1. **(15 points) Do exercise 14.9.6 on page 346.**
   Exercise 14.9.6 In a random sample of 746 individuals being treated in Veterans Affairs primary care clinics, 86 were determined to have post-traumatic stress disorder (PTSD) by diagnostic interview [242].
   - **(a) What is a point estimate for p, the proportion of individuals with PTSD among the population being treated in Veterans Affairs primary care clinics?**
   - **(b) Construct and interpret a 95% confidence interval for the population proportion.**
   - **(c) Construct a 99% confidence interval for p. Is this interval longer or shorter than the 95% confidence interval? Explain.**
   - **(d) Suppose that a prior study had reported the prevalence of PTSD among patients seen in primary care clinics in the general population to be 7%. You would like to know whether the proportion of individuals being treated in Veterans Affairs primary care clinics who have PTSD is the same. What are the null and alternative hypotheses of the appropriate test?**
   - **(e) Conduct the test at the 0.01 level of significance, using the normal approximation to the binomial distribution.**
   - **(f) What is the p-value? Interpret this p-value in words.**
   - **(g) Do you reject or fail to reject the null hypothesis? What do you conclude?**
   - **(h) Now conduct the test using the exact binomial method of hypothesis testing. Do you reach the same conclusion?**

Solution: (a) The point estimate for the population proportion of individuals with PTSD among those being treated in Veterans Affairs primary care clinics is simply the sample proportion:

$$\hat{p} = \frac{86}{746} = 0.1152 = 11.5\ \%$$

(b) To construct a 95% confidence interval for the population proportion, we can use the formula:

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where z is the critical value from the standard normal distribution for a 95% confidence interval, which is approximately 1.96 for a large sample size, and n is the sample size.

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.1152 \pm 1.96\sqrt{\frac{0.1152(1-0.1152)}{746}}$$

$$0.0848 < p < 0.1456$$

We can interpret this as: we are 95% confident that the true proportion of individuals with PTSD among those being treated in Veterans Affairs primary care clinics is between 8.48% and 14.56%.

(c) To construct a 99% confidence interval, we use the same formula but with a different critical value, approximately 2.58 for a 99% confidence interval:

$$\hat{p} \pm 2.58\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.1152 \pm 2.58\sqrt{\frac{0.1152(1-0.1152)}{746}}$$

$$0.0751 < p < 0.1553$$

This interval is longer than the 95% confidence interval because we are more confident that the true proportion of individuals with PTSD is contained within it.

(d) The null hypothesis is that the proportion of individuals being treated in Veterans Affairs primary care clinics who have PTSD is equal to 7%, while the alternative hypothesis is that it is not equal to 7%.

$$H_0: p = 0.07$$

$$H_{\mathrm{a}}: p \neq 0.07$$

(e) To conduct the test at the 0.01 level of significance, we can use the normal approximation to the binomial distribution. We first calculate the test statistic:

$$z = \frac{(\hat{p} - p_0)}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

$$z = \frac{(0.1152 - 0.07)}{\sqrt{\frac{0.07(1-0.07)}{746}}}$$

$$z = 4.84$$

where p_0 is the hypothesized proportion under the null hypothesis. The critical value for a two-tailed test at the 0.01 level of significance is approximately ±2.58, so we reject the null hypothesis if |z| > 2.58. In this case, |z| = 4.84, which is greater than 2.58, so we reject the null hypothesis.

(f) The *p-value* is the probability of obtaining a test statistic as extreme or more extreme than the observed one, assuming the null hypothesis is true. In this case, the p-value is the probability of getting a z-score of 4.84 or more extreme in a standard normal distribution:

$$p - value = 2 \times P(Z > 4.84) < 0.0001$$

This means that if the null hypothesis were true, the probability of observing a sample proportion as extreme or more extreme than the observed one is less than 0.01%.

(g) We reject the null hypothesis because the *p-value* is less than the significance level of 0.01. We can conclude that the proportion of individuals with PTSD among those being treated in Veterans Affairs primary care clinics is significantly different from 7%.

(h) To conduct the exact binomial method of hypothesis testing, we need to calculate the probability of observing 86 or more individuals with PTSD in a sample of 746 individuals, assuming the null hypothesis is true.

This is the probability of getting 86 or more successes in 746 trials with a success probability of 0.07:

$$p - value \ = \ P(X \ >= \ 86)$$

$$= \ 1 \ - \ P(X \ <= \ 85)$$

$$= \ 1 \ - \ pbinom(85, size \ = \ 746, prob \ = \ 0.07)$$

$$< \ 0.0001$$

The $p - value$ is again less than the significance level of 0.01, so we reject the null hypothesis and conclude that the proportion of individuals with PTSD among those being treated in Veterans Affairs primary care clinics is significantly different from 7%. This conclusion is the same as in part (g), but the exact binomial method gives a more precise $p - value$ than the normal approximation.

