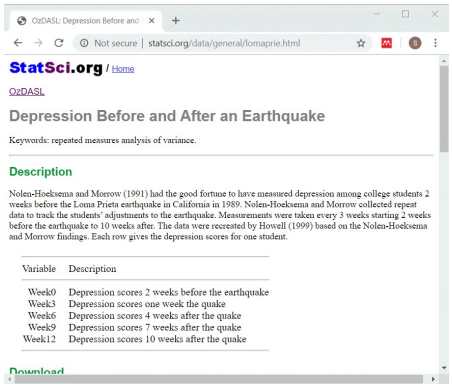
Gathering and spreading

Suman Sahil, Steve Simon

Creation date: 2019-11-20

Description of lomaprie data set



OzDASL: Depression Before and

Not secure | statsci.org/data/general/lomaprie.html

StatSci.org / Home

OzDASL

Depression Before and After an Earthquake

Keywords: repeated measures analysis of variance.

Description

Nolen-Hoeksema and Morrow (1991) had the good fortune to have measured depression among college students 2 weeks before the Loma Prieta earthquake in California in 1989. Nolen-Hoeksema and Morrow collected repeat data to track the students' adjustments to the earthquake. Measurements were taken every 3 weeks starting 2 weeks before the earthquake to 10 weeks after. The data were recreated by Howell (1999) based on the Nolen-Hoeksema and Morrow findings. Each row gives the depression scores for one student.

Variable	Description
Week0	Depression scores 2 weeks before the earthquake
Week3	Depression scores one week the quake
Week6	Depression scores 4 weeks after the quake
Week9	Depression scores 7 weeks after the quake
Week12	Depression scores 10 weeks after the quake

Download

Description of data

Speaker notes:

This data set displays depression scores among college students in a study planned for other purposes but which ended up with the baseline measurements collected just before a major California earthquake. The researchers seized the opportunity to assess changes in depression immediately before and after the earchquake as well as the recovery process longer term.

The source for this data set is

http://www.statsci.org/data/general/lomaprie.html

lomaprie_db, listing of original table

– SQL code

```
select *
  from lomaprie
  limit 5

select count(*) as n_records
   from lomaprie
```

lomaprie_db, listing of original table

– Output

```
##   id Week0 Week3 Week6 Week9 Week12
## 1  1     6    10     8     4      6
## 2  2     2     4     8     5      6
## 3  3     2     4     8     5      6
## 4  4     4     5     8    10      7
## 5  5     4     7     9     7     12
##   n_records
## 1        25
```

Speaker notes:

This is what the original data looks like. Notice that there is no primary key in this table. Normally this is a major flaw in any reasonable database, but let's ignore that for now.

Gathering into a single column

– SQL code

```
select
  id, week0 as depression, 0 as time
  from lomaprie
union select
  id, week3 as depression, 3 as time
  from lomaprie
union select
  id, week6 as depression, 6 as time
  from lomaprie
```

Gathering

– SQL code, continued

```
union select
  id, week9 as depression, 9 as time
  from lomaprie
union
  select id, week12 as depression, 12 as time
  from lomaprie
limit 13
```

Speaker notes:

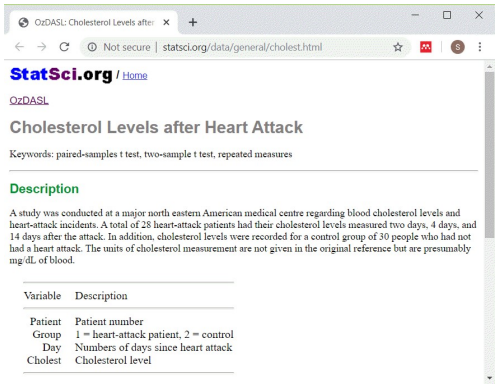
Gathering

– Output

```
##    id depression time
## 1   1          4    9
## 2   1          6    0
## 3   1          6   12
## 4   1          8    6
## 5   1         10    3
## 6   2          2    0
## 7   2          4    3
## 8   2          5    9
## 9   2          6   12
## 10  2          8    6
## 11  3          2    0
## 12  3          4    3
## 13  3          5    9
```

Speaker notes:

Description of cholestg data set



OzDASL: Cholesterol Levels after

Not secure | statsci.org/data/general/cholest.html

StatSci.org / Home

OzDASL

Cholesterol Levels after Heart Attack

Keywords: paired-samples t test, two-sample t test, repeated measures

Description

A study was conducted at a major north eastern American medical centre regarding blood cholesterol levels and heart-attack incidents. A total of 28 heart-attack patients had their cholesterol levels measured two days, 4 days, and 14 days after the attack. In addition, cholesterol levels were recorded for a control group of 30 people who had not had a heart attack. The units of cholesterol measurement are not given in the original reference but are presumably mg/dL of blood.

