drug Administration (FDA) on July 2. The clinical performance of the test is summarized at https://www.fda.gov/media/139755/download (see p. 12). The stated "sensitivity" is based on the correct identification of 26 positive specimens out of 31 known positive specimens. Using this data, complete the following.

i) (2 points) Define $\hat{\pi}$ as the estimated sensitivity. Compute this value.

```
w <- 26
n <- 31
sensitivity <- w/n
```

$\hat{\pi} = 0.84$

ii) (5 points) Calculate a 95% confidence interval for the sensitivity . While normally only one interval would be calculated in practice, calculate the interval using the Wilson, Agresti-Coull, and Clopper-Pearson expressions. Compare the intervals.

```
binom.confint(x = w, n = n, conf.level = 0.95,
              methods = c("wilson", "ac", "exact")) %>% # Clopper-Peason is "exact"
  mutate(width = upper - lower) %>%
  knitr::kable(digits = 3)
```

| method | x | n | mean | lower | upper | width |
|---|---|---|---|---|---|---|
| agresti-coull | 26 | 31 | 0.839 | 0.669 | 0.934 | 0.265 |
| exact | 26 | 31 | 0.839 | 0.663 | 0.945 | 0.283 |
| wilson | 26 | 31 | 0.839 | 0.674 | 0.929 | 0.255 |

The confidence intervals for all methods are wider than 0.25, which indicates that our estimate of the sensitivity of the BD veritor test is quite uncertain. The Wilson interval is narrowest, then

the Agresti-Coull, and the Clopper-Pearson (CP) is widest. It makes sense that the CP interval is widest since it *always* has coverage $\geq 1-\alpha$ (i.e., meets or exceeds the nominal confidence level), whereas the Wilson and Agresti-Coull can have true coverage probabilities lower than the nominal coverage especially when the $\hat{\pi}$ is near 0 or 1 (0.84 in this case).

iii. Interpret the Wilson interval from ii in the context of the test.

We are 95 % confident that the sensitivity, the probability that a sample that tested positive using RT-PCR will test positive using BD Veritor, is between 0.67 - 0.93.

iv) (3 points) The confidence interval for the sensitivity is quite wide. Suppose a larger sample size is taken and $\hat{\pi}$ remains the same to two decimal places. What is approximately the smallest sample size that would result in a Wilson interval no wider than 0.05? You may use a trial and error method to find this sample size.

```
# function to compute the wilson interval at a desired width given n and p
# will return ~ 0 when at desired width
wilson_width <- function(n, p, alpha = 0.05, width = 0.05) {
  # compute corrected probability of success
  w <- round(n*p)
  pi.hat <- w/n
  p.tilde <- (w + qnorm(p = 1-alpha/2)^2 /2) / (n+qnorm(1-alpha/2)^2)
  # compute lower and upper wilson
  lower.wilson <- p.tilde - qnorm(p = 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 <- p.tilde + qnorm(p = 1-alpha/2) * sqrt(n) /
    (n+qnorm(1-alpha/2)^2) * sqrt(pi.hat*(1-pi.hat) + qnorm(1-alpha/2)^2/(4*n))
  # get width of interval - desired width
  return(upper.wilson - lower.wilson - width)
}

# find minimum sample size to get desired width
n_minimum <- uniroot(wilson_width, c(31, 1000), p = sensitivity)$root %>% ceiling()

# get confidence interval with n_minimum
binom.confint(x = round(sensitivity*n_minimum), n = n_minimum, methods = "wilson") %>%
  mutate(width = upper - lower) %>%
  knitr::kable(digits = 3)
```

| method | x | n | mean | lower | upper | width |
|---|---|---|---|---|---|---|
| wilson | 697 | 831 | 0.839 | 0.812 | 0.862 | 0.05 |

The minimum necessary sample size to get a Wilson confidence interval of width 0.05 is 831.

v) (3 points) Examine the interval given for the positive percent agreement (PPA) measure on p. 12 (Table 1) of the download from the FDA. BD uses this value as the sensitivity of the test (see their definition of PPA and Table 2). What confidence level and interval did BD use in their calculations? If you cannot decide on only one interval, list all intervals that it could be.

```r
false_neg <- 5
ppa <- w/(w + false_neg)
ci_ppa_bd <- c(0.67, 0.93)
ci_bd_print <- paste(round(ppa,2), "(" , ci_ppa_bd[1], ", " , ci_ppa_bd[2], ")")
```

