STATISTICS 475: Applied Discrete Data Analysis

MIDTERM 1

Feb 28, 2022

SOLUTIONS

INSTRUCTIONS: **DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO!**

- Answer all parts of all problems in the space provided.
- Place final answers in the blanks when they are provided.
- **Show work!** If you don't have a calculator, it is enough to write out the exact equation with the right numbers so that I could calculate the result myself.
- If you need more space (I don't think you should), indicate clearly where the problem continues.
- Use $\alpha = 0.05$ unless told otherwise

PROBLEM 1 (3 pts)
A professor gives a true-false final exam with 10 questions on it. A student who has paid no attention to the course has to guess randomly at each question. They need to get at least 8 questions right to pass the course.

1. **(2 pts)**What is the probability that they pass? (If you don't have a calculator, it is enough to write out the exact equation with the right numbers that I could calculate myself.)

   Number correct=$W \sim Bin(10, 0.5)$. Find $P(W \geq 8) = P(8) + P(9) + P(10)$

   $P(w) = (n!/w!(n-w)!)\pi^w(1-\pi)^n = 10!/(w!(10-w)!)0.5^w(1-0.5)^{10-w}$
   $P(8) = (10 * 9/2) * .5^{10} = 0.0439$
   $P(9) = (10/1) * .5^{10} = 0.0098$
   $P(10) = (1) * .5^{10} = 0.0010$
   $0.0439 + 0.0098 + 0.0010 = 0.0547$
   P(Pass)=0.0547

2. **(1 pt)** In order to solve this problem we have to assume "independent trials". What does this assumption mean in terms of the student taking the final exam?

   Independent trials means here that getting one question right or wrong does not influence the chance that another question is right or wrong.

BONUS: (1 pt) What is the name of the NFL placekicker who had the record for longest field goal for many years, despite having no toes on his foot? (Circle one)

1. Garo Yepremian
2. Justin Tucker
3. **Tom Dempsey**
4. Sebastian Janikowski
5. Jason Elam

PROBLEM 2 (5 pts)
See the Covid Data Description provided before the test. The description is repeated below for convenience.

(All data from the BCCDC Covid Surveillance Dashboard)
From January 11 through February 10, 1167 "Seniors" (people aged 70+) were hospitalized with Covid-19, of whom 891 were "fully vaccinated" (2 or more shots). Anti-vaccine commentators like to use the fact that over half of hospitalizations are among vaccinated people as "proof" that the vaccines "don't work". In BC, roughly 670,000 out of 720,000 people ages 70 and over are fully vaccinated.

1. **(1 pt)** *Among seniors in the hospital during this period*, estimate the probability that the senior is vaccinated.
   1167 are hospitalized; 891 of those are seniors. $891/1167 = 0.76350$

   Answer: 0.76

2. **(1 pt)** *Among vaccinated seniors*, estimate the odds that the senior is hospitalized during this period.
   674532 seniors are vaccinated, 891 are hospitalized. Estimate odds as $w/(n-w)$
   $891/(670000 - 891) = 0.001332$
   Answer: 0.0013

3. **(1 pt)** *Among unvaccinated seniors*, estimate the odds that the senior is hospitalized during this period.
   $720000 - 670000 = 50000$ seniors are unvaccinated. 1167-891=276 were hospitalized
   $276/(50000 - 276) = 0.00555$
   Answer: 0.0056

4. **(1 pt)** Report the estimated odds ratio in a way that compares the risk of hospitalization for an unvaccinated senior against the same risk for a vaccinated senior, treating vaccination as the "standard".
   Odds of hosp for unvax / odds of hosp for vax
   $0.00552/0.001332 = 4.14$
   Answer: 4.14

5. **(1 pt)** Regardless of your previous answers, suppose that a 95% confidence interval for this odds ratio was from 2 to 3. What would that tell you about the relationship between vaccination and the risk of hospitalization for seniors. Specifically, does there appear to be any benefit to the vaccines?
   Since 1 is not included in this interval, then it is clear that unvaccinated seniors have higher odds of hospitalization than vaccinated seniors. (The vaccine appears to provide protection against hospitalization.)

PROBLEM 3 (11 pts)
Refer to the **Low Birth Weight** problem that was described before this test, and the code provided for this analysis.

1. (1 pt) In the context of the problem what is the definition of "success" that is used in all of the logistic regression models that we fit in our R program?