2. **(10 points) Do exercises 14.9.9 on page 346.**
   Exercise 14.9.9 You plan to conduct a study to compare the prevalence of electronic cigarette use in the past 30 days by high school students in grade 10 to the prevalence in grade 12. Based on a previous study, you assume that the prevalence of e-cigarette use among students in grade 12 is 25%. You intend to conduct a two-sided, one-sample test of proportions at the 0.05 level of significance.
   a) **If the prevalence of e-cigarette use among 10th graders is postulated to be 20%, what sample size would be required to have 80% power?**
   b) **If the prevalence of e-cigarette use among 10th graders is postulated to be 15%, what sample size would be required to have 80% power?**
   c) **If the prevalence of e-cigarette use among 10th graders is postulated to be 15% but you plan to perform your test at the 0.01 level of significance, what sample size would be required to have 80% power?**

Solution: (a) Assuming a two-sided, one-sample test of proportions at 0.05 level of significance with a 80% power to detect a difference in e-cigarette use prevalence between students in the 10th and 12th grades, with presumptive prevalence rates of 20% and 25%, respectively, the required number of samples would be:

$$n \ = \ \left(z_{\frac{\alpha}{2}} \ + \ z_{\beta}\right)^2 \ \times \ \frac{p(1 \ - \ p)}{d^2}$$

where *p* the postulated prevalence rates of e-cigarette use in 10th grade students, and $z_{\frac{\alpha}{2}}$ is the critical value of the standard normal distribution at the 0.025 level of significance, which is 1.96, and z_{β} is the critical value of the standard normal distribution at 80% power, which is 0.84 and d is the margin error, which is the difference between the postulated proportion and the true proportion that we want to detect.

When we substitute the values, we obtain:

$$n \ = \ (1.96 \ + \ 0.84)^2 \ \times \ \frac{0.20(0.80)}{(0.20 \ - \ 0.25)^2} \ = \ 501.76$$

Hence, a total sample size of 502 students would be required.

(b) Assuming the same conditions as in (a) but with postulated prevalence rates of 15% for 10th graders, respectively, the required sample size would be:

$$n = (1.96 + 0.84)^2 \times \frac{0.15(0.85)}{(0.15 - 0.25)^2} = 99.96$$

Hence, a total sample size of 100 students would be required.

(c) Assuming the same conditions as in (b) but with 0.01 level of significance, the required sample size would be:

$$n = (2.576 + 0.84)^2 \times \frac{0.15(0.85)}{(0.15 - 0.25)^2} = 148.78$$

Hence, a total sample size of 149 students would be required.

3. **(10 points) Do exercises 14.9.10 on page 347.**
   Exercise 14.9.10 You plan to conduct a study to compare the prevalence of electronic cigarette use in the past 30 days by high school students in grade 10 to the prevalence in grade 12. You do not have any information about prevalence in either group. You intend to conduct a two-sided, two-sample test of proportions at the 0.05 level of significance and wish to have 90% power for your test.
   - **(a) If the prevalence of e-cigarette use is postulated to be 20% for 10th graders and 25% for 12th graders, what sample size would be required? Assume that you want to enroll equal numbers of 10th and 12th graders.**
   - **(b) If the prevalence of e-cigarette use is postulated to be 15% for 10th graders and 25% for 12th graders, what sample size would be required?**
   - **(c) If the prevalence of e-cigarette use is postulated to be 15% for 10th graders and 25% for 12th graders, and you are satisfied with 80% power rather than 90%, what sample size would be required?**

Solution: (a) Assuming a two-sided, two-sample test of proportions at 0.05 level of significance with a 90% power to detect a difference in e-cigarette use prevalence between students in the 10th and 12th grades, with presumptive prevalence rates of 20% and 25%, respectively, the required number of samples would be:

$$n = 2 \times \left(z_{\frac{\alpha}{2}} + z_{\beta}\right)^2 \times \frac{p_1(1 - p_1) + p_2(1 - p_2)}{(p_1 - p_2)^2}$$

where p_1 and p_2 are the postulated prevalence rates of e-cigarette use in 10th and 12th grade students, respectively, and $z_{\frac{\alpha}{2}}$ is the critical value of the standard normal distribution at the 0.025 level of significance, which is 1.96, and z_{β} is the critical value of the standard normal distribution at 90% power, which is 1.28.

