Caltech Department of Mathematics

Ma 3/103
Introduction to Probability and Statistics

KC Border
Winter 2017

Final Examination

Due Thursday, March 16, 2017 at 4:00 pm
in the Exam Drop Box outside 253 Sloan.

Instructions

This exam has 2 questions on 2 pages, not including this cover page. The questions will require some thought on your part on how to deal with data in a context slightly different from what you have not seen before. There is not necessarily only one "correct" answer. There is **no time limit**, and you do not have to complete it in one sitting. **You will need access to the course web site and 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.
- No collaboration is allowed.
- You may use the textbooks (Pitman; Larsen and Marx), any supplementary text listed on the course web page, homework solutions, lecture notes, and other handouts from the current Ma 3 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 may not use any internet sources other than the current course web site.**
- I recommend you use a program such as R or Mathematica to look up values of the cumulative distribution functions and quantile functions for the standard normal, the Student t, the χ^2, and the F distributions.

  You may also use statistical software to compute t-tests or regressions, etc., **but 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 and what the results are.

1 Theory

1. 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. A sample of size M of these fish is caught, tagged, and released back into the lake. A while later, another sample of size S is taken with replacement, and it is found that X of them are tagged.

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

   (b) (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?)

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

   (d) (10 pts) What is the Method of Moments Estimator of N?

   (e) (5 pts) Are there problems with these estimators? If so, what might you do?

   (f) (20 pts) What would change in your analysis if the second sample had been taken without replacement?

2 Practice

2. The Poisson distribution is named for the French mathematician Siméon Denis Poisson (1781–1840). Poisson is also the French word for fish. If you have ever done any angling (fishing for the fun of it) you may have observed that actually catching a fish is a rare event. Ladislaus von Bortkiewicz in his *Das Gesetz der kleinen Zahlen* [2] popularized the notion of the Poisson distribution for the number of occurrences of rare events.

   Thompson [1] reports data from the 1969 Creel Census of the Lower Current River in Missouri. These data were collected by game wardens and represent the actual number of fish caught and kept by a sample of 911 anglers. Here is a summary of the data:

Catch size	Number of anglers
0	515
1	65
2	60
3	66
4	53
5	55
6	27
7	25
8+	45

   I have also put the data in a TAB-separated text file on the course web site. Be aware that most computers do not recognize 8+ as a number.

(a) (30 pts) Propose and carry out an appropriate test for the hypothesis that the catch sizes are characterized by the Poisson distribution. Be sure to explain how you treated 8+ and why.

  i. Estimate the Poisson parameter μ by the method of maximum likelihood.
  ii. Carefully describe your null hypothesis.
  iii. Describe your test statistic.
  iv. What is its distribution under the null hypothesis? (How many degrees of freedom?)
  v. Describe your critical region.
  vi. What is your conclusion?

(b) (40 pts) An alternative to the Poisson model that has found wide use is the ZIP, or Zero-Inflated Poisson, model. In this model, there are two kinds of fishermen. With probability π the angler is not interested in catching fish, and never catches any. With probability $1 - \pi$ the anglers is a Poissonian fisherman who catches a number of fish according to a Poisson distribution with pmf p_μ. Thus the pmf p_{ZIP} for the ZIP model

$$p_{\text{ZIP}}(k) = P\left(k \text{ fish are caught}\right) = \begin{cases} \pi + (1-\pi)e^{-\mu}, & k = 0 \\ (1-\pi)e^{-\mu}\frac{\mu^k}{k!}, & k \geqslant 1. \end{cases}$$

  i. For the ZIP model, the log-likelihood function is

$$\ln L(\mu, \pi; k_1, \dots, k_n) =$$
$$n_0 \ln(\pi + (1-\pi)e^{-\mu}) + (n - n_0)\ln(1-\pi) - (n-n_0)\mu + \ln(\mu)\sum_{i=1}^n k_i - \sum_{i=1}^n \ln k_i!,$$

  where n_0 is the number of i's with $k_i = 0$. You cannot solve for the maximum likelihood estimator analytically, but just as with the World Series data, you can use a numerical solution. (Hint: If your algorithm needs help finding a starting place for μ, try the sample average. And you might try starting with $\pi = n_0/n$.)
  ii. Use the MLE estimates of the ZIP model predict the expected number of fishermen with k fish in their creel, which are different predictions from the straightforward Poisson model. Perform a goodness-of-fit test of the ZIP model.
  iii. Comment on the performance of the ZIP model vis-a-vis the Poisson model.

References

[1] W. A. Thompson. 1976. Fisherman's luck. *Biometrics* 32(2):265–271.
http://www.jstor.org/stable/2529497

[2] L. von Bortkiewicz. 1898. *Das Gesetz der kleinen Zahlen [The law of small numbers]*. Leipzig: B.G. Teubner. The imprint lists the author as Dr. L. von Bortkewitsch.