Caltech Department of Mathematics

Ma 3/103
Introduction to Probability and Statistics

KC Border
Winter 2017

Assignment 8: Specification Testing

Due Tuesday, March 7 by 4:00 p.m. at 253 Sloan

How to answer these questions

These questions require you to use sophisticated computer programs to answer simple questions. When you are asked for a graph or a number, you should describe what commands you used (include the code), and what you expect they should compute. Just reporting a number is not sufficient—you have to explain why it is the right number.

When you produce graphs, make sure you label the axes, and explain what it is a graph of.

In other words, write up your answers as if you were writing for publication in a journal, not as if you are mindlessly churning out problem set answers.

No collaboration is allowed on optional exercises.

Exercise 1 (Earthquakes)

Do earthquakes follow a Poisson process? That is, is the time between earthquakes independently and exponentially distributed? Or equivalently, is the number of earthquakes each year distributed according to a Poisson distribution?

For the first part I am going to ignore everything seismologists know about earthquakes, and treat the data as if it were generated by an unknown stochastic process. Clearly we are going to have to make a number of modeling choices. The most important are the size of the quakes to consider, and the geographic area we restrict attention to. I am going to restrict attention to Southern California, since that is where I live and work.

As for magnitudes, according to Kate Hutton, Jochen Woessner, and Egill Hauksson [2], the Southern California Seismic Network (SCSN) has recorded over 470,000 recorded quakes of magnitude 3.25+ in Southern California since 1932.[1] The 1952 Kern County earthquake sequence produced so many earthquakes, the cataloging effort could not keep up [2, p. 437], so this actually an undercount.

Keeping this in mind I decided to restrict attention to earthquakes of magnitude at least 4.5. (Below that they seem too puny for me to worry about.)

The Southern California Earthquake Data Center at Caltech (http://www.data.scec.org) has an Earthquake Catalog at http://www.data.scec.org/eq-catalogs/date_mag_loc.php. It lists 830 earthquakes of magnitude at least 4.5 in Southern California (see the web page for the latitude and longitude range) from January 1, 1933 through February 28, 2017. See [2] for a more detailed description of what is in the Catalog.

I have created a simplified Catalog to use for this exercise. It is a plain ASCII tab-separated file with Unix-style newline characters. The first field gives the date and time of the earthquake as a fractional number of days since the start of 1933. The first earthquake occurred on 1933/03/11 at 01:54:09.34, so this corresponds to 69.079 days after the start of 1933 (00:00 on January 1). The second field is the magnitude of the quake. The third and fourth fields are the latitude and longitude of the epicenter. The last two fields are the original date and time fields from the catalog. You can find the simplified catalog on the Ma 3 web site

In order to find the waiting times in days between quakes, you merely need to subtract consecutive dates. (See the hints below for how to do this in R and Mathematica.)

1. (10 pts) What is the relationship between the mean and the standard deviation of an exponential distribution?

2. (10 pts) Create a histogram of the inter-arrival times.

3. (10 pts) Find the mean and standard deviation of the inter-arrival times. Do they come close to satisfying the relationship in part 1?

[1] [2, pp. 438–439]: During the past 77 years, the SCSN has recorded more than 470,000 earthquakes. Most of these events were detected in the past two decades because more stations were deployed, data processing procedures improved, and the 1992 M_w 7.3 Landers, the 1994 M_w 6.7 Northridge, and the 1999 M_w 7.1 Hector Mine sequences occurred. However, the number of $M \geqslant 3.25$ events has remained similar throughout the whole time period, except for increased activity during large aftershock sequences. Thus, the earthquake monitoring capabilities for moderate-sized or large events ($M_c \geqslant 3.25$) have remained similar since the 1930s.

4. (10 pts) What is the log-likelihood function for a sample $x_1, \ldots, x_n$ drawn from an Exponential(λ) distribution?

5. (10 pts) Assuming the earthquake inter-arrival times are exponentially distributed with parameter λ, what is the maximum likelihood estimate of λ?

6. (10 pts) Create a Q-Q plot of the quantile of the empirical cdf vs the quantiles of an Exponential distribution with parameter $\hat{\lambda}_{\text{MLE}}$. *Do not create a Normal Q-Q plot.* How does it look?

7. (10 pts) Use a Kolmogorov–Smirnov test to test the null hypothesis that your data are exponentially distributed with parameter $\hat{\lambda}_{\text{MLE}}$ versus the "two-sided" alternative hypothesis that the distributions are different. Does it agree with your visual assessment?