The estimate of PPA (sensitivity) reported by BD is 0.84 ( 0.67 , 0.93 ).

```r
# Confidence levels to try
try_conf <- seq(0.8, 0.99, by = 0.01)

# get intervals for all confidence levels for all methods
# extract methods that gave CI within 2 digits of BD's interval
lapply(seq_along(try_conf), function(x)
  binom.confint(x = w, conf.level = try_conf[x], n = (false_neg + w), methods = "all" ) %>%
  add_column(`confidence level` = try_conf[x])) %>%
  bind_rows() %>%
  mutate(lower = round(lower, 2),
         upper = round(upper, 2)) %>%
  filter(near(lower, ci_ppa_bd[1]), near(upper, ci_ppa_bd[2])) %>%
  knitr::kable(digits = 2)
```

| method | x | n | mean | lower | upper | confidence level |
|---|---|---|---|---|---|---|
| logit | 26 | 31 | 0.84 | 0.67 | 0.93 | 0.94 |
| agresti-coull | 26 | 31 | 0.84 | 0.67 | 0.93 | 0.95 |
| logit | 26 | 31 | 0.84 | 0.67 | 0.93 | 0.95 |
| wilson | 26 | 31 | 0.84 | 0.67 | 0.93 | 0.95 |

The above table gives methods and confidence levels that could have been used to obtain the confidence interval in the BD report. To obtain this table, I searched over confidence levels between 0.8-0.99 by increments of 0.01 and used all available methods in binom::binom_confint(). Then, I filtered out the combinations of method and confidence level which yielded intervals within two decimal places of the one reported in the BD Veritor document.

vi. (3 points) Why is it incorrect for BD not to state a confidence level in their document?

If the confidence level is not reported it will be unclear to the reader how they should interpret the interval estimate. For instance, an unscrupulous company could report an interval with low confidence level, say 50 % confidence level, to get a narrow confidence interval. If they do not report the confidence level they used, it will likely be misleading to the reader who will probably assume a 95 % confidence level based on convention.

vii. (3 points) Fitzpatrick et al. (The Lancet, 2021, p. 24-5) warned readers of inflated accuracy claims by manufacturers of these antigen type tests. Read this paper and summarize the problem examined in this paper.

Fitzpatrick et al. explained that manufacturers of rapid antigen tests (including BD Veritor) have overstated the sensitivity of their tests because they assumed that the positive results from RT-PCR tests were “true positives” even though it is known that RT-PCR is imperfect (produces false negatives). To understand how the sensitivity of the RT-PCR tests used as a reference by rapid antigen test manufacturers, we would need to know the sensitivity of the RT-PCR test

they used as a reference which was not reported. Also, FDA emergency use authorizations (EUA) only require sample sizes of at least 30 positive cases which is not large enough to obtain precise estimates of PPA.

viii. (2 points) Suppose you had a test performed by the BD Veritor system. Would you be comfortable with the accuracy of your test result?

I would not feel comfortable with the accuracy of my test result because of the wide confidence interval of their PPA estimate and because I do not know the accuracy of the RT-PCR test they used as a reference.

## 2. In or out?

(28 total points) The two major golf governing bodies, the United States Golf Associations and the R&A, modified a number of rules for golf play in 2019. One of these changes gave golfers the choice between leaving the flagstick in the hole while putting or taking it out. Golfers had always been required to take the flagstick out of the hole prior to the rule change. Now, with the choice, golfers would like to know what strategy–flagstick in or flagstick out–will result in a larger probability of success.

A number of golfing groups performed experiments in 2019 to determine the best strategy. Interestingly, these groups did not all come to the same conclusion! Bilder (Chance, 2020, v. 4, p. 56-61) examined data from these experiments to resolve these differences. We will focus in this project on the data in Table 1 of the paper corresponding to slightly off-center putts that approach the hole at a medium speed:

```
# create data set
golf <- data.frame(flagstick = c(rep("out", 73), rep("in", 45), rep("out", 27), rep("in", 55)),
                   outcome = c(rep("success", 73 + 45), rep("failure", 27 + 55)))
# print contingency table
c.table <- golf %>% table() %>% addmargins()
c.table
```

```
##          outcome
## flagstick failure success Sum
##       in       55      45 100
##       out      27      73 100
##       Sum      82     118 200
```

a) (3 points) The Edoardo Molinari Golf Academy performed the experiment that resulted in the data above. Based on the observed proportions only, they developed their own conclusions. No statistical inference methods were used. Why is this a poor way to develop conclusions?

