STA286 Lecture 01

Neil Montgomery

Last edited: 2017-01-11 09:39

admin

contact, notes

date format	YYYY-MM-DD – *All Hail ISO8601!!!*
instructor	Neil Montgomery
email	`neilm@mie.utoronto.ca`
office	BA8137
office hours	W11-1
website	portal (announcements, grades, suggested exercises, etc.)
github	`https://github.com/sta286-winter-2017` (lecture material, code, etc.)

Lecture notes and other course timing matters will be organized by *lecture number* and not *lecture date*, due to two lecture sections.

evaluation, book, tutorials

what	when	how much
midterm 1	2017-02-13	20%
midterm 2	2017-03-27	30%
exam	TBA	50%

The book is Walpole, R.E., Myers, R.H., Myers, S.L., Ye, K., 2012. *Probability & statistics for engineers & scientists.* 9th edition.

I will suggest exercises from this book each week. Your TA will work through some of them in tutorial each week.

Tutorials start TBA.

software

The course begins and ends with data analysis, with a long stretch of probability theory in the middle.

Data analysis requires a computer. Also, some concepts can be illustrated using simulation, which also requires a computer.

We will be using `R`. It's pretty good at data analysis.

language	interpreter	integrated development environment
`R`	`R`	`RStudio`

Some detailed instructions and suggestions for installation and configuration appear on the course website.

I will try to impart some data analysis workflow wisdom throughout the course. Some already appears in the detailed instructions.

MATLAB SUCKS!

what is a dataset?

most datasets are rectangles

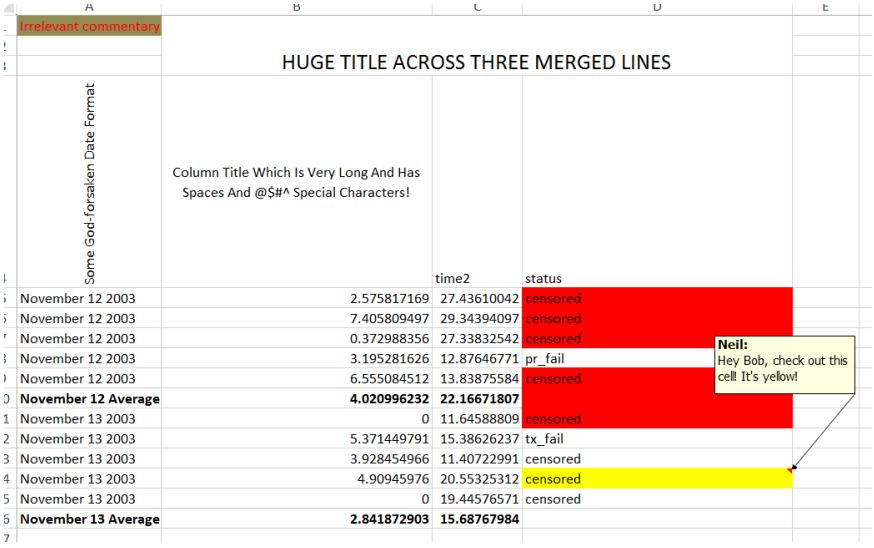
Columns are the *variables*.

The top row has the names of the variables; possibly chosen wisely.

Rows are the *observations* of measurements taken on *units*.

There are no averages, no comments (unless in a "comment" variable), no colors, no formatting, no plots, no capes!

not a dataset



	A	B	C	D	E
1	Irrelevant commentary				
2		HUGE TITLE ACROSS THREE MERGED LINES			
3					
4	Some God-forsaken Date Format	Column Title Which Is Very Long And Has Spaces And @$#^ Special Characters!	time2	status	
5	November 12 2003	2.575817169	27.43610042	censored	
6	November 12 2003	7.405809497	29.34394097	censored	
7	November 12 2003	0.372988356	27.33832542	censored	
8	November 12 2003	3.195281626	12.87646771	pr_fail	
9	November 12 2003	6.555084512	13.83875584	censored	
10	**November 12 Average**	**4.020996232**	**22.16671807**		
11	November 13 2003	0	11.64588809	censored	
12	November 13 2003	5.371449791	15.38626237	tx_fail	
13	November 13 2003	3.928454966	11.40722991	censored	
14	November 13 2003	4.90945976	20.55325312	censored	
15	November 13 2003	0	19.44576571	censored	
16	**November 13 Average**	**2.841872903**	**15.68767984**		
17					

Neil:
Hey Bob, check out this cell! It's yellow!