   [Make sure you understand the output of your computer program. Mathematica and R (by default) both compute the same test statistic, which is what R refers to as the two-sided test statistic. (R gives you the option to do a one-sided test for stochastic dominance.) Mathematica gives you no choice to sidedness of the test. The finite-sample distribution of the K–S test statistic is rather complicated and Mathematica and R use different methods to approximate it. As result, the p-values they give for the same statistic differ by about 60%. Don't be alarmed, at any reasonable significance level (5%, 1% 0.1%), they agree on whether to reject the null hypothesis. With either program, the appropriate response is to reject the null hypothesis if the p-value is less than the α-level of significance.]

Now I am willing to let just a tiny bit of science sneak into the pure statistics. Recall that earthquakes often come with "foreshocks" and "aftershocks."[2]

Perhaps if we viewed all the quakes separated by say fewer than four days as a single "event," we would get a better fit to the exponential model.

8. (30 pts)

   Redo parts (2)–(7) with the smaller data set obtained by simply discarding all inter-arrival times less than four days. *Make sure to recompute your means and standard deviations, and your estimate of λ!*

[2]From [2, Abstract]: The three largest earthquakes recorded were 1952 M_w 7.5 Kern County, 1992 M_w 7.3 Landers, and 1999 M_w 7.1 Hector Mine sequences, and the three most damaging earthquakes were the 1933 M_w 6.4 Long Beach, 1971 M_w 6.7 San Fernando, and 1994 M_w 6.7 Northridge earthquakes. All of these events ruptured slow-slipping faults, located away from the main plate boundary fault, the San Andreas fault. Their aftershock sequences constitute about a third of the events in the catalog.

How does the exponentiality hypothesis stand up now?

9. (10 pts) There is no real justification for the four-day minimum above. Suggest a more intelligent, but more time-consuming, approach to deciding which are aftershocks and foreshocks. (Hint: Look at the list of references.)

10. (10 pts) How long do you think we should expect to wait for the next magnitude 4.5+ quake?

11. (20 pts) Construct an appropriate test of the hypothesis that the number of quakes in a given time period follows a Poisson distribution. Be sure to

    - justify your choice of time period.
    - explain your choice of null hypothesis.
    - explain your critical region.

□

Exercise 2 (The World Series, once again)

A World Series can last 4, 5, 6, or 7 games. Recall that our model predicted the probability that a given World Series lasts m games is

$$P(m) = \binom{m-1}{m-4}\Big(p^4(1-p)^{m-4} + (1-p)^4 p^{m-4)}\Big),$$

where p is the probability that the better team wins. The 108 best-of-seven World Series produced the following results:

Length	4	5	6	7	Total
Number	21	25	24	38	108

Previously you used the method of maximum likelihood to estimate p at 0.5972 (at least that's what I got), which predicts the following expected numbers (rounded to one decimal place) of each length.

Length	4	5	6	7
$\hat{p}$	0.16	0.27	0.30	0.28
Expected	16.8	29.0	32.3	29.9

Use a chi-square test [3, Section 10.4] to provide a specification test of the model we have been using.

1. (5 pts) Carefully state the null hypothesis.

2. (15 pts) Write out by hand the formula for the test statistic. (Hint: All the numbers you need are in the two tables above.) What is the value of the test statistic? (You may use a computer/calculator to evaluate the formula.)

3. (5 pts) Should you use a two-sided test or a one-sided test? Why?

4. (10 pts) Explain how many degrees of freedom you should use. (Remember, p was estimated by MLE.) What is the critical value of the test statistic? Why?

5. (5 pts) Draw a rough sketch of the pdf to illustrate the critical value for a test at the $\alpha = 0.05$ level of significance.

6. (5 pts) What is the p-value of the test statistic you computed?

7. (5 pts) Do you reject or fail to reject the null hypothesis at the $\alpha = 0.05$ level of significance?

