Caltech Department of Mathematics

Ma 3/103
Introduction to Probability and Statistics

KC Border
Winter 2020

Final Exam

Due Tuesday, March 24 by 4:00 pm PDT *via email*.
Use the address CaltechMath3Work@gmail.com.

There is **no time limit**, but it shouldn't take more than four hours.
No collaboration is allowed.

This exam has five multipart questions. Some questions will require thought on your part on how to deal with data. There is not necessarily only one "correct" answer, but there are bad answers. Since you will be using software, there is **no time limit**, and you do *not* have to complete it in one sitting.
You will need to use statistical software during the exam.

- If you have any questions about these instructions, consult a TA for the course (they are listed on the course web page) or the professor (kcb@caltech.edu).
- Write legibly in complete sentences and explain yourself. Part of scientific communication is letting others know why you are correct. Attach any printouts or charts, making sure they are labeled in a way that makes clear what they are and which questions they pertain to.
- You may use the textbooks (Pitman [3]; Larsen–Marx [2]), any supplementary text listed on the course web page, homework solutions, lecture notes, and other handouts from the current Ma 3 web site and auxiliary web site, your own notes and homework, your midterm, and TA notes. You may also use someone else's notes that you have copied by hand.
- You are allowed to use statistical software (including Wolfram Alpha) for computations, and you may refer to the course web site or the auxiliary course web site. Since the software's documentation often resides on-line, you may use on-line documentation. **You may not use any other internet resources.**
- **Simply attaching a printout of your computer session is not an acceptable answer**. You must write a narrative to explain in your own words what you did, why you did it, and what the results are.

Exercise 1 (15 pts) As a rule, I hate True/False and multiple choice exams, but there has been a lot of discussion lately of the state of statistical literacy, e.g., [1] so I am including a few standard questions that appear on surveys.

1. A researcher conducts an experiment, analyzes the data, and reports:

   The 95% confidence interval for the mean μ is [0.1, 0.4].

   Which of the following are accurate, and why?

   (a) (5 pts) The null hypothesis $H_0\colon \mu = 0$, that the true mean equals 0, would be rejected at the 5% level of significance.

   (b) (5 pts) If we were to repeat the experiment over and over, then 95% of the time, the true mean falls between 0.1 and 0.4.

   (c) (5 pts) If we were to repeat the experiment over and over, then 95% of the time, the estimated mean would fall between 0.1 and 0.4. □

Exercise 2 (25 pts) In HW 8 you used your favorite software to report the results of a multiple linear regression analysis of one of Anscombe's quartet, and its standard report tells you that the your estimate $\hat{\beta}$, your estimated coefficient on x is equal to 0.5, that its standard error is 0.12, its t-value is 4.2 with 9 degrees of freedom, and p-value equal to 0.002. (I have rounded off the results you should have gotten in HW 8.)

1. (5 pts) What is the null hypothesis regarding β that your software assumes you are interested in testing? What is the result of that test at the $\alpha = 0.05$ significance level?

2. (5 pts) What is the relationship between the estimate $\hat{\beta}$, the t-value t, and the standard error $\hat{s}$ as reported?

3. (15 pts) Suppose you want to test the null hypothesis $H_0\colon \beta = 1$ at the $\alpha = 0.01$ significance level. Have I given enough information from the experiment? If not, what else do you need? What is the result of your test? □

Exercise 3 (30 pts) L&M [2, p. 528] cite a study of the relation the blood pressure of children and their fathers. Blood pressure was categorized according to whether it was in the lower third, the middle third or the upper third of their cohort. Here are the data they report.

		Child's BP		
		Lower	Middle	Upper
Father's BP	Lower	14	11	8
	Middle	11	11	9
	Upper	6	10	12

How would you test the null hypothesis H_0 that the child's classification is stochastically independent from the father's?

Describe the test statistic you would use, its distribution under the null hypothesis, and the critical value(s) for the test. Should it be one- or two-sided?

Now carry out the test. What is your decision? □

Exercise 4 (30 pts) Remember the earthquake data from HW 8? If the waiting times between quakes are independent, letting w_n be the time between earthquake $n-1$ and earthquake n, then a regression of the form

$$w_n = \beta_0 + w_{n-1}\beta_1 + e_n,$$

should have $\beta_1 = 0$ and have little explanatory power. (although there could be other ways the data could be dependent.) To see if this is true of the data, I decided to take only quakes with magnitude $\geqslant 6.0$, and found the times between them. (Why 6.0? You'll definitely notice when one occurs.) I then created a y variable by dropping the first value, and x variable by dropping the last value, so that both variables have 31 observations. Thus the regression equation

$$y_n = \beta_0 + x_n\beta_1 + e_n$$

corresponds to the regression of w_n on w_{n-1}. I put these on the website at http://www.its.caltech.edu/kcborder/Courses/Ma3/Data/EarthQuake6Lags.txt. The file contains x in the first column an y in the second, and there is a header line. The format is the same as the files for Anscombe's quartet in HW 8.

Analyze these data, and convince of you conclusions. First explain why we care if the times are independent. Then sure to explain your estimates, what tests you performed, at what level of significance, and how you drew your conclusions. □

Exercise 5 (65 pts) A standard technique used by field biologists to estimate animal populations is the **capture-recapture** method. To be concrete, imagine a lake containing an unknown number N of fish. Some of these fish are caught, tagged, and released back into the lake. The total number of tagged fish is T. A while later, another sample of size S is taken **without replacement**, and it is found that X of them are tagged.

1. (10 pts) What is the probability distribution of X? That is, give a formula for $P(X = x)$ in terms of T, N, x, and S.

2. (5 pts) Discuss the assumptions on the nature of fish and the procedure for sampling that you used to justify the distribution above. (For instance, what if tagging a fish is so traumatic that it dies soon thereafter?)

3. (5 pts) Give a formula for the likelihood $L(N; x)$.

4. (15 pts) What is the Maximum Likelihood Estimator of N?

   Hint: Look at the ratio $L(N; x)/L(N-1; x)$. When is this ratio > 1?

5. (5 pts) Are there potential problems with this estimator?

6. What would change in your analysis if the second sample had been taken with replacement? In particular,

   (a) (10 pts) What is the new probability distribution of X?

   (b) (15 pts) What is the new Maximum Likelihood Estimator of N?

□

References

[1] R. Hoekstra, R. D. Morey, J. N. Rouder, and E.-J. Wagenmakers. 2014. Robust misinterpretation of confidence intervals. *Psychonomics Bulletin Review* 21(5):1157–1164. DOI: 10.3758/s13423-013-0572-3

[2] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[3] J. Pitman. 1993. *Probability*. Springer Texts in Statistics. New York, Berlin, and Heidelberg: Springer.