While the sample proportion of success is an estimate of the probability of success ($\hat{\pi}$), drawing conclusions from point estimates without any consideration of the uncertainty of the estimate can lead to conclusions that may not be robust to sampling error.

b) (5 points) Find the 95% Agresti-Caffo confidence interval for the difference in the success probabilities. Interpret the interval in the context of the data problem.

```
wald2ci(x1 = c.table["in","success"], n1 = c.table["in","Sum"],
        x2 = c.table["out","success"], n2 = c.table["out","Sum"],
        conf.level = 0.95, adjust = "AC")
```

```
##
##
##
## data:
##
## 95 percent confidence interval:
##  -0.4042220 -0.1447976
## sample estimates:
## [1] -0.2745098
```

The 95 % Agresti-Caffo confidence interval for the difference in success probability ($\hat{\pi}_{in} - \hat{\pi}_{out}$) is -0.27 (-0.40, -0.14). This means we are 95 % confident that the probability of a successful put when the flagstick is in the hole is between 0.14 - 0.4 less than when the flagstick is out.

c) (5 points) The paper uses a score confidence interval rather than an Agresti-Caffo confidence interval. A score confidence interval for the difference between success probabilities in general is explained in Exercise #24 on p. 57 of my book. Using the information available for this exercise, compute the 95% score confidence interval for the difference in the success probabilities. Compare this interval to what was obtained for part b).

```
diffscoreci(x1 = c.table["in","success"], n1 = c.table["in","Sum"],
        x2 = c.table["out","success"], n2 = c.table["out","Sum"],
        conf.level = 0.95)
```

```
##
##
##
## data:
##
## 95 percent confidence interval:
##  -0.4050691 -0.1450295
```

The 95 % confidence score interval for difference in success probabilities is -0.41 to -0.15 which is almost the same as the Agresti-Caffo interval (-0.40 to -0.14). The conclusion drawn from either interval is the same: we are 95 % confident that the true difference in putting success probability excludes zero and that taking out the flagstick is better than leaving it in.

d) (4 points) Perform the score test for the difference in probabilities of success. Discuss how your conclusions agree with those from the score confidence interval.

```
pi.hat1 <- (c.table["in", "success"]/c.table["in","Sum"])
pi.hat2 <- (c.table["out", "success"]/c.table["out","Sum"])
pi.bar <- sum(c.table[c("in","out"), "success"])/sum(c.table[c("in","out"), "Sum"])
n1 <- c.table["in","Sum"]
n2 <- c.table["out","Sum"]

# score test statistic for pi_1 - pi_2
Z_0 <- (pi.hat1 - pi.hat2) / sqrt(pi.bar * (1 - pi.bar) * (1/n1 + 1/n2))

# crit Z value and p value
Z_crit <- qnorm(1 - 0.05/2)
p.value <- 1 - pnorm(abs(Z_0))
```

Based on a 0.05 level score test of the hypotheses $H_0 : \hat{\pi}_{in} - \hat{\pi}_{out} = 0$ vs $H_1 : \hat{\pi}_{in} - \hat{\pi}_{out} \neq 0$ we conclude that the difference between the probabilities of success is significantly different from zero ($Z_0 = $ -4.026, p = $2.8 \times 10^{-5}$). This is consistent with the inference from the score interval which did not include zero.

e) (4 points) Estimate the relative risk and calculate the corresponding 95% confidence interval for it. Interpret the results using the phrasing discussed in the course notes.

```
RR.hat <- pi.hat1 / pi.hat2

alpha <- 0.05
var.log.rr <- (1-pi.hat1) / (n1*pi.hat1) + (1-pi.hat2) / (n2*pi.hat2)
RR.CI <- exp(log(pi.hat1/pi.hat2) + qnorm(p = c(alpha/2, 1- alpha/2)) * sqrt(var.log.rr))
round(RR.CI, 2)
```

```
## [1] 0.48 0.79
```

The 95 % CI is 0.48 < RR < 0.79. Therefore, with 95 % confidence, the probability of a successful putt is 0.48 to 0.79 times as large when the flagstick is in compared to when the flagstick is removed.

f) (4 points) Estimate the odds ratio and calculate the corresponding 95% confidence interval for it. Interpret the results using the phrasing discussed in the course notes.

```
OR.hat <- (pi.hat1 * (1 - pi.hat2)) / (pi.hat2 * (1 - pi.hat1))

alpha <- 0.05
var.log.or <- 1/c.table[1,1] + 1/c.table[1,2] + 1/c.table[2,1] + 1/c.table[2,2]
OR.CI <- exp(log(OR.hat) + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(var.log.or))
round(OR.CI, 2)
```

```
## [1] 0.17 0.55
```

The 95 % CI is 0.17 < OR < 0.79. Therefore, with 95 % confidence, the odds of a successful putt are 0.17 to 0.55 times as large when the flagstick is in compared to when the flagstick is removed.

g) (3 points) Should one take the flagstick out or leave it in the hole while putting in this situation? Use your statistical inference results from this problem to explain your answer.

In this situation, the data-driven golfer should remove the flagstick before putting the ball because the the odds of a successful putt are 0.17 to 0.55 times as large when the flagstick is in compared to when the flagstick is removed (at the 95 % confidence level).