   "Success" is if a baby is born with low birthweight (less than 2500g)

2. (2 pts) In my description of the problem, I thought that `smoke`, `race`, and `ht` would be important variables. I nonetheless started by fitting a model with all variables in it. Would it be reasonable to eliminate all variables other than these three? Report and interpret results of a test that can answer this question.
   See the LR test comparing `mod.fit2` to `mod.fit1`
   Test Statistic: 15.045

   p-value: 0.01017

   Conclusion: The additional variables contribute significantly to the logistic regression model (they should not all be dropped).

3. (2 pts) Do the effects of smoking on low birth weight differ depending on the race of the mother? Cite evidence from the output and answer this question.

   `mod.fit3` has the `smoke:race` interaction added. The LR test for this effect is not significant ($-2\log(\Lambda) = 2.59$,, p-value=0.274), so the evidence does not support a different smoking effect for different races.

4. (3 pt) Regardless of the results so far, consider a three-factor model with no interactions. In the table below, write out the indicator variables and their values for each combination of `smoke`, `race`, and `ht`. There are more columns than you should need. Use variable names that clearly distinguish between the factors (e.g., s_1,$s_2, \ldots$ for `smoke`, $r_1, r_2, \ldots$ for `race`, and $h_1, h_2, \ldots$ for `ht`). **Include ALL indicators, and not just the ones that R uses.**

			Indicators											
`smoke`	`race`	`ht`	s_1	s_2		r_1	r_2	r_3		h_1	h_2			
0	1	0	1	0		1	0	0		1	0			
0	1	1	1	0		1	0	0		0	1			
0	2	0	1	0		0	1	0		1	0			
0	2	1	1	0		0	1	0		0	1			
0	3	0	1	0		0	0	1		1	0			
0	3	1	1	0		0	0	1		0	1			
1	1	0	0	1		1	0	0		1	0			
1	1	1	0	1		1	0	0		0	1			
1	2	0	0	1		0	1	0		1	0			
1	2	1	0	1		0	1	0		0	1			
1	3	0	0	1		0	0	1		1	0			
1	3	1	0	1		0	0	1		0	1			

5. (2 pt) Using the indicators you laid out above, write out the mathematical model that R used for the model fit in `mod.fit3`? Use symbols for parameters that are clear and easy to understand (e.g., β^S for smoke parameters, and so forth).

   $\text{logit}(\pi) = \beta_0 + \beta_2^S s_2 + \beta_2^R r_2 + \beta_3^R r_3 + \beta_2^H h_2 + \beta_{22}^{SR} s_2 r_2 + \beta_{23}^{SR} s_2 r_3$

6. (1 pts) Consider `mod.fit2`. If I wanted to find a likelihood-ratio confidence interval for the odds ratio relating to the effect of hypertension vs no hypertension on low birth weight, what coefficients would I need to use in `mcprofile()`?
   $(\beta_0 + \beta_2^S s_2 + \beta_2^R r_2 + \beta_3^R r_3 + \beta_2^H 1)$
   $-(\beta_0 + \beta_2^S s_2 + \beta_2^R r_2 + \beta_3^R r_3 + \beta_2^H 0)$
   $a_1 =$ 0 , $a_2 =$ 0 , $a_3 =$ 0 , $a_4 =$ 0 , $a_5 =$ 1

7. (1 pt Bonus) Consider `mod.fit2`. If I wanted to find a likelihood-ratio confidence interval for the *probability of low birth weight* for a nonsmoking white mother with no hypertension, what coefficients would I need to use in `mcprofile()`?
   $\beta_0 + \beta_2^S 0 + \beta_2^R 0 + \beta_3^R 0 + \beta_2^H 0$

   $a_1 =$ 1 , $a_2 =$ 0 , $a_3 =$ 0 , $a_4 =$ 0 , $a_5 =$ 0

PROBLEM 4 (7 pts)
Two batches of insects are sprayed with different concentrations of a new pesticide. At a concentration of 1.1%, 3 out of 63 survive. At a concentration of 0.65%, 12 out of 72 survive.

1. (3 pts) Name the method you would use for these situations:

   (a) Suppose you wanted to quickly calculate a confidence interval for the probability of insect survival at 1.1% concentration without a computer.

   Agresti-Coull

   (b) Suppose you had a computer and wanted to calculate a confidence interval for the probability of insect survival at 1.1% concentration using the method that is generally most accurate without being too conservative.

   (Wilson) Score

   (c) Suppose you wanted to calculate a confidence interval for the probability of insect survival at 1.1% concentration. What method would be most likely to give you an endpoint below 0?

   Wald ("asymptotic" in `binom.confint`)

2. (2 pt) **Refer to the Pesticide script given to you before the test.** Perform the score test of the null hypothesis that the survival rate of insects is the same at both concentrations against the alternative that the lower concentration has a higher survival rate.

