Caltech Department of Mathematics

Ma 3/103 KC Border
Introduction to Probability and Statistics Winter 2020

Assignment 8

Due Tuesday, March 10 by 4:00 p.m.
in the dropbox in the lobby of Linde Hall.

Instructions:

When asked for a probability or an expectation, give both a formula and an explanation for why you used that formula, and also give a numerical value when available.

When asked to plot something, use informative labels (even if handwritten), so the TA knows what you are plotting, attach a copy of the plot, and, if appropriate, the commands that produced it.

No collaboration is allowed on optional exercises.

Exercise 1 (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 111 best-of-seven World Series produced the following results:

Length	4	5	6	7	Total
Number	21	26	24	40	111

Previously you used the method of maximum likelihood to estimate p at 0.5906 (at least that's what I got, rounded to four decimal places), which predicts the following expected numbers (rounded to two decimal places) of each length.

Length	4	5	6	7
$\hat{p}$	0.150	0.266	0.302	0.283
Expected	16.62	29.48	33.51	31.38

Use a chi-square test [4, 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 2 (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. A map of the region designated as Southern California by Southern California Earthquake Data Center is shown in Figure 1.

As for magnitudes, according to Kate Hutton, Jochen Woessner, and Egill Hauksson [3], 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 [3, p. 437], so this is 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.)

[1] [3, pp. 438–439]: During this time, 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.

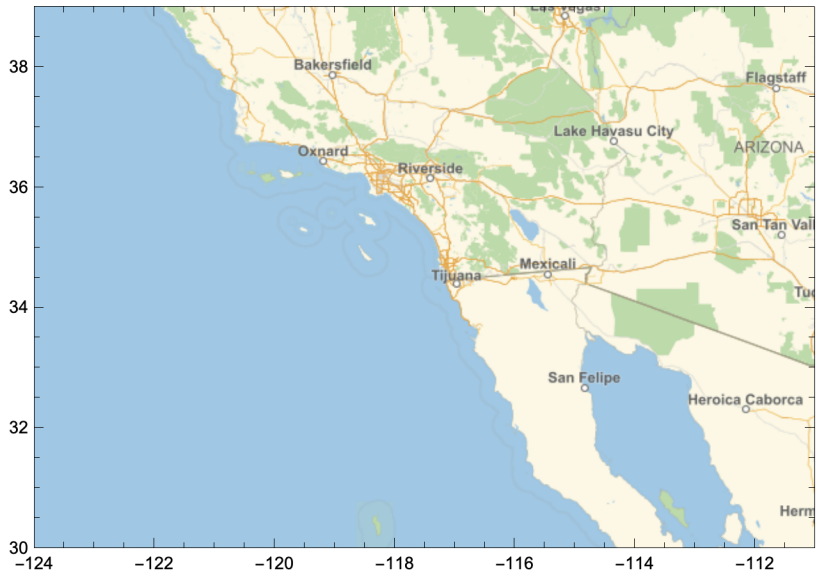

Figure 1. The region covered by the earthquake data.

The Southern California Earthquake Data Center has an Earthquake Catalog at http://service.scedc.caltech.edu/eq-catalogs/date_mag_loc.php. It lists 867 earthquakes of magnitude at least 4.5 in Southern California (see Figure 1) from January 1, 1933 through February 22, 2020. See [3] 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 1933/02/24 at 19:33:17.51 so this corresponds to 54.81478 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 auxiliary 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?

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.

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!*

   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 should we expect to wait for the next magnitude 4.5+ quake? □

Exercise 3 (Anscombe's quartet)

Francis Anscombe [1] created and presented four small sets of data. You can find them on the course auxiliary web site at http://www.its.caltech.edu/~kcborder/Courses/Ma3/Data/Anscombe1.txt, Anscombe2, Anscombe3, and Anscombe4. Each data set has 11 observations on two variates labeled X and Y. (Each file has a header line.)

[2]From [3, 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.

1. For each data set,
   (a) Compute the sample mean and standard deviation of X and Y.
   (b) Regress Y on X and a constant term. Explain the computations you are doing. (If you omit this information, we shall be unable to award partial credit if your calculations are incorrect.)
2. (5 pts per table cell) Summarize your results in Table 1 and turn it in with your assignment. In the Table, β_0 refers to the coefficient on the constant (or intercept term) and β_1 refers to the coefficient on X.
3. (10 pts) Do these results allow you to conclude anything about the similarities and/or differences in the relationship between X and Y in these different data sets?
4. (10 pts) For each data set, make a scatter plot of Y against X, showing the regression line. (Label and turn in the plots.)
5. (10 pts) For each data set, create a Normal Q-Q plot of the residuals, and perform a Kolmogorov–Smirnov test for normality of the residuals. For which data sets you reject the hypothesis? Put it in the Table. (Label and turn in the plots.)
6. (5 pts) Do these scatter plots allow you to conclude anything about the similarities and/or differences in the relationship between X and Y in these different data sets?
7. (10 pts) Which data set(s) come closest to satisfying the assumptions of the standard linear model? □

Exercise 4 (10 pts) How much total time did you spend on the preceding exercises? **Please put the answer to this exercise on the *front page* of your answers and identify it as such.** □

Exercise 5 (Optional Exercise) (40 pts) An urn contains 4 balls each of a distinct color. At each step we draw two balls randomly, and change the color of the second one to the color of the first one, then we return the balls to the urn. What is the expected time of arriving to the case where all balls have the same color? (The process of drawing the two balls and replacing them takes place in one time period.) □

