Chapter 1

Analyzing a binary response, part 1: introduction

1. (a) The independence and constant probability assumptions may be in question. The main reason is that the infant may learn how to better pick up the Cheerio over the successive trials. Additionally, the Cheerio needs to be placed in the same location on the food tray.

   (b) The independence and constant probability assumptions may be in question. The main reason is that the person may learn how to better break the balls or change strategies (e.g., aim at different locations) to break the balls each time. Additionally, other issues include using the same pool stick, having the same quality of rack each time, ...

   (c) Satisfying the assumptions should not be too difficult. Given they are in the same plot of land, they should generally receive the same amounts of water, weather conditions, and soil composition. There may nonetheless be small deviations within the plot that could allow neighboring seeds to share a local advantage or disadvantage, resulting in a lack of independence. Of course, extrapolating here to other more diverse situations (say, a different climate) may be in question.

   (d) Focusing on one intersection may not provide a valid estimate for the entire town, unless all residents pass through the intersection during the observation period, or the probability of an alternative-fuel car is constant throughout the town (unlikely if there are income differences in different parts of town). Independence may be an issue due to how the cars interact at the intersection or even prior to it.

2. 95% intervals:

Part	Wald	Wilson	Agresti-Coull	Clopper-Pearson
(a)	(0.2320, 0.6680)	(0.2582, 0.6579)	(0.2581, 0.6581)	(0.2306, 0.6847)
(b)	(0.4080, 0.7920)	(0.4074, 0.7660)	(0.4070, 0.7664)	(0.3867, 0.7887)
(c)	(0.6439, 0.8561)	(0.6318, 0.8399)	(0.6309, 0.8408)	(0.6260, 0.8498)
(d)	(0.0567, 0.1673)	(0.0679, 0.1792)	(0.0667, 0.1804)	(0.0626, 0.1808)

   Example interpretation for (c) and the Wilson interval: We would expect 95% of all similarly constructed intervals (repeat the sampling process and use the same type of interval) to contain the true germination rate. Alternatively, with 95% confidence, the true germination rate is between 0.63 and 0.84.

3. (a) An assumption is that the position of the Kisses in the cup will not affect the result. For example, do these Kisses interfere with each other as they are poured out of the cup, which would be a violation of the independence assumption? Also, because there are 10 separate cups, one needs to assume that there are no differences among the cups and how they are poured onto the table.

   (b) 95% intervals:

Wald	Wilson	Agresti-Coull	Clopper-Pearson
(0.2944, 0.4856)	(0.3002, 0.4880)	(0.3001, 0.4881)	(0.2940, 0.4927)

   For example, with 95% confidence, the probability that the kiss will land on its base is between 0.30 and 0.49 (using a Wilson interval).

   (c) All confidence intervals are likely o.k. to use due to the sample size. Plus, notice how close the limits are among the intervals.

   (d) 95% intervals:

Wald	Agresti-Caffo
$(-0.0338, 0.2338)$	$(-0.0347, 0.2307)$

The intervals above are for $\pi_{\text{no almonds}} - \pi_{\text{almonds}}$. Because 0 is within the intervals (in practice, only one interval would be used), there is not sufficient evidence to indicate a difference exists.

(e) Both intervals generally should be o.k. to use due to the sample sizes for each type of kiss. Again, notice how close the limits are among the intervals.

4. (a) The chlamydia status of each woman needs to be independent of all others. For example, groups of women who have common sexual partners are not deliberately included in the sample together. Also, the probability that each women has chlamydia needs to be the same. This could be a problem if some women engage in "risky behavior" while others do not.

(b) Due to the large sample size, all intervals are likely to be similar even with the lower observed prevalence. The safest intervals to use are the Wilson, Agresti-Coull, and the Clopper-Pearson. The 95% Wilson interval is $0.0486 < \pi < 0.0838$. We would expect 95% of all similarly constructed intervals to contain the prevalence. The 95% Clopper-Pearson interval is a little wider with an interval of $0.0476 < \pi < 0.0840$.