   Test Statistic: 4.842 or $\sqrt{4.842} = 2.20$

   p-value: 0.014

   Conclusion: Reject H_0 and conclude that the lower concentration does have a higher survival rate.

3. (2 pt) Estimate the percentage by which the odds of survival decrease when the higher concentration is used compared to the lower concentration, and what is the most appropriate confidence interval for this parameter?

   Estimated Percentage: 75% ($\widehat{OR} = 0.25$)

   Confidence Interval: 12% to 93% (CI for OR is 0.072 to 0.878)

PROBLEM 5 (4 pts)
We ask 50 right-handed people to catch an object thrown to them. One time they are to use their right hand and the other time they use their left. The order is randomized for each person. We record whether they catch the object each time. We speculate that right-handed people will be better at catching the object with their right hand than their left.

The data look like this

		Left Hand		
		Catch	No Catch	Total
Right Hand	Catch	25	11	36
	No Catch	3	11	14
	Total	28	22	50

1. (2 pts) Calculate a test statistic that would allow us to see whether these data support our initial beliefs.

   This is a paired data test, so we can use McNemar or the score Z on the two off-diagonal counts, 11 and 3
   McNemar: $(11-3)^2/(11+3) = 64/14 = 4.57$
   Score $Z_0 = (11/14 - 0.5)/\sqrt{0.5 * 0.5/14} = 2.13$
   Test Statistic: Either 4.57 or 2.13 (A Wald test would be inferior, Z_w =2.61)

2. (2 pts) Explain how you would use the statistic to compute a p-value and draw a conclusion (I don't need R code, and you don't need to do the calculation, but you can tell me what you would do with the test statistic to calculate the p-value).

   p-value calculation: If McNemar's, this is a 2-sided test and we need a 1-sided p-value. Compute p-value from probability (area) above 4.57 on χ^2_1 and divide it by 2, since observed proportion is in the expected direction ($11/14 > 0.5$).

   If Score or Wald test, compute probability of a value larger than 2.13 (or 2.61 for Wald) on a standard normal distribution.

STATISTICS 475: Applied Discrete Data Analysis
MIDTERM 1
Feb 28, 2023

INSTRUCTIONS: **DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO!**

- Answer all parts of all problems in the space provided.
- Place final answers in the blanks when they are provided.
- **Show work! For all calculations**, it is enough to write out the exact equation with the right numbers so that I could calculate the result myself.
- If you need more space (I don't think you should), indicate clearly where the problem continues.
- Use $\alpha = 0.05$ unless told otherwise

PROBLEM 1 (1 pts)
A complicated model has a parameter that is NOT estimated using maximum likelihood. Someone has developed a new method for testing hypotheses about this parameter. How can you find a confidence interval for the parameter?
Invert the hypothesis test. Specifically, repeatedly run the test to find the boundary values of the parameter between rejecting and not rejecting the null hypothesis. These boundaries are the interval endpoints.

PROBLEM 2 (4 pts)
See the Fellowship Problem Description provided before the test and below.

> Nine companions of several different races form a fellowship and embark on a dangerous quest. They will travel together for the most part, but there will be times that they break off into smaller groups, and some will occasionally travel as individuals. There are many deathly perils along the quest. The dangers are different for different individuals and groups along different parts of the journey.

Consider the random variable counting the total number of companions who survive the quest.

1. **(1 pt)** Based only on this brief problem description, are the assumptions of the binomial distribution reasonably satisfied here? Of so, explain why. If not, explain which assumption(s) might you be most concerned about and why.
   We should be concerned about independence and equal probability. Since members sometimes face dangers together, then their survival or deaths may be correlated. And different members face different perils, so not all have exactly the same probability of survival

2. Regardless of your previous answer, suppose that all of the assumptions are satisfied. Assume further that each member of the fellowship has a 2/3 chance of surviving the quest.

   (a) **(2 pt)** What is the probability that at least 8 survive the quest? (Provide a calculation-ready formula, but you do not have to actually calculate the final answer.)
   Need to find $P(W \geq 8)$ *when* $W \sim Bin(9, 2/3)$. $P(W \geq 8) = P(W = 8) + P(W = 9)$ *and* $p(W = w) = \frac{n!}{w!(n-w)!}\pi^w(1-\pi)^{n-w}$
   Answer is $\frac{9!}{8!1!}(2/3)^8(1/3)^1 + \frac{9!}{9!0!}(2/3)^9(1/3)^0$

   (b) **(1 pt)** What is the expected number of surviving participants?
   $n\pi = 9 * 2/3 = 6$

PROBLEM 3 (6 pts)
See the Raccoon Problem Description provided before the test. The description is repeated below for convenience.