When we substitute the values, we obtain:

$$n = (1.96 + 1.28)^2 \times \frac{0.20(0.80) + 0.25(0.75)}{(0.20 - 0.25)^2} = 1460 \, per \, group$$

Hence, a total sample size of 2920 students (1460 per group) would be required.

(b) Assuming the same conditions as in (a) but with postulated prevalence rates of 15% and 25% for 10th and 12th graders, respectively, the required sample size would be:

$$n = (1.96 + 1.28)^2 \times \frac{0.15(0.85) + 0.25(0.75)}{(0.15 - 0.25)^2} = 332 \, per\ group$$

Hence, a total sample size of 664 students (332 per group) would be required.

(c) Assuming the same conditions as in (b) but with 80% power instead of 90%, the required sample size would be:

$$n = (1.96 + 0.84)^2 \times \frac{0.15(0.85) + 0.25(0.75)}{(0.15 - 0.25)^2} = 248 \, per\ group$$

Hence, a total sample size of 496 students (248 per group) would be required.

4. **(10 points) Do exercise 14.9.13 on page 348.**
   Exercise 14.9.13 Suppose you are interested in investigating factors that affect the prevalence of tuberculosis among intravenous drug users. In a group of 97 individuals who admit to sharing needles, 24.7% had positive tuberculin skin test results; among 161 drug users who deny sharing needles, 17.4% had positive test results [246].
   - **(a) Assuming that the population proportions of positive skin test results are in fact equal, estimate their common value p.**
   - **(b) Test the null hypothesis that the proportions of intravenous drug users who have positive tuberculin skin test results are identical for those who share needles and those who do not.**
   - **(c) What is the probability distribution of the test statistic?**
   - **(d) What is the p-value? What do you conclude?**
   - **(e) Construct a 95% confidence interval for the true difference in proportions.**

Solution: (a) Assuming that the population proportions of positive skin test results are in fact equal, the common value p can be estimated using the following formula:

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Where p_1 and p_2 are proportion of individuals who admit to sharing needles has positive tuberculosis skin test result among 97 individuals and drug users who denied sharing needles has positive test result among the 161 drug users, respectively. n_1 and n_2 are their sample sizes respectively.

When we substitute the values, we obtain:

$$p = \frac{97 \times 0.247 + 161 \times 0.174}{97 + 161} = 0.2015$$

Hence, the estimate of their common value p is 0.2015.

(b) The null hypothesis is that the proportion of intravenous drug users who have positive skin test results are identical for those shared the needles and those who don’t share needle.

$$H_0 : p_1 = p_2$$

$$H_a : p_1 \neq p_2$$

Under H_0 , the test statistic is

$$z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Where $q = 1 - p$

$$z = \frac{0.247 - 0.174}{\sqrt{0.2015 \times 0.7985\left(\frac{1}{97} + \frac{1}{161}\right)}} = 1.4147$$

Using a standard normal distribution table, the critical value for a two-sided test at the 0.05 level of significance is ±1.96. Since 1.4147 < 1.96, we fail to reject the null hypothesis.

(c) The probability distribution of the test statistic is approximately normal, since the sample sizes are large enough and the proportion of positive test results is not too close to 0 or 1.

(d) The *p-value* is 0.157. The *p-value* of the two-sided test is greater than 0.05. The above test failed to reject the null hypothesis. Hence, it may be concluded that the proportions are identical in the two populations.

(e) To construct 95% confidence intervals for the true difference in proportions, we can use the formula:

$$\text{C. I.} = (p_1 - p_2) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