5. The 95% Agresti-Min interval is $0.0051 < \pi_{+1} - \pi_{1+} < 0.0294$. McNemar's test gives a value of $M = 8.0476$ with a p-value of 0.0046. At the $\alpha = 0.05$ significance level, there is sufficient evidence to indicate a difference in the positive diagnostic rates. Notice that $\hat{\pi}_{+1} - \hat{\pi}_{1+} > 0$ indicating that the MagNA observes more positive results. Also, notice the confidence interval is entirely above 0 and a 1-sided test of $H_0 : \pi_{+1} - \pi_{1+} \leq 0$ vs. $H_0 : \pi_{+1} - \pi_{1+} > 0$ leads to a p-value of $P(Z > +\sqrt{8.0476}) = 0.0023$.

6. (a) The estimates are close to 1, so they want to make sure the interval is ≤ 1. While the Wilson interval will guarantee this too, this interval's performance for π values very close to 1 can be poor. Also, because humans are being tested here, there may be an interest in wanting to make sure that at least a $(1-\alpha)100\%$ confidence level is achieved.

(b) 95% intervals: (0.9282, 0.9856) for sensitivity and (0.9489, 0.9824) for specificity; we would expect 95% of all similarly constructed intervals to contain the test's true sensitivity and specificity.

(c) 95% intervals:

Disease	Gender	Specimen	Symptoms_Status	Se	Se.low	Se.up	Sp	Sp.low	Sp.up
Chlamydia	Male	Swab	Symptomatic	0.96	0.93	0.99	0.97	0.95	0.98
Chlamydia	Male	Swab	Asymptomatic	0.95	0.87	0.99	0.98	0.96	0.99
Chlamydia	Male	Urine	Symptomatic	0.99	0.96	1.00	0.98	0.97	0.99
Chlamydia	Male	Urine	Asymptomatic	0.96	0.89	0.99	0.99	0.97	1.00
Chlamydia	Female	Swab	Symptomatic	0.92	0.87	0.96	0.97	0.95	0.98
Chlamydia	Female	Swab	Asymptomatic	0.98	0.91	1.00	0.99	0.97	1.00
Chlamydia	Female	Urine	Symptomatic	0.94	0.89	0.97	0.99	0.98	0.99
Chlamydia	Female	Urine	Asymptomatic	0.97	0.89	1.00	0.99	0.98	1.00
Gonorrhea	Male	Swab	Symptomatic	0.99	0.97	1.00	0.99	0.97	1.00
Gonorrhea	Male	Swab	Asymptomatic	1.00	0.78	1.00	0.97	0.94	0.98
Gonorrhea	Male	Urine	Symptomatic	0.98	0.96	0.99	1.00	0.99	1.00
Gonorrhea	Male	Urine	Asymptomatic	1.00	0.75	1.00	0.99	0.98	1.00
Gonorrhea	Female	Swab	Symptomatic	1.00	0.96	1.00	0.98	0.97	0.99
Gonorrhea	Female	Swab	Asymptomatic	0.97	0.84	1.00	1.00	0.99	1.00
Gonorrhea	Female	Urine	Symptomatic	0.93	0.85	0.97	0.99	0.98	1.00
Gonorrhea	Female	Urine	Asymptomatic	0.88	0.71	0.96	0.99	0.98	1.00

7. (a) Because the first box at 10°C had 0 hatch, a Wald interval results is an interval of (0,0). An Agresti-Coull interval results in a lower bound less than 0 at 10°C . The Wilson interval is liberal for a small π (unless the small change mentioned in Section 1.1.3 is implemented). Thus, a Clopper-Pearson interval may be best to use. Below are the Wilson and Clopper-Pearson intervals.

Temperature	Wilson	Clopper-Pearson
10°C	(0.0000, 0.1135)	(0.0000, 0.1157)
15°C	(0.0059, 0.1667)	(0.0008, 0.1722)
20°C	(0.6644, 0.9266)	(0.6528, 0.9436)

(b) Because the interval at 20°C does not overlap with the intervals at the two other temperatures, we can informally conclude that there is a difference in the probability of hatching at these temperatures.

8. (a) In case different boxes cause different hatch probabilities, the replication ensures that we can account for box-to-box variability in our comparison of hatch rates at different temperatures.

(b) Formatted output from `binom.confint()` for 95% Clopper-Pearson intervals:

method	x	n	mean	lower	upper
exact	1	30	0.03	0.00	0.17
exact	2	30	0.07	0.01	0.22
exact	4	30	0.13	0.04	0.31
exact	1	30	0.03	0.00	0.17
exact	0	30	0.00	0.00	0.12
exact	0	30	0.00	0.00	0.12
exact	0	30	0.00	0.00	0.12
exact	12	30	0.40	0.23	0.59
exact	0	30	0.00	0.00	0.12
exact	2	30	0.07	0.01	0.22

(c) The eighth box gave results quite different from the other boxes.