> Raccoons living in urban areas frequently forage for food in humans' trash bins, earning them the nickname, "trash pandas". *Researchers studying the behaviour of urban raccoons wanted to determine whether raccoons were more attracted to fresh meat or meat that was spoiled from sitting out too long*, which often happens in trash bins. They identified 30 locations in an urban area that had trash bins where raccoons were known to forage. On two nights, one week apart, they placed 1kg of meat in a loose paper wrapping in the bottom of each trash bin. One night the meat was fresh, and the other night it was spoiled from sitting unrefrigerated for two days prior. For each location they flipped a coin to determine which week would have the fresh meat. They put cameras on each trash bin for 24 hours to determine whether raccoons took the meat in that time.

The data are summarized in the table below. For example, there were 8 locations where raccoons took both the fresh meat and the spoiled meat, and 5 locations where they took only the spoiled meat.

		Fresh		
		Take	Leave	Total
Spoiled	Take	8	5	13
	Leave	11	6	17
	Total	19	11	30

1. **(2 pts)** The researchers' main question is in italics above. Considering the data above, define symbols to represent the population parameters of interest in this problem and rewrite this question into null and alternative hypotheses using these symbols.

   H_0 : $\pi_{1+} - \pi_{+1} = 0$ H_a : $\pi_{1+} - \pi_{+1} \neq 0$
   Symbol definitions (what quantity does each parameter symbol represent?):
   π_{1+} = *probability that raccoons take spoiled meat*
   π_{+1} = *probability that raccoons take fresh meat*

2. **(2 pts)** Write out the formula for the test statistic that is most appropriate for performing this test, plugging in the numbers from the data. The final answer should be numbers in a formula that one could easily punch into a calculator or program into R. You don't have to do the calculation.
   These are paired data, because each location is measured twice under two different conditions. Using McNemar's Test:
   $M = \frac{(n_{01}-n_{10})^2}{n_{01}+n_{10}} = \frac{(11-5)^2}{11+5} = 36/16$

3. Suppose that the p-value from this test were 0.08.

   (a) **(1 pt)** Using a 5% type 1 error rate, what conclusion would you draw from the hypothesis test about raccoons' preference for fresh or spoiled meat?
   *We would conclude that there is **not** sufficient evidence of a preference for either type of meat. (We would NOT reject H_0)*

   (b) **(1 pt)** Suppose that the researchers had reason to believe that raccoons preferred spoiled meat and wanted to specifically test THAT hypothesis instead. What would the p-value be for that test?
   We would change $H_a : p_{1+} - p_{+1} > 0$. However, the direction of the difference ($13/30 - 19/30$*) points in the opposite direction, so take 1-sided p-value as* $1 - 0.08/2$*=0.96*

PROBLEM 4 (2 pts)
Someone proposes to use a log link for a binomial regression model.

1. What concern should you have about this plan that you wouldn't have with a logit link?
   The logit link forces probabilities to remain between 0 and 1 for all values of linear predictor. The log link would allow probabilities to be > 1 *for large values of the linear predictor*

2. If the linear predictor is estimated to be 1.0, what is the estimated probability of success?

   ANSWER: *e^1 (notice that this is >0)*

BONUS: (1 pt) In which movie franchise was the phrase "I don't know...fly casual" spoken?

1. Avalon
2. Fast and Furious
3. Harry Potter
4. Star Trek
5. ***Star Wars***
6. Top Gun

PROBLEM 5 (5 pts)
See the Horseshoe Crab Data Description provided before the test. Refer to the program and output from the program *Midterm 1 2023 HorseshoeCrabs.R.* Refer to the models using `Weight` and `Width` as possible explanatory variables.

1. **(1 pt)** Is there evidence that the effect of the female's width on the probability of a satellite depends on the weight of the crab? Explain.
   No. From the LR tests on mod4, the test for interaction is not significant (p=0.34)

2. **(1 pt)** In the model where `Weight` is the only variable, it appears to be significant, but in the model where `Weight` and `Width` are both in the model, it is not. Explain why this outcome can happen (Hint: what hypotheses are tested by these two tests?).
   In the model with both variables, the test for `Weight` assumes that `Width` is in the reduced model, but in the model with just `Weight`, the reduced model is empty. `Weight` may not explain as much when `Width` is in the model as when it is alone (e.g. if they are correlated)

3. **(1 pt)** The function `Anova(mod4)` produces 3 tests, each of which is a comparison of two of the fitted models, `mod1`, `mod2`, `mod3`, and `mod4`. In the test for `Weight:Width`, which two models are the full and reduced models? (Use model labels "mod*" as answers)