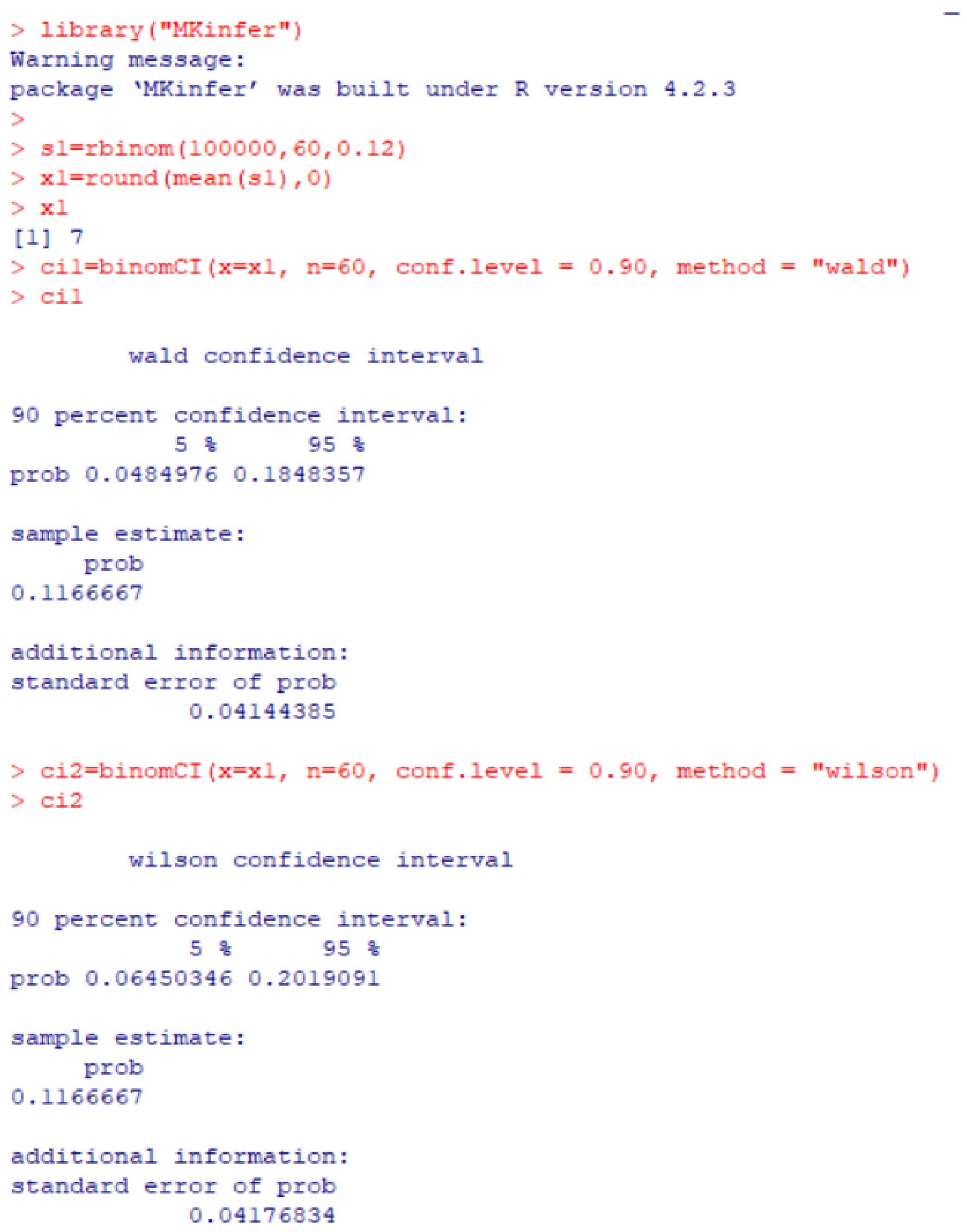
$$z_{\frac{\alpha}{2}} = 1.96 \text{ at } 95\% \text{ confidence interval}$$

$$\therefore -0.0307 < p < 0.1767$$

Hence, the 95% confidence interval is $(-0.0307, 0.1767)$. This interval contains the true population proportion with 95% confidence.

5. **(10 points) Run a Monte Carlo simulation to check the coverage of the Wald interval and the Wilson interval mentioned in lecture (or see equations (1) and (4) in the reference paper at link http://projecteuclid.org/download/pdf_1/euclid.ss/1009213286, a pdf file is posted on Canvas also). Simulate the 90% confidence intervals for n=60 and p=0.12. Do 100,000 simulation runs to calculate the empirical coverages. Do either of the intervals have correct coverage probability? Submit your R commands and your numerical estimates of the two coverage probabilities, with your answer to above question.**

Solution: R Code



```
> library("MKinfer")
Warning message:
package 'MKinfer' was built under R version 4.2.3
>
> s1=rbinom(100000,60,0.12)
> x1=round(mean(s1),0)
> x1
[1] 7
> ci1=binomCI(x=x1, n=60, conf.level = 0.90, method = "wald")
> ci1

        wald confidence interval

90 percent confidence interval:
           5 %       95 %
prob 0.0484976 0.1848357

sample estimate:
     prob
0.1166667

additional information:
standard error of prob
            0.04144385

> ci2=binomCI(x=x1, n=60, conf.level = 0.90, method = "wilson")
> ci2

        wilson confidence interval

90 percent confidence interval:
            5 %       95 %
prob 0.06450346 0.2019091

sample estimate:
     prob
0.1166667

additional information:
standard error of prob
            0.04176834
```

From the above output, both Wald and Wilson intervals cover $p = 0.12$ with error levels 0.04144385 and 0.04176834, respectively. Since the error level of Wald interval is less than that of Wilson, then Wald interval is even better than Wilson interval to cover $p = 0.12$.