(d) There could be variable conditions within the chamber. For example, the temperature might vary among different positions in the chamber, and this box might have been in a favorable spot. Also, different boxes influence hatch rates as suggested in (a).

(e) The boxes' combined results may be analyzed as a single binomial sample only if the conditions of a binomial distribution are satisfied. In particular, it must be assumed that the boxes have constant hatch probabilities, and that eggs hatching within a box are independent (e.g., first hatches do not promote or interfere with future hatching). Otherwise, it is best to analyze the data as 10 separate binomials with $n = 30$ using techniques described in Chapters 2, 5 (overdispersion), or 6 (clustered data).

9. (a) $0.9632 < \pi < 1.0118$

(b) $0.9325 < \pi < 0.9978$

(c) No. The sample contained one person who does not oppose the tax, so it is impossible for the population probability to be 1. Any confidence interval that includes the value 1 (or above) is erroneous.

(d) The Wilson interval is preferred because its limits can not exceed 1.

10. Equating Z_0 to $Z_{1-\alpha/2}$ gives us

$$\frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = Z_{1-\alpha/2}.$$

Squaring both sides and simplifying leads to

$$n(\hat{\pi} - \pi_0)^2 - \pi_0(1-\pi_0)Z^2_{1-\alpha/2} = 0.$$

Grouping like terms results in

$$\pi_0^2(n + Z^2_{1-\alpha/2}) - \pi_0(2n\hat{\pi} + Z^2_{1-\alpha/2}) + n\hat{\pi}^2 = 0.$$

Using the quadratic formula, we have

$$\begin{aligned}
\pi_0 &= \frac{2n\hat{\pi} + Z^2_{1-\alpha/2} \pm \sqrt{(2n\hat{\pi} + Z^2_{1-\alpha/2})^2 - 4n\hat{\pi}^2(n + Z^2_{1-\alpha/2})}}{2(n + Z^2_{1-\alpha/2})} \\
&= \frac{n\hat{\pi} + Z^2_{1-\alpha/2}/2}{n + Z^2_{1-\alpha/2}} \pm \frac{\sqrt{4n^2\hat{\pi}^2 + 4n\hat{\pi}Z^2_{1-\alpha/2} + Z^4_{1-\alpha/2} - 4n^2\hat{\pi}^2 - 4n\hat{\pi}^2 Z^2_{1-\alpha/2}}}{2(n + Z^2_{1-\alpha/2})} \\
&= \frac{w + Z^2_{1-\alpha/2}/2}{n + Z^2_{1-\alpha/2}} \pm \frac{Z_{1-\alpha/2}}{n + Z^2_{1-\alpha/2}} \frac{\sqrt{n\hat{\pi} + Z^2_{1-\alpha/2}/4 - n\hat{\pi}^2}}{2} \\
&= \tilde{\pi} \pm \frac{Z_{1-\alpha/2}\sqrt{n}}{n + Z^2_{1-\alpha/2}} \sqrt{\hat{\pi} + \frac{Z^2_{1-\alpha/2}}{4n} - \hat{\pi}^2} \\
&= \tilde{\pi} \pm \frac{Z_{1-\alpha/2}\sqrt{n}}{n + Z^2_{1-\alpha/2}} \sqrt{\hat{\pi}(1-\hat{\pi}) + \frac{Z^2_{1-\alpha/2}}{4n}}.
\end{aligned}$$

11. An advantage is the Clopper-Pearson interval is guaranteed to have a true confidence level greater than or equal to $(1-\alpha)100\%$. The other intervals all can have a true confidence level less than $(1-\alpha)100\%$. A disadvantage is the Clopper-Pearson interval is often wider than the other intervals.

12. The plots can be constructed by simply changing the `n` and `alpha` values in the corresponding program that produced the figure for $n = 40$ and $\alpha = 0.05$.

    (a) We leave this up to the reader to choose particular examples to investigate further.

