Why the mean is the same but the median is different

May 17, 2015

1 Introduction

In the first peer-reviewed assignment for Reproducible Research there is a problem where some data is missing. We were supposed to compute the mean and the median of the existing data, then impute the missing values and recompute the mean and the median.

The data was in columns labeled *steps*, *date*, and *interval*. The *date* is a date. The *interval* is a five minute interval, and the *steps* are the number of steps taken during that five minute interval on that date.

R can be used to confirm that there are 8 dates with missing step values, and that for those 8 dates, **all** the step values are missing. Let's call these dates the "missing data" days.

To impute the missing values, the *steps* for an *interval* were set equal to the mean of the steps on that interval, where the mean was taken for that interval over all days that had data.

For the sake of consistent terminology, let's define a few things.

n	number of days that originally had data
μ	mean total number of steps, taken over the original n days
m	median total number of steps, taken over the original n days

The values μ and m can be computed in R. In the following explanations we will use d to denote the day and i to denote the interval. The index d ranges from 1 to n, and the index i ranges from 1 to 288 (the number of 5-minute intervals in a day).

After filling in missing values and recomputing the mean and median, the mean was the same as it was before missing values were filled in, while the median changed slightly. Why does this make sense?

2 Why the mean stayed the same

The mean total number of steps per day, μ, is given by the formula

$$\mu = \frac{1}{n}\left(\sum_{d=1}^{n}\sum_{i=1}^{288} x_i^d\right)$$

where n is the number of days that originally had data and x_i^d is the number of steps taken on day d during interval i. The inner summation adds up the steps for each interval during day d to find the total number of steps taken on day d. The outer summation adds up the total number of steps on each day, over all days. Finally we divide by $\frac{1}{n}$ to find the mean total steps taken per day.

Since the summations are finite, we can swap them and distribute the coefficient $\frac{1}{n}$ to get the equivalent formula

$$\mu = \sum_{i=1}^{288}\left(\frac{1}{n}\sum_{d=1}^{n} x_i^d\right).$$

This is still the mean total number of steps per day. However, now the inner sum is taken over all days while interval i is held fixed. The multiplication by $\frac{1}{n}$ gives the mean number of steps taken during interval i over all days. The outer sum takes the total, over all intervals, of the mean number of steps per interval.

For the missing data days, we filled in the missing data for each interval by setting it equal to the mean number of steps taken on that interval (where the mean was taken over all the days that originally had data). Since **all** the data was missing on the missing data days, the number of steps for each interval i was filled in with

$$\left(\frac{1}{n}\sum_{d=1}^{n} x_i^d\right).$$

When we take the sum over all intervals, we find that the total number of steps taken on a missing data day will now come out to be

$$\sum_{i=1}^{288}\left(\frac{1}{n}\sum_{d=1}^{n} x_i^d\right).$$

As shown above, this is exactly the mean total steps we computed *before* filling in the missing data.

To summarize: the total number of steps taken on a missing data day will now come out to be the same as the mean number of total steps per day we calculated when we were ignoring missing data.

R can be used to confirm that each of the missing data days has the same total number of steps after the missing data is filled in, and that this number is equal to the original mean μ.

The only thing left is to show that if we have a collection of n numbers with mean μ and we add to that collection 8 copies of μ, the mean won't change.

Call the n numbers $t_1, \ldots t_n$. Then by definition

$$\frac{1}{n}\sum_{i=1}^{n} t_i = \mu$$

so

$$\sum_{i=1}^{n} t_i = \mu n.$$

If we toss in 8 copies of μ and take the mean of the new collection, we are now finding the mean of $n+8$ numbers and we get

$$\frac{1}{n+8}\left(\sum_{i=1}^{n} t_i + 8\mu\right) = \frac{1}{n+8}\left(n\mu + 8\mu\right)$$

which simplifies to

$$\frac{1}{n+8} \cdot \mu(n+8) = \mu.$$

Thus, the mean does not change when we recompute it for all days with the missing data filled in via the method described above.

3 Why the median changed

Think of the data for total number of steps per day as a list of 53 numbers

$$t_1, t_2, \ldots, t_{52}, t_{53}.$$

The median is the "middle" number. There were originally 53 days with data, and so if we sort the data the median is position 27. We're also interested in the numbers to either side of the median.

$$t_1, \ldots, t_{26}, t_{27} = m, t_{28}, \ldots, t_{53}.$$

The interesting part is that, as one can check with R,

$$m < \mu < t_{28}.$$

When we fill in the missing data days, we're adding 8 copies of μ to the list of total steps taken per day. In the new sorted list, these will occur after the original median m but before any of the other values. Originally we had

$$t_1, \ldots, t_{26}, m, t_{28}, \ldots, t_{53}.$$

After replacing missing values the sorted list of total steps values is

$$t_1, \ldots, t_{26}, m, \mu, \mu, \mu, \mu, \mu, \mu, \mu, \mu, t_{28}, \ldots, t_{53}.$$

This new list is $53+8 = 61$ elements long. The new median, or "middle number" is now the 31st element, shown in parentheses:

$$t_1, \ldots, t_{26}, m, \mu, \mu, \mu, (\mu), \mu, \mu, \mu, \mu, t_{28}, \ldots, t_{53}.$$

Thus it makes sense that the median changed.