   Full: mod4 Reduced: mod3

4. **(1 pt)** Consider the model

$$\text{logit}(\pi) = \beta_0 + \beta_1 \texttt{Weight} + \beta_2 \texttt{Width} + \beta_3 \texttt{Weight} \times \texttt{Width}.$$

   Compute a likelihood ratio statistic for a single test of $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ against H_a : at least one of these parameters is not zero. Report the test statistic value (or the formula with all numbers in place) and the degrees of freedom for the χ^2 approximation for its distribution.
   Compare the residual deviances from mod4 and mod0

5. Test Stat: 225.76−198.98 DF: 3 (=172-169)

6. **(1 pt)** Find a 95% confidence interval for the effect of a 1-unit increase in weight on the odds that a female crab has a satellite, assuming that this effect is constant for all values of width, and that width is held constant. If calculations are required, answer may be the formula with numbers.
   Need to exponentiate confidence interval for `Weight` parameter in model with `Width` but no interaction (`mod3`). The LR confidence interval from `confint(mod3)` is -0.48 to 2.20, so the CI on the OR is $\exp(-0.48), \exp(2.20)$

PROBLEM 6 (7 pts+1 bonus)
See the Horseshoe Crab Data Description provided before the test. Refer to the program and output from the program *Midterm 1 2023 HorseshoeCrabs.R.* This problem refers to the logistic regressions modeling the probability that a female has at least one satellite male, using the Color factor `colfac` and Spine factor `spfac` as possible explanatory variables.

1. **(2 pts)** In the table below, write out the indicator variables and their values for each combination of the factors for Color and Spine (`colfac` and `spfac`, respectively). There are more columns than you should need. Use variable names for the indicators that clearly distinguish between the factors (e.g., $c_1, c_2, \ldots$ for Color and $s_1, s_2, \ldots$ for Spine). **Include ALL indicators, and not just the ones that R uses.**

		Indicators												
`Color`	`Spine`	c_1	c_2	c_3	c_4		s_1	s_2	s_3					
1	1	1	0	0	0		1	0	0					
1	2	1	0	0	0		0	1	0					
1	3	1	0	0	0		0	0	1					
2	1	0	1	0	0		1	0	0					
2	2	0	1	0	0		0	1	0					
2	3	0	1	0	0		0	0	1					
3	1	0	0	1	0		1	0	0					
3	2	0	0	1	0		0	1	0					
3	3	0	0	1	0		0	0	1					
4	1	0	0	0	1		1	0	0					
4	2	0	0	0	1		0	1	0					
4	3	0	0	0	1		0	0	1					

2. **(2 pt)** Using the indicators you laid out above, write out the mathematical model that R used for the model fit in `modc1`. Use symbols for parameters that are clear and easy to understand (e.g., β^S for spine parameters, and so forth).

   $\text{logit}(\pi) = \beta_0 + \beta_2^C c_2 + \beta_3^C c_3 + \beta_4^C c_4 + \beta_2^S c_2 + \beta_3^S c_3$

3. **(2 pts)** Interpret the following parameter estimates. (What do they say about the relationship between the presence of satellites and either Color or Spine? Interpret the *estimate*, not the test.)

   (a) `colfacdark` in `modc1`
   Difference in log odds of satellite or logits (or log OR) between a dark crab and a medium light crab when the spine is good. Dark crabs have 2.19 lower log odds of satellite than medium-light when the spine condition is good.

   (b) `colfacdark:spfacworn` in `modc2`
   The log OR between a dark crab and a medium light crab with a worn spine is 15.3 lower than the same log-OR for a crab with a good spine.

4. **(1 pts)** Consider `modc1`. If I wanted to find a likelihood-ratio confidence interval for the *probability* of satellites for a medium dark female with a good spine, what coefficients would I need to use in `mcprofile()`?
   $\beta_0 + \beta_3^C$

   $a_1 =$ 1 , $a_2 =$ 0 , $a_3 =$ 1 , $a_4 =$ 0 , $a_5 =$ 0 $a_6 =$ 0

5. (1 pt Bonus) Consider `modc2`. If I wanted to find a likelihood-ratio confidence interval for the *odds ratio* comparing dark to medium color for females with a broken spine what coefficients would I need to use in `mcprofile()`?

   See ordering in output.

   $a_1 =$ 0 , $a_2 =$ −1 , $a_3 =$ 0 , $a_4 =$ 1 , $a_5 =$ 0 $a_6 =$ 0 , $a_7 =$ 0 , $a_8 =$ 0 , $a_9 =$ 0 , $a_{10} =$ −1 $a_{11} =$ 0 , $a_{12} =$ 1