□

Exercise 3 (10 pts) How much time did you spend on the previous exercises? (Do not include time spent on optional exercises.) □

Exercise 4 (Optional Exercise) (40 pts) Alex tosses a fair coin n independent times and Blair tosses a fair coin m independent times. Find an *elegant* or clever argument to compute the probability that they have equal numbers of Tails. (I will be the judge of whether the argument is elegant, but it had better not involve any lengthy sums.) □

References

[1] J. K. Gardner and L. Knopoff. 1974. Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian? *Bulletin of the Seismological Society of America* 64(5):1363–1367.
http://bssa.geoscienceworld.org/content/64/5/1363.full.pdf+html

[2] K. Hutton, J. Woessner, and E. Hauksson. 2010. Earthquake monitoring in Southern California for seventy-seven years (1932–2008). *Bulletin of the Seismological Society of America* 100(2):423–446. DOI: 10.1785/0120090130

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

[4] P. Teetor. 2011. *R cookbook*. O'Reilly Media.
http://shop.oreilly.com/product/9780596809164.do

Mathematica hints for Problem 1

Try importing the data with something like

```
quakes = Import[NotebookDirectory[] <> "SimplifiedEarthquakeCatalog", "Table",
   "HeaderLines" -> 0]
```

The dates are in column 1, so

```
dates = quakes[[All, 1]]
```

gives you an array of just dates. To get the inter-arrival times you have to look at the successive differences:

```
times = Differences[dates]
```

Now you have the inter-arrival times in days.

Now you can use `Histogram` on `times`. You probably want to convert those integers to floating point with the `N` function before using `Mean` and `StandardDeviation`.

Now you have to figure out the maximum likelihood estimator $\hat{\lambda}$. Then you can substitute that into your Kolmogorov–Smirnov Test

```
KolmogorovSmirnovTest[times, ExponentialDistribution[ estimated value goes here ]]
```

Make sure you know what the output of the test is. Maybe you should check out the documentation on `HypothesisTestData`:

```
htd = KolmogorovSmirnovTest[times,
  ExponentialDistribution[ESTIMATED LAMBDA], "HypothesisTestData"]

htd["TestConclusion"]
htd["TestDataTable"]
```

To drop values less than 4 from `times`, you can use:

```
times2 = Select[times, # > 4 &]
```

Mathematica hints for Problem 2

To find critical values use the `InverseCDF` function. For instance to find the value x^* such that 95% of the probability in a Chi-square with 8 degrees of freedom lies to the left of x^* use:

```
xstar = InverseCDF[ChiSquareDistribution[8], 0.95]
```

To find the p-value of the test statistic for a one-sided Chi-square test with 8 with degrees of freedom, when the test statistic has value `X`, use

```
1 - CDF[ChiSquareDistribution[8], X]
```

R hints for Problem 1

These hints have been tested with R Studio on a Macintosh.

Input the data. Note the quotation marks around file and path names. Look at it.

```
setwd("your/path/goes/here")
quakes = read.table("SimplifiedEarthquakeCatalog",header=T)
quakes

quakes$DateSerial
```

gives just the list of dates.

To get the inter-arrival times the `diff()` function generates successive differences along a vector:

```
times = diff(quakes$DateSerial)
```

Now we have the list of inter-arrival times.

Now you can use the `mean`, `sd`, and `hist` functions. Don't forget to compute the Maximum Likelihood Estimate of λ.

Creating a Q-Q plot versus the exponential is a bit tricky. I found this method in *The R Cookbook* [4, p. 255].

`plot(qexp(ppoints(times), rate=`$\hat{\lambda}$`), sort(times))`

`qexp` is the quantile function (inverse cdf) for the exponential family. `ppoints` scales the vector `times` to fit into $(0, 1)$, and `sort` sorts it. Actually, you probably want to label things, and draw in a line of slope 1, so you want to use

```
plot(qexp(ppoints(times), rate=ESTIMATE OF LAMBDA GOES HERE),
  sort(times), main="Exponential Q-Q Plot",
  xlab="Theoretical Quantiles", ylab="Sample Quantiles")

abline(a=0, b=1)
```

For a Kolmogorov–Smirnov Test of exponentiality, you can use

```
ks.test(times, pexp, rate=ESTIMATE OF LAMBDA GOES HERE)
```

(`pexp` is the exponential cdf.)

To select the times greater than 4, use:

```
times2 = times[times > 4]
```

Note the square brackets.

R hints for Problem 2

To find critical values, use the quantile function. The quantile function for the Chi-square is `qchisq`. For instance, to find the value x^* such that 95% of the probability in a Chi-square with 8 degrees of freedom lies to the left of x^*, use:

```
xstar = qchisq(0.95, 8)
```

To find the p-value of the test statistic, use the cdf function. The cdf for a Chi-square is `pchisq`. For a one-sided Chi-square test with 8 with degrees of freedom, when the test statistic has value `X`, use:

```
1 - pchisq(X, 8)
```