Introduction to data management and cleaning

Dr. Juliet Pulliam
Department of Biology and
Emerging Pathogens Institute
University of Florida

Clinic on the Meaningful Modeling of Epidemiological Data
African Institute for the Mathematical Sciences
Muizenberg, South Africa
13653273600

Levels of data aggregation

Aggregated data

De-identified data

Personally identifying data } Individual-level

Read in the dataset

```
# cases.R
# JRCP 06 Jan 2010
#
# Last modified: 09 June 2015

rm(list=ls())

require(foreign)

cases <- read.spss('cases.sav',to.data.frame=T,trim.factor.names=T)

dim(cases)

[1] 135  91

names(cases)

[1] 'idno'  …   'sex'   …  'age'  'occup'  'dor'  'dtadm'  'dtill'
…  'fever'  'dtfever'  …  'cough'  'dtcough'  …  'headache'
… 'dthead'  … 'outcome'  …  'dtdeath'  … 'igm1'  'igg1'
'pcr1ser'     'dtserum1'    'Igm2'        'Igg2'        'dtserum2'
… 'virusiso' … 'CXR'   'ARDS'  …
```

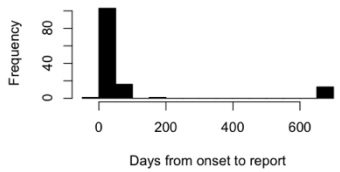
Look at the dataset - format

```
head(cases$dor)

[1] 13266288000 13266288000 13266374400

cases$dor <- as.Date('1582-10-14') + cases$dor/60/60/24

range(cases$dor,na.rm=T)

[1] "2003-03-06"  "2008-03-15"
```

Categorical data - factor (levels)
Continuous data - numeric (range)
Dates - Date, POSIXct (range)
Binary data - T, F

Look at the dataset - logic

```
range(cases$dor-cases$dtill)
[1] NA NA
range(cases$dor-cases$dtill,na.rm=T)
[1]  -7 686
subset(cases,dor-dtill<0)$idno
AT567
```




```
hist(cases$dor-cases$dtill,20,xlab='Days from onset to report',main='',col=1)
```

Inconsistencies

IDNO: SF999

dtill given as 16 Feb 2005
dtfever given as 16 Feb 2003

Correction: Change year of dtfever from 2003 to 2005

IDNO: GW321, Year: 2003

dtill given as 12 Jan 2003
dtcough given as 12 Mar 2003
dthead given as 12 Mar 2003
dtdeath given as 15 Jan 2003

Correction: Change month of dtcough and dthead from 12 Mar to 12 Jan

Inconsistencies

IDNO: LP309, Year: 2002

dtill given as 12 Aug 2002
dthead given as 11 Aug 2002

Correction: none

IDNO: AT567, Year: 2005

dor given as 6 May 2005
dtill given as 14 May 2005

Correction: delete case (sample negative when tested at CDC)

Correct the dataset

```
# DATABASE CORRECTIONS #

cases$dtfever[cases$idno=='SF999'] <- as.Date('2005-01-16') # Onset 2005; dtfever was 2003

cases$dthead[cases$idno=='GW321'] <- as.Date('2003-01-12') # Onset/death Jan; dtfever was March
cases$dtcough[cases$idno=='GW321'] <- as.Date('2003-01-12') # Onset/death Jan; dtcough was March

cases <- subset(cases,idno!='AT567') # sample negative when tested at CDC
```

Save corrected dataset as new file

```
now <- format(Sys.time(),'%d%b%Y_%H%M')

fn <- paste('casesClean',now,'.Rdata',sep='')
fn

[1] 'casesClean07Jan2010_2048.Rdata'

save(cases,now,file=fn)
```