not a dataset

ASSETNUM	MOVEDATE_1	FROM_LOCATION1	TO_LOCATION1	MOVEDATE_2	FROM_LOCATION2	TO_LOCATION2	MOVEDATE_3	FRC
0201011	2005-12-16	NO_LOCATION	RSREPAIR					
0209679	2006-01-16	NO_LOCATION	RSREPAIR	2006-01-30	RSREPAIR	DN4VNCR	2014-02-14	DN4
0209680	2005-05-17	NO_LOCATION	RSREPAIR	2005-08-03	RSREPAIR	WY172UCR	2013-11-08	WY
0209709	2005-05-20	NO_LOCATION	WY92WEPR	2011-10-07	WY92WEPR	RSREPAIR	2013-11-08	RSR
0209711	2011-10-07	WY91WEPR	RSREPAIR	2013-11-08	RSREPAIR	WY174VNCR		
0209714	2003-12-15	NO_LOCATION	RSREPAIR					
0209720	2011-10-07	WY95WEPR	RSREPAIR	2013-06-25	RSREPAIR	WY70ASPR		
0209722	2011-10-07	WY106WEPR	RSREPAIR	2013-06-27	RSREPAIR	WY144BSUSR		
0209728	2011-10-07	WY94WEPR	RSREPAIR	2013-11-08	RSREPAIR	WY143NWCPR		
0209729	2006-01-16	NO_LOCATION	RSREPAIR	2006-01-30	RSREPAIR	DN12ASRA	2014-04-04	DN1
0209737	2005-01-11	NO_LOCATION	DN15NWCRB	2006-03-21	DN15NWCRB	RSREPAIR	2006-03-31	RSR
0209739	2011-10-07	WY144WEPR	RSREPAIR	2013-12-09	RSREPAIR	WY178TPR		
0209740	2011-10-07	WY143WEPR	RSREPAIR	2012-09-12	RSREPAIR	DNSPARE	2014-05-30	DN9
0209741	2006-01-16	NO_LOCATION	RSREPAIR	2006-01-30	RSREPAIR	DN10BHR	2014-09-05	DN1

an oil readings dataset (wide version)

```
## # A tibble: 612 × 17
##     Ident                Date WorkingAge   TakenBy    Fe    Al    Cu
##     <chr>               <dttm>      <dbl>     <chr> <dbl> <dbl> <dbl>
## 1  448576 1999-05-10 19:00:00        243 EMPL_0917    13     5    14
## 2  448576 1999-07-26 19:00:00        569 EMPL_0917    18     6    25
## 3  448576 1999-09-29 19:00:00        830 EMPL_9375    26     6    35
## 4  448576 1999-10-08 19:00:00        862 EMPL_0917    15     9    14
## 5  448576 1999-11-02 19:00:00        946 EMPL_9375    14     4    19
## 6  448576 1999-12-09 19:00:00       1088 EMPL_0917    18     5    23
## 7  448576 1999-12-27 19:00:00       1157 EMPL_9375    24     8    25
## 8  448576 2000-01-14 19:00:00       1238 EMPL_9375    27     9    34
## 9  448576 2000-02-15 19:00:00       1376 EMPL_9375    16     8    17
## 10 448576 2000-03-11 19:00:00       1492 EMPL_0917    20     8    20
## # ... with 602 more rows, and 10 more variables: Cr <dbl>, Si <dbl>,
## #   Pb <dbl>, Ph <dbl>, Ca <dbl>, Zn <dbl>, Mg <dbl>, Mo <dbl>,
## #   Sn <dbl>, Na <dbl>
```

oil readings with Ident and TakenBy properly treated

```
## # A tibble: 612 × 17
##     Ident                Date WorkingAge   TakenBy    Fe    Al    Cu
##    <fctr>              <dttm>      <dbl>    <fctr> <dbl> <dbl> <dbl>
## 1  448576 1999-05-10 19:00:00        243 EMPL_0917    13     5    14
## 2  448576 1999-07-26 19:00:00        569 EMPL_0917    18     6    25
## 3  448576 1999-09-29 19:00:00        830 EMPL_9375    26     6    35
## 4  448576 1999-10-08 19:00:00        862 EMPL_0917    15     9    14
## 5  448576 1999-11-02 19:00:00        946 EMPL_9375    14     4    19
## 6  448576 1999-12-09 19:00:00       1088 EMPL_0917    18     5    23
## 7  448576 1999-12-27 19:00:00       1157 EMPL_9375    24     8    25
## 8  448576 2000-01-14 19:00:00       1238 EMPL_9375    27     9    34
## 9  448576 2000-02-15 19:00:00       1376 EMPL_9375    16     8    17
## 10 448576 2000-