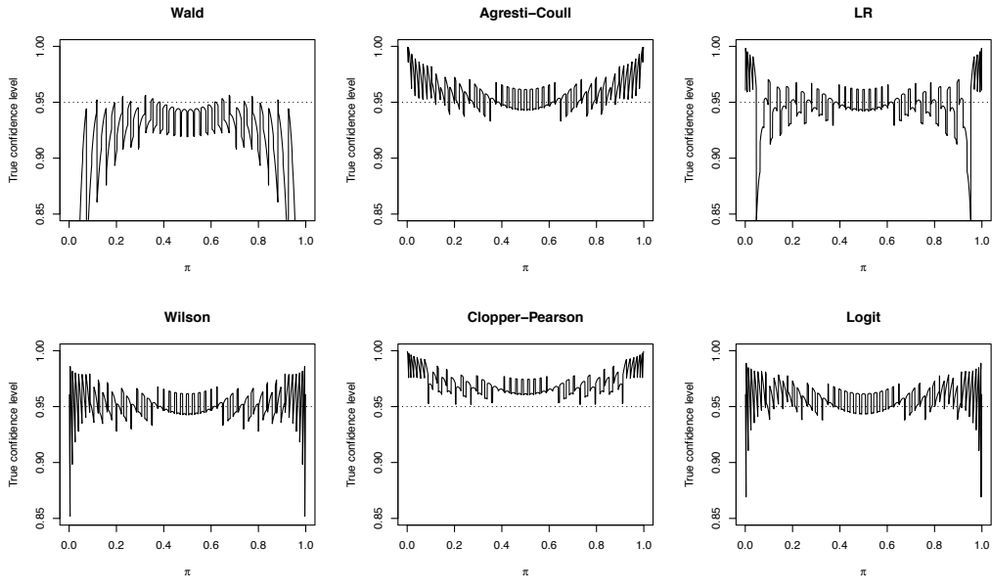
    (b) The confidence levels generally get closer to the true confidence level as n increases. The areas where some confidence intervals have problems achieving the true confidence level (e.g., values of π close to 0 or 1) become narrower in range with respect to π. There are more (less) up-and-down spikes on the plots as n increases (decreases).

    (c) Estimated true confidence levels for some intervals are closer to $(1-\alpha)100\%$ for smaller α (e.g., Clopper-Pearson). The reverse is generally true for larger α. The Wald interval never achieves $(1-\alpha)100\%$ for some smaller levels of α (e.g., $\alpha = 0.01$ with $n = 40$).

    The recommendations given by the paper appear to be appropriate for different levels of α and n than those examined in the text.

13. (a) $0.1456 < \pi < 0.7000$

    (b) Below is the plot which also includes the same intervals as previously along with the logit interval.



    (c) The interval's true confidence level is similar to the Wilson interval from π values of approximately 0.2 to 0.8. There is a large reduction in the true confidence level close to 0 and 1, like the Wilson interval, but this reduction occurs farther away from 0 and 1 than it does for the Wilson interval. The LR interval can be quite conservative, like the Agresti-Coull interval, very close to 0 and 1. Overall, the LR interval is a good interval (much better than the Wald), but the Wilson interval generally is a little better from 0 to 0.2 and 0.8 to 1.

14. (a) The numerical range for $\log(\pi/(1-\pi))$ is $-\infty$ to $+\infty$. The numerical range for $\exp(\cdot)/[1+\exp(\cdot)]$ is 0 to 1, which is ideal because this is the range for a probability.

    (b) Define $g(\pi) = \log(\pi/(1-\pi)) = \log(\pi) - \log(1-\pi)$. Then $g'(\pi) = \pi^{-1} + (1-\pi)^{-1} = [\pi(1-\pi)]^{-1}$ and $g'(\hat{\pi}) = [\pi(1-\pi)]^{-1}\Big|_{\pi=\hat{\pi}} = [\hat{\pi}(1-\hat{\pi})]^{-1}$. Previously, we have seen that $\widehat{Var}(\hat{\pi}) = n^{-1}\hat{\pi}(1-\hat{\pi})$. Therefore, $\widehat{Var}(g(\hat{\pi})) = g'(\hat{\pi})^2\widehat{Var}(\hat{\pi}) = [n\hat{\pi}(1-\hat{\pi})]^{-1}$.

    (c) $0.1583 < \pi < 0.7026$

    (d) The problem is that $\log(\pi/(1-\pi))$ is undefined for $\hat{\pi} = 0$ or 1. For example, when $\hat{\pi} = 0$, we obtain $\log(0)$. The logit interval code in `binom.confint()` starts with the line of `if (any(method == "logit") || all.methods)`. The next line in the function checks if $w = 0$ or $w = n$. If this occurs, the Clopper-Pearson interval is used. There are other potential solutions (such as add 0.5 to the numerator and denominator of $\hat{\pi}$ as suggested by Brown et al. 2001), but this is reasonable.

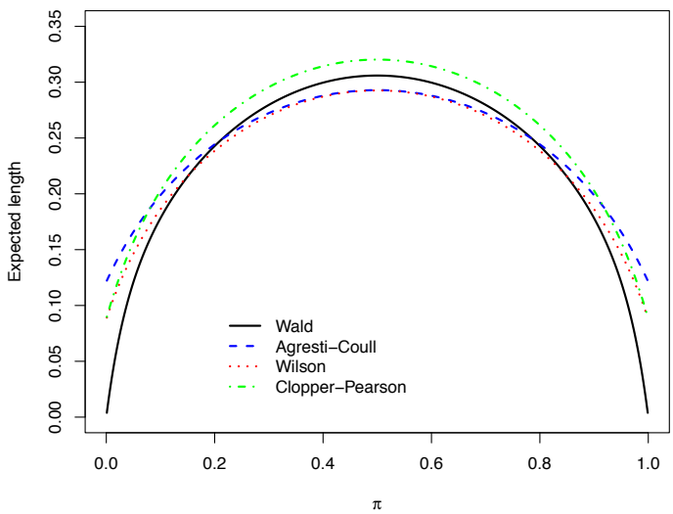
    (e) Please see the plot given in the answer for the LR interval exercise. Note that we used the `binom.confint()` function to calculate the interval.

(f) The interval's true confidence level is similar to the Wilson interval. At times, the interval is generally conservative (e.g., see the range from 0.1 to 0.3).

15. (a) We would like an interval to consist of the smallest range in order to provide a more precise set of possible values for π, while still having $(1-\alpha)100\%$ confidence. The expected length of an interval should not be examined without taking into account the interval's true confidence level. An interval can be really short, but yet be quite liberal. For example, the Wald interval is often the shortest interval, but we have seen that its true confidence level is also often quite far from the stated level.

(b) When finding the true confidence level of a Wald interval in ConfLevel.R, replace the value of `pi` with 0.16 and the value of `save` with `upper - lower` in order to find the expected length.

(c) Below is one plot with all of the expected lengths.



The Clopper-Pearson interval is usually the longest except for at values of π close to 0 and 1. The Wald interval is usually the shortest for values of π close to 0 and 1, which is likely contributing to its liberalness in that location. The Agresti-Coull interval is similar to the Wilson interval, except for values of π close to 0 and 1. Overall, the Wilson interval is usually the shortest or second shortest interval.

(d) Along with the Wilson interval's good coverage properties, we see that its expected length is among the best as well. Thus, the Wilson interval is recommended for use with $n = 40$ and $\alpha = 0.05$.

(e) i. Simulate many data sets (e.g., 1000) using the `rbinom()` function with a specified value for n and π. For example, use $n = 40$ and $\pi = 0.16$ as in part (b).

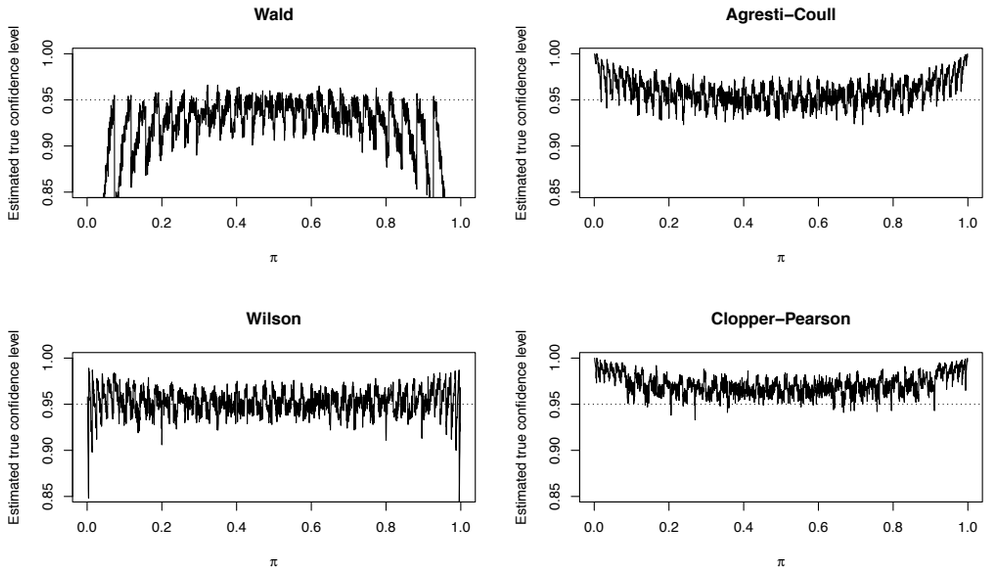
ii. Calculate the $(1-\alpha)100\%$ interval for each sample.

