**Chapter 1 practice problems**

The answers given here are sometimes only partial answers. Please see the answer keys for projects and tests for examples of full answers.

Note: Many of the practice problems for this course are based on exercises given in Alan Agresti’s “Introduction to Categorical Data Analysis” book.

1. Below is how you can prove E(W) = nπ. Note that you are NOT responsible for this proof other than perhaps as an extra credit problem.



= 

= 

=  since w = 0 does not contribute to the sum

= 

=  where x=w-1

=  because a binomial distribution with n-1 trials is inside the sum.

1. When the General Social Survey asked subjects whether they would be willing to accept cuts in their standard of living to protect the environment, 344 of 1170 subjects said “yes”
   1. Estimate the population proportion who would say “yes” by finding 

> w <- 344

> n <- 1170

> pi.hat <- w/n

> pi.hat

[1] 0.2940171

* 1. Conduct a hypothesis test to determine whether a majority or minority of the population would say “yes”.

Using a score test for H0:π = 0.5 vs. Ha:π ≠ 0.5

> prop.test(x = w, n = n, p = 0.5, conf.level = 0.95, correct = FALSE)

1-sample proportions test without continuity correction

data: w out of n, null probability 0.5

X-squared = 198.5675, df = 1, p-value < 2.2e-16

alternative hypothesis: true p is not equal to 0.5

95 percent confidence interval:

0.2686193 0.3207630

sample estimates:

p

0.2940171

Using a LRT for H0:π = 0.5 vs. Ha:π ≠ 0.5

> #Sometimes, computations work better with finding log(L) instead of L

> log.L.Ho <- w\*log(0.5)+(n-w)\*log(1-0.5) #log(L) under Ho

> log.L.Ho.Ha <- w\*log(pi.hat)+(n-w)\*log(1-pi.hat) #log(L) under Ho or Ha

> test.stat <- -2\*(log.L.Ho-log.L.Ho.Ha)

> data.frame(log.L.Ho, log.L.Ho.Ha, stat = test.stat)

log.L.Ho log.L.Ho.Ha stat

1 -810.9822 -708.68 204.6043

> 1-pchisq(q = test.stat, df = 1) #p-value

[1] 0

Because the p-value is small, there is sufficient evidence to indicate that π > 0.5 or π < 0.5.

* 1. Find the 99% Wald, Agresti-Coull, Wilson, and Clopper-Pearson intervals. Why do you think the intervals are similar?

> alpha < -0.01

> library(package = binom)

> binom.confint(x = w, n = n, conf.level = 1-alpha, methods = "all")

method x n mean lower upper

1 agresti-coull 344 1170 0.2940171 0.2609272 0.3294300

2 asymptotic 344 1170 0.2940171 0.2597081 0.3283261

3 bayes 344 1170 0.2941930 0.2603652 0.3288722

4 cloglog 344 1170 0.2940171 0.2601360 0.3286321

5 exact 344 1170 0.2940171 0.2601828 0.3295516

6 logit 344 1170 0.2940171 0.2609106 0.3294518

7 probit 344 1170 0.2940171 0.2606715 0.3292094

8 profile 344 1170 0.2940171 0.2605299 0.3290594

9 lrt 344 1170 0.2940171 0.2605403 0.3290754

10 prop.test 344 1170 0.2940171 0.2682044 0.3212000

11 wilson 344 1170 0.2940171 0.2609468 0.3294104

In summary, the intervals are:

|  |  |  |
| --- | --- | --- |
| **Name** | **Lower** | **Upper** |
| Wald | 0.2597 | 0.3283 |
| Agresti and Coull | 0.2609 | 0.3294 |
| Wilson | 0.2609 | 0.3294 |
| Clopper-Pearson | 0.2602 | 0.3296 |

1. Using calculus, it is often easier to derive the maximum of the log of the likelihood function than the maximum of the likelihood function itself. Both functions have the maximum at the same value, so it is sufficient to do either. Calculate the log likelihood function for the binomial distribution and find the maximum likelihood estimation using calculus methods. Students are NOT responsible for this proof other than perhaps as an extra credit problem.

This is like what I did in a video, but now start with a binomial. Note that there is just one observation here.



where W is the number of successes. The MLE can be found from



Then



Setting this equal to 0 and solving for π produces

 ⇒  ⇒  ⇒  ⇒ 

Thus, the maximum likelihood estimator of π is .

1. Derive the limits for the Wilson confidence interval. Students are NOT responsible for this proof other than perhaps as an extra credit problem.

The limits of the Wilson confidence interval come from “inverting” the score test for π. This means finding the set of π0 such that

 is satisfied.

Working with an equality produces,

.

Then



Using the quadratic formula produces



Thus the limits of Wilson interval are .

1. For a past project in STAT 875, I had my students investigate the “logit confidence interval” for the probability of success. See p. 114-5 of Brown, Cai, and DasGupta (2001) for more information about it. The interval is found as



where I use ea to represent exp(a). Complete the following.

* 1. Find the confidence interval for when w = 4 and n = 10.

