A Vignette on Laminations and Bering Sea Levels

Christina Knudson, PhD

03/15/2018

Overview

This vignette will demonstrate the process and functions used to answer the following research question: Are laminated intervals primarily associated with high sea levels?

In this vignette, we first cover the hypotheses, the test statistic, and the intuition behind the test. Next, we demonstrate the `Bering` functions necessary to conduct the test. Then, we demonstrate visualizing the results using the `Bering` output. Finally, we discuss the computational cost of this procedure.

Hypotheses, test statistics, and intuition

The null hypothesis is that sea levels and laminations are unrelated; the mean sea level during laminated intervals is similar to the mean sea level overall. The alternative hypothesisis that the sea level is relatively high during laminated intervals.

We select the mean sea level during laminations as our test statistic. Using Monte Carlo, we simulate 10000 data-sets and test statistics under the null hypothesis. To simulate each data-set, we randomly assign each measurement to one of two states (laminated or non-laminated) using a Markov chain with empirical transition probabilities calculated using the original data set.

If the test statistic is relatively large compared to the distribution of simulated test statistics, we will reject the null hypothesis and conclude that the mean sea level is significantly higher during laminations. In particular, our p-value is the proportion of test statistics greater than or equal to the original data's test statistic.

This vignette uses `mocksea`, which has real relative sea levels, real measurement ages, and and simulated laminations. We will update the `Bering` package and this vignette to include the true lamination data if/when our manuscript is published.

Exploring the data

Begin by invoking the `Bering` package and `mocksea` data.

```r
library(Bering)
data(mocksea)
```

At this point, I highly recommend conducting exploratory data analysis to become acquainted with the data. In particular, I recommend using the `summary` command (in base R) to understand the data.

```r
summary(mocksea)
```

```
##       age             sealevel        lam
##  Min.   :   0.10   Min.   :-132.80   0:1210
##  1st Qu.:  90.19   1st Qu.: -74.94   1: 989
##  Median : 291.74   Median : -49.90
##  Mean   : 293.44   Mean   : -49.37
##  3rd Qu.: 418.20   3rd Qu.: -26.16
##  Max.   :1208.09   Max.   :  42.53
```

Our data frame `mocksea` has three variables: `age`, `sealevel`, and `lam`. Information on these variables can be found by typing the following into the `R` console:

```
?mocksea
```

After completing exploratory data analysis, we ensure that the measurements are in order by age. The code below demonstrates one way to perform this check. TRUE indicates the measurements are indeed in order.

```
with(mocksea, sum(age == sort(age)) == length(age))
```

```
## [1] TRUE
```

Conducting the test

To prepare for the hypothesis test, we calculate two empirical probabilities:

- the empirical probability of a laminated measurement following a nonlaminated measurement
- the empirical probability of a laminated measurement following another laminated measurement.

The `peas` function from `Bering` calculates these probabilities using a single input: the vector indicating which measurements are laminated.

```
mypea <- peas(mocksea$lam)
(pswitchL <- mypea$pswitchL)
```

```
## [1] 0.1057851
```

```
(pstayL <- mypea$pstayL)
```

```
## [1] 0.8715875
```

Next, we calculate the test statistic (the mean sea level during laminated intervals) of the original data. The code below demonstrates one way to calculate this statistic.

```
teststat <- with(mocksea, mean(sealevel[lam==1]))
```

Finally, we calculate the p-value using the `onemeanRtest` function from `Bering`. The first argument is the vector of sea levels, the second is the test statistic, the third is the empirical probability of a laminated measurement following a nonlaminated measurement, the fourth is the empirical probability of a laminated measurement following another laminated measurement, and the fifth is the Monte Carlo sample size. The Monte Carlo sample size represents the number of test statistics that will be simulated under the null hypothesis (no association between sea levels and laminations).

```
m <- 10^4
out <- onemeanRtest(mocksea$sealevel, teststat, pswitchL, pstayL,  m)
out$pvalue
```

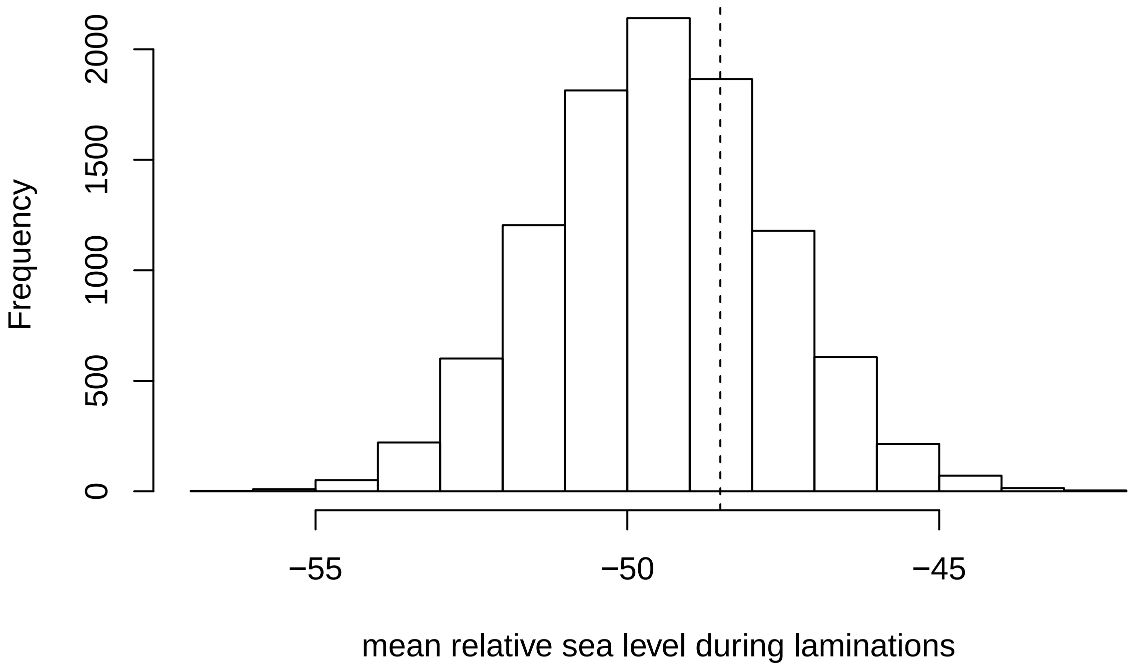
```
## [1] 0.2953705
```

Because the p-value from this test (using mock data) is quite large, we do not reject our null hypothesis. In other words, the mean sea level is not significantly higher during laminations than it would be if sea levels and laminations were truly related.

Visualizing the results

To better understand our results, we can visualize the simulated means and compare them to the original data's mean. We first create a histogram using the simulated means (returned by our test) and then we draw a vertical dashed line at our original data's mean.

```
hist(out$lammean, main=NULL, xlab="mean relative sea level during laminations")
abline(v=teststat, lty=2)
```



This plot makes it clear that our original data's test statistic is not at all unusual compared to the test statistics simulated under the hypothesis that sea levels and laminations are unrelated.

Computational cost

Compiling this vignette on a 'normal' computer (Windows 10 with 8 GB of RAM) took about 34 seconds. Therefore, running all the `R` code in this vignette should take less than 34 seconds.