iii. Calculate the length of each interval.

iv. Calculate the mean length across all simulated data sets. This is the estimated expected length of the confidence interval.

16. There are a couple of ways to complete this problem. First, for a specified value of π, the probabilities for $W = 0, \ldots, n$ could be estimated via simulation. These estimated probabilities then can be used in the same manner as the actual probabilities to find the estimated true confidence level. Second, for a specified value of π again, a confidence interval for each simulated sample could be calculated and then checked for whether or not π is in the interval. The proportion of times a π is in the intervals is the estimated true confidence level. Both ways will result in the same estimated confidence level as long as the same samples are chosen.

Below is the plot. The extra jitteriness in comparison to the actual true confidence level plots is due to Monte Carlo simulation variability.



17. (a) 95% intervals for $\pi_{\text{no time-out}} - \pi_{\text{time-out}}$:

Wald	Agresti-Caffo
$(-0.0536, 0.4959)$	$(-0.0558, 0.4765)$

For example, with 95% confidence, the difference in the probability of success for the no time-out strategy vs. the time-out strategy is between −0.0558 and 0.4765 when using the Agresti-Caffo interval.

(b) Two-sided tests:

	Score	Pearson chi-square	LRT
Test statistic	$Z_0 = 1.63$	$X^2 = 2.67$	$-2\log(\Lambda) = 2.61$
P-value	0.1022	0.1022	0.1062

At the $\alpha = 0.05$ level, there is not sufficient evidence to indicate a difference between the probabilities of success for the two strategies.

(c) The estimated relative risk is 1.35, and the 95% confidence interval for the relative risk is (0.90, 2.05). With 95% confidence, the probability of success is between 0.90 and 2.05 times as large when there is not a time-out called prior to the field goal than when there is a time-out called.

(d) The estimated odds ratio is 3.3, and the 95% confidence interval for the odds ratio is (0.76, 14.34). With 95% confidence, the odds of a success is between 0.76 and 14.34 times as large when there is not a time-out called prior to the field goal than when there is a time-out called.

(e) There is at most marginal evidence that icing the kicker is a good strategy. For example, while the 95% confidence interval for the relative risk contains 1, the lower limit is somewhat close to 1. Also, the p-value is close to 0.10. One could also perform a one-sided test of $H_0 : \pi_{\text{no time-out}} - \pi_{\text{time-out}} \leq 0$ vs. $H_a : \pi_{\text{no time-out}} - \pi_{\text{time-out}} > 0$ because the goal here is to determine if the time-out strategy *lowers* the probability of success. In this case, the p-value would be 0.0511 which is very close to rejecting at the $\alpha = 0.05$ level. One-sided confidence intervals could also be constructed as well.

18. (a) 95% intervals for $\pi_{\text{never}} - \pi_{\text{ever}}$:

Wald	Agresti-Caffo
$(-0.5846, -0.1909)$	$(-0.5674, -0.1870)$

For example, with 95% confidence, the difference in the probability of being HIV positive for those never using a condom vs. those using a condom is between −0.0558 and 0.4765 when using the Agresti-Caffo interval.

(b) Two-sided tests:

	Score	Pearson chi-square	LRT
Test statistic	$Z_0 = -4.28$	$X^2 = 18.32$	$-2\log(\Lambda) = 15.48$
P-value	0.00002	0.00002	0.00008

Because the p-value is very small, there is sufficient evidence to indicate a difference between the probabilities of being HIV positive for the two types of condom use.

(c) The estimated odds ratio is 0.19, and the 95% confidence interval for the odds ratio is (0.08 0.44). With 95% confidence, the odds of being HIV positive is between 0.08 and 0.44 times as large when a condom is never used than when a condom is used. Inverting the interval, we obtain (2.29, 12.52). Thus, the odds of being HIV positive are between 2.29 and 12.52 as large when condom is ever used than when it is never used.

(d) The results do not agree with the belief that condom use helps to prevent HIV transmission. The odds of someone having HIV are larger if they have used a condom than those who have not. This is an example of where there may be some additional variable not being taken into account. For example, perhaps prostitutes are far more likely to use condoms than the rest of the population. Because they are in a high-risk profession, prostitutes still may be very susceptible to HIV infection (recall that the question asks whether they have *ever* used condoms). If this is the case, then the fact that condoms were used mainly in the high-risk group might cause there to appear to be a greater proportion of HIV-positives among condom users than among non-users, even if the relative proportions of positives in both risk groups are reduced by condom use. This would then be an example of Simpson's paradox—the marginal association between two variables is different from the association between these two variables conditional on a third variable. Alternatively, perhaps people who already know that they are HIV-positive use condoms to prevent further spread of the disease. It is important to remember that in an observational study, association does not imply causation.

(e) Please see the paper.

19. (a) The sample consists of 16,402 people who are 18-30 years old in the Rayong and Chon Buri provinces in Thailand. Note that the sample given in the contingency table has 7 people removed because they had HIV-1 infection at the time the study began. The population is all 18-30 year old healthy people in the Rayong and Chon Buri provinces in Thailand. While this is the population for this study, the hope is the results for this population would extend to the entire world. Notice that this trial is not focused on people with high risk of HIV infection. This is what the "community-based" part of the trial means.

(b) The estimated odds ratio is 0.69, and the corresponding 95% confidence interval is 0.4805 to 0.9832. The score test p-value is 0.0390 for a two-sided test. The estimated relative risk is 0.69, and the corresponding 95% confidence interval is 0.4831 to 0.9834. The upper bounds on the two confidence intervals are less than 1 (although quite close to 1) and the p-value is less than $\alpha = 0.05$ (although quite close to 0.05). Due to these measures being close to their boundaries of being declared "non-significant," we are hesitant to make an absolute judgment that indeed the treatment type affects HIV status. Rather, we would prefer to say that there is marginal evidence the treatment type affects HIV status. One could also perform a one-sided test or calculate a one-sided confidence interval. For example, the one-sided test leads to a p-value of 0.0195 where the alternative hypothesis is that the vaccine reduces the probability of HIV infection. However, in a different clinical trial (HVTN 502) a few years prior to this one, the reverse effect than what was intended (vaccine led to a higher HIV prevalence) occurred, so two-sided tests and confidence intervals may be better to examine.

(c) One possible report without getting into the statistical details:

> A recently completed AIDS vaccine clinical trial has shown moderate evidence that the vaccine can prevent HIV infection. The vaccine combined two other vaccines—AIDSVAX and ALVAC—that had failed clinical trials in the past. Their combined effect may partially prevent individuals from coming down with HIV. Overall, 8,197 were given the vaccine with 51 still becoming infected with HIV, and 8,198 were given the placebo with 74 becoming infected with HIV. The difference between the vaccine and placebo groups was sufficient to conclude a benefit from the vaccine; however, the observed difference in this trial was small enough that it may still be due to chance. Still, there is a glimmer of hope where there has not been hope for a long time. More studies will need to be performed to determine if the combination of the two AIDS vaccines work to prevent infection.

20. (a) Intent-to-treat: The population is the same as in the previous problem, but with the added assumption that they were all HIV-negative prior to the study.
Per-protocol: The population is the same as in the previous problem, but includes only those individuals who complete all treatments.

(b) Odds ratios and score tests:

	95% confidence interval for the odds ratio	Two-sided score test p-value
Intent-to-treat	(0.52, 1.04)	0.0803
Per-protocol	(0.48, 1.14)	0.1694

The p-values are now greater than $\alpha = 0.05$, and the intervals contain 1. Using a strict $\alpha = 0.05$ level, one would conclude there is not sufficient evidence that the vaccine works. However, with respect to the intent-to-treat data, our "marginal evidence" conclusion from the previous problem would not change.

21. (a) We conjecture that the results would have been less publicized. It may be more interesting to read about a success than something which is somewhat inconclusive.