References

[1] F. J. Anscombe. 1973. Graphs in statistical analysis. *The American Statistician* 27(1):17–21. http://www.jstor.org/stable/2682899

[2] 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

	Data set 1	Data set 2	Data set 3	Data set 4
Mean of X				
Std. Dev. of X				
Mean of Y				
Std. Dev. of Y				
$\hat{\beta}_0$				
t-statistic for $\hat{\beta}_0$				
p-value of t-statistic for $\hat{\beta}_0$				
Reject $H_0\colon \beta_0 = 0$? Y/N				
$\hat{\beta}_1$				
t-statistic for $\hat{\beta}_1$				
p-value of t-statistic for $\hat{\beta}_1$				
Reject $H_0\colon \beta_1 = 0$? Y/N				
R^2				
Adjusted $\bar{R}^2$				
F-statistic for the regression				
p-value of F-statistic				
Sum of squared residuals				
Reject H_0: Normal residuals? Y/N				

Table 1. Summary of Anscombe's quartet. (You may round to two or three decimal places if you wish.)

[3] 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

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

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

Mathematica hints for Problem 1

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]
```

Mathematica hints for Problem 2

Try importing the data with something like

```
quakes = Import[NotebookDirectory[] <> "SimplifiedEarthquakeCatalog2020.txt", "Table",
   "HeaderLines" -> 1]
```

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 &]
```

R hints for Problem 1

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)
```

R hints for Problem 2

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("SimplifiedEarthquakeCatalog2020.txt",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.

To create a Q-Q plot versus the exponential:
`qqplot(qexp(ppoints(times), rate=`$\hat{\lambda}$`), times)`
Here `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

```
qqplot(qexp(ppoints(times), rate=ESTIMATE OF LAMBDA GOES HERE),
  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.

Introduction to simple regression with R

R is an object-oriented language, and regression is carried out via an `lm` (or linear model) object. The oddest thing about an `lm` object is its `formula` property. A typical `formula` specification looks like `formula=y~x`, which means to regress y on x and a constant. (Constants are supplied by default. To get rid of the constant you could say `formula=y~-1+x`, but we won't need that for this assignment.) The `lm` object also requires a `data` specification. The are other properties that could be set, but the default values are adequate for us.

First you read in the data, as usual. You may need to supply a full path or use `setwd` to set the working directory to where you data files are kept. Let's call the data `a`, as in Anscombe.

```
a=read.table("Anscombe1.txt", header=TRUE)
```

You can use `sapply` to calculate the means and standard deviations. (`sapply` applies a function to a list and returns a vector.)

```
sapply(a,mean)
sapply(a,sd)
```

Now you can create a linear model object. Let's call it `mod`. Because we read in the data with headers, R knows what X and Y are.

```
mod=lm(formula=y~x, data=a)
```

Much useful output is obtained by simply asking for a summary:

```
summary(mod)
```

This should have provided you with estimated coefficients, t-statistics, the R^2, the F-statistic for the regression, and some other stuff. This is all you need for the first problem.

There are other useful things in the object. I suggest you try:

```
coef(mod)
residuals(mod)
```

The sum of squared residuals is not automatically reported, but it's not hard to calculate.

```
e=residuals(mod)
ssr=sum(e*e)
ssr
```

You can compute the s^2 statistic $\dfrac{e'e}{T-K}$, and take its square root:

```
T=length(a$x)
K=length(coef(mod))
sqrt(ssr/(T-K))
```

Where have you seen that number before?

How do you create a scatter plot with the regression line? Here's one way:

```
plot(a)
abline(mod)
```

And here's a normal Q-Q plot of the residuals, with a Kolmogorov–Smirnov test for normality thrown in for good measure.

```
e=residuals(mod)
qqnorm(e)
abline(a=0,b=1)
ks.test(e,pnorm)
```

Introduction to simple regression with Mathematica

With Mathematica, regression is carried out via a `LinearModelFit` object. First you read in the data, as usual. You may use `SetDirectory` to set the working directory to where you data files are kept. Let's call the data `a`, as in Anscombe.

```
a = Import["Anscombe1.txt", "Table", "HeaderLines" -> 1];
```

Then

```
Mean[a]
StandardDeviation[a]
```

Now you can create a linear model object. Let's call it `mod`.

```
mod = LinearModelFit[a, x, x]
```

To get the fitted model, the fitted model the following returns a function that can be used with `Plot`.

```
Normal[mod]
```

To get information on the estimated coefficients and on the overall regression:

```
mod["ParameterTable"]
mod["RSquared"]
mod["AdjustedRSquared"]
```

To get the F-statistic for the regression and its p-value:

```
mod["ANOVATableFStatistics"]
mod["ANOVATablePValues"]
```

To get the sum of squared residuals:

```
e = mod["FitResiduals"];
ssr = Total[e e]
```

You can compute the s^2 statistic $\dfrac{e'e}{T-K}$, and take its square root:

```
T = Length[a];
Sqrt[ssr/(T-2)]
```

How do you create a scatter plot with the regression line? Here's one way:

```
Show[ListPlot[a], Plot[Normal[mod], {x, Min[a[[All,1]]], Max[a[[All,1]]]}] ]
```

And here's a normal Q-Q plot of the residuals,

```
QuantilePlot[e]
```

To test the normality of the residuals, you should normalize them by dividing by their standard deviation and using a KS-test for the standard normal distribution. See Subsection 24.2.2 of the notes for a discussion of some of Mathematica's anomalous behavior with regard to KS tests.