> w < -4

> n <- 10

> alpha <- 0.05

> pi.hat <- w/n

> #LOGIT interval

> num <- exp(log(pi.hat /(1-pi.hat)) + qnorm(p = c(alpha/2, 1-alpha/2)) \*

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

> num/(1 + num)

[1] 0.1583420 0.7025951

> binom.confint(x = 4, n = 10, conf.level = 0.95, methods = "all")

method x n mean lower upper

1 agresti-coull 4 10 0.4000000 0.16711063 0.6883959

2 asymptotic 4 10 0.4000000 0.09636369 0.7036363

3 bayes 4 10 0.4090909 0.14256735 0.6838697

4 cloglog 4 10 0.4000000 0.12269317 0.6702046

5 exact 4 10 0.4000000 0.12155226 0.7376219

6 logit 4 10 0.4000000 0.15834201 0.7025951

7 probit 4 10 0.4000000 0.14933907 0.7028372

8 profile 4 10 0.4000000 0.14570633 0.6999845

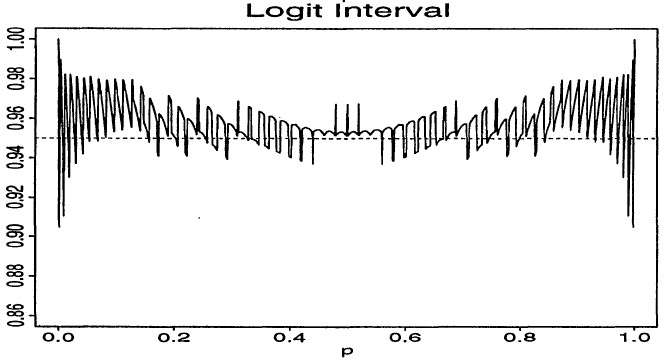
9 lrt 4 10 0.4000000 0.14564246 0.7000216

10 prop.test 4 10 0.4000000 0.13693056 0.7263303

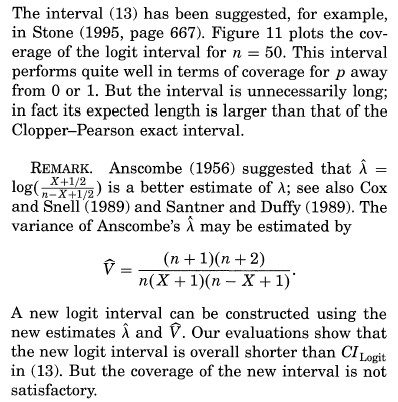
11 wilson 4 10 0.4000000 0.16818033 0.6873262

The interval is (0.1583, 0.7026).

* 1. Below is Figure 11 from Brown, Cai, and DasGupta (2001) demonstrating the true confidence level for n = 50 and α = 0.05. Comment on how well the confidence interval performs with respect to its stated confidence level and the intervals examined in class.



Excerpt from Brown et al. (2001):





The logit interval is mostly conservative for 0.05 < π < 0.2 and 0.8 < π < 0.95 and has a true confidence level close to 0.95 for middle values of π. The Wilson interval is generally closer to the 0.95 level for these values of π. The logit interval has very erratic coverage for values of π very close to 0 or 1 - it can be close to 1 or close to 0.9. This is somewhat similar to the erratic coverage found with the Wilson interval. The Agresti-Coull interval would be a more conservative choice for these values of π.

* 1. Create the plot in b) using R.

This can be done by using code from ConfLevel4Intervals.R as a template. Replace one of the intervals calculated in it with the logit interval.

1. Below are results from a study on myocardial infraction and whether or not an individual took aspirin over a particular time period:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Myocardial infraction | |
|  |  | Yes | No |
| Group | Placebo | 189 | 10,845 |
| Aspirin | 104 | 10,933 |

* 1. Find and interpret the estimated relative risk and corresponding confidence interval. Note that the interval is (1.43, 2.31) or equivalently (0.43, 0.70) for an inverted RR.
  2. Find and interpret the estimated odds ratio and corresponding confidence interval.
  3. Discuss the differences in the interpretations. Why are the computed values similar?

1. Below are results from a study comparing radiation therapy with surgery in treating cancer of the larynx.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Cancer | |
|  |  | Controlled | Not controlled |
| Treatment | Surgery | 21 | 2 |
| Radiation therapy | 15 | 3 |

Let π1 be the probability that cancer was controlled for the surgery treatment, and let π2 be the probability that cancer was controlled for the radiation therapy treatment.

* 1. Perform a Pearson chi-square, LR, and score test to determine if there is a difference between the treatments.

> c.table <- array(c(21, 15, 2, 3), dim = c(2, 2), dimnames = list(Trt =

c("Surgery", "Radiation"), Cancer = c("Controlled", "Not Controlled")))

> c.table

Cancer

Trt Controlled Not Controlled

Surgery 21 2

Radiation 15 3

> library(package = vcd)

> assocstats(x = c.table)

X^2 df P(> X^2)

Likelihood Ratio 0.59476 1 0.44058

Pearson 0.59915 1 0.43890

Phi-Coefficient : 0.121

Contingency Coeff.: 0.12

Cramer's V : 0.121

* 1. Should we be concerned about if the sample sizes are large enough for the large sample approximation to hold? Note that we will explore this more in Chapters 3 and 6.

1. Another way to write the Pearson chi-square statistic is



Answer the following questions:

* 1. Show the second equality holds.
  2. Notice the role of nj in this second expression. Suppose as nj increases,  and  change very little. What will happen to X2?
  3. Relate your answer to b) to what will happen to the corresponding hypothesis test result. Discuss why this is a problem with hypothesis testing.

1. Show that π1 = π2 is equivalent to OR = 1 and RR = 1.
2. What happens to the Pearson chi-square statistic and the -2log(Λ) statistic if the rows were interchanged? Do the values the stay the same or change?

1. Suppose the odds ratio between treatment (A, B) and response (death, survival) equals 2.0 with the following contingency table set-up.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Response | |
|  |  | Death | Survival |
| Treatment | A |  |  |
| B |  |  |

* 1. Explain what is wrong with the interpretation, “The probability of death with treatment A is twice as large as this probability for treatment B.” Similarly, explain what is wrong with the interpretation, “death is twice as likely when someone receives treatment A instead of treatment B.” Give the correct interpretation.
  2. When is the quoted interpretation in (a) correct?