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 hypothesis, the test statistic used, and the intuition behind the test. Next, we demonstrate the `Bering` functions necessary to conduct the test. Then, we demonstrate how to use the test output to visualize our results. 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 and the alternative hypothesisis that high sea levels are associated with laminations.

To answer this research question, we select the mean sea level during laminations as our test statistic. We compare this test statistic to 10000 test statistics that are simulated under the null hypothesis. These test statistics are simulated using Monte Carlo. The sea levels and dates of measurements remain unchanged, but we randomly assign each measurement to one of two states (laminated or non-laminated) using a Markov chain with empirical transition probabilities.

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.

We present this method using simulated laminations, but we will update the `Bering` package and this vignette to include the true data if/when our manuscript is published.

Conducting the test

Begin by invoking the package and the sea levels data.

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

At this point, I highly recommend you conduct exploratory data analysis to acquaint yourself with the data.

After completing your exploratory data analysis, 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
```

Next, calculate the empirical probability of a laminated measurement following a nonlaminated measurement and the empirical probability of a laminated measurement following another laminated measurement. We use the `peas` command from `Bering`. The only argument for this function is the vector of laminated states.

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

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

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

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

Now calculate the test statistic (the mean sea level during laminated intervals) of the original data. You can calculate this in many different ways, but here we demonstrate the calculation using indexing.

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

Finally, 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).

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

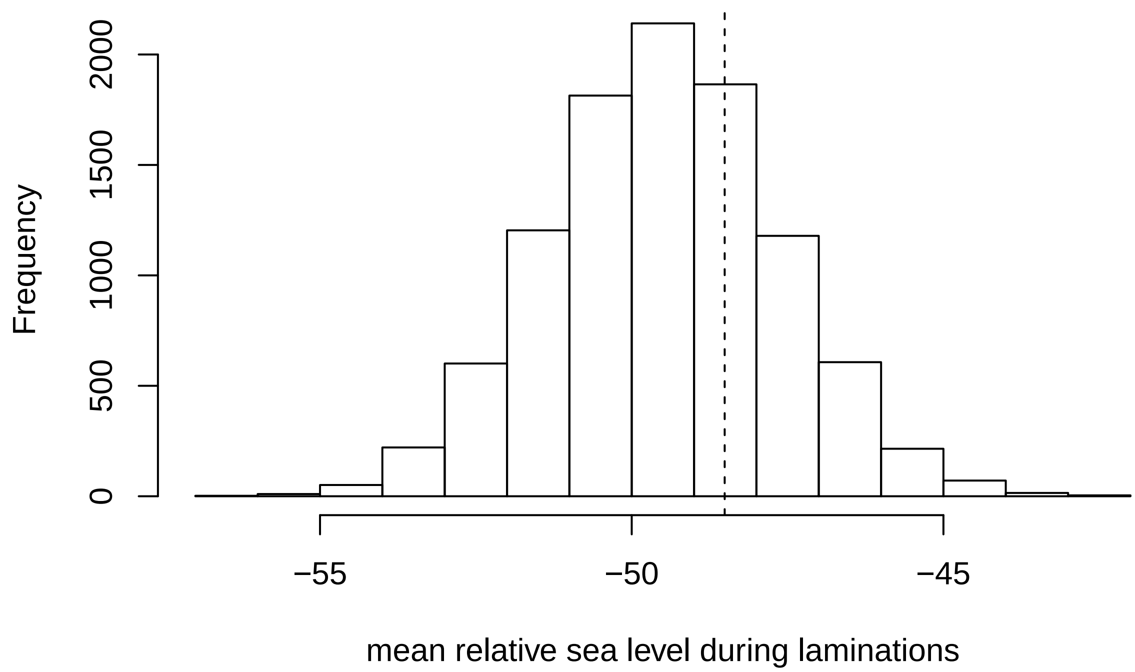
```
## [1] 0.2953705
```

Because the sea level's lamination data are simulated, the p-value from this test is quite large and we do not reject our null hypothesis. According to this data set, the mean sea level is not significantly higher during laminations than it would be if sea levels and laminations were truly unrelated.

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.

```r
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.