Variable	Description
Patient	Patient number
Group	1 = heart-attack patient, 2 = control
Day	Numbers of days since heart attack
Cholest	Cholesterol level

Description of data

Speaker notes:

This data set shows cholesterol levels for patients after a heart attack. I simplified the data by removing the control patients.

The source for this data set is

STatSci.org

cholesterol, listing of original table

– SQL code

```
select *
  from cholesterol
  order by patient, day
  limit 10


select count(*) as n_records
   from cholesterol
```

cholestg_1_db, listing of original table

– Output

```
##    patient day cholest
## 1        1   2     270
## 2        1   4     218
## 3        1  14     156
## 4        2   2     236
## 5        2   4     234
## 6        3   2     210
## 7        3   4     214
## 8        3  14     242
## 9        4   2     142
## 10       4   4     116
##   n_records
## 1        75
```

Speaker notes:

This is what the original data looks like.

cholestg_1_db, count of days

– SQL code

```
select day, count(day) as n_days
  from cholesterol
  group by day
```

– Output

```
##   day n_days
## 1   2     28
## 2   4     28
## 3  14     19
```

Spreading across multiple columns

– SQL code

```
select d2.patient, d2.cholest as chol02,
  d14.cholest as chol14
  from cholesterol as d2
  left join cholesterol as d14
  on d2.patient=d14.patient
  where d2.day=2 and d14.day=14
  limit 10
```

Spreading across multiple columns

– Output

```
##    patient chol02 chol14
## 1        1    270    156
## 2        3    210    242
## 3        6    272    256
## 4        7    160    142
## 5        8    220    216
## 6        9    226    248
## 7       11    186    168
## 8       12    266    236
## 9       14    318    200
## 10      15    294    264
```

Spreading across multiple columns, Take 2

– SQL code

```
select d2.patient, d2.cholest as chol02,
  d14.cholest as chol14
  from cholesterol as d2
  left join cholesterol as d14
  on d2.patient=d14.patient and
    d2.day=2 and d14.day=14
  limit 10
```

Spreading across multiple columns, Take 2

– Output

```
##    patient chol02 chol14
## 1        1    270    156
## 2        2    236     NA
## 3        3    210    242
## 4        4    142     NA
## 5        5    280     NA
## 6        6    272    256
## 7        7    160    142
## 8        8    220    216
## 9        9    226    248
## 10      10    242     NA
```

Entity-Attribute-Value format

```
##    entity attribute  value
## 1       1       age  9.000
## 2       1       fev  1.708
## 3       1        ht 57.000
## 4       1       sex  0.000
## 5       1     smoke  0.000
## 6       2       age  8.000
## 7       2       fev  1.724
## 8       2        ht 67.500
## 9       2       sex  0.000
## 10      2     smoke  0.000
## 11      3       age  7.000
## 12      3       fev  1.720
```

Advantages of Entity-Attribute-Value format

- Universal format
- Easy to add fields
- Ideal for sparse matrices
  - Sparse: most entries are zero

Spreading the EAV format

– SQL code

```
select
  d1.entity,
  d1.value as age,
  d2.value as fev,
  d3.value as ht,
  d4.value as sex,
  d5.value as smoke
  from eav as d1
inner join eav as d2
  on d1.attribute='age' and
     d2.attribute='fev' and
     d1.entity=d2.entity
```

Spreading the EAV format

– SQL code, continued

```
inner join eav as d3
  on d3.attribute='ht' and
     d1.entity=d3.entity
inner join eav as d4
  on d4.attribute='sex' and
     d1.entity=d4.entity
inner join eav as d5
  on d5.attribute='sex' and
     d1.entity=d5.entity
  limit 10
```

Speaker notes:

Spreading the EAV format

– Output

```
##    entity age   fev   ht sex smoke
## 1       1   9 1.708 57.0   0     0
## 2       2   8 1.724 67.5   0     0
## 3       3   7 1.720 54.5   0     0
## 4       4   9 1.558 53.0   1     1
## 5       5   9 1.895 57.0   1     1
## 6       6   8 2.336 61.0   0     0
## 7       7   6 1.919 58.0   0     0
## 8       8   6 1.415 56.0   0     0
## 9       9   8 1.987 58.5   0     0
## 10     10   9 1.942 60.0   0     0
```

Speaker notes:

Spreading the EAV format

– SQL code

```
select
  entity, value as age
  from eav
  where entity in (
    select entity from eav
    where attribute='smoke' and value=1
  ) and attribute='age'
limit 10
```

Spreading the EAV format

– Output

```
##    entity age
## 1     191   9
## 2     332  14
## 3     358  14
## 4     366  13
## 5     369  11
## 6     370  14
## 7     372  13
## 8     381  12
## 9     384  14
## 10    388  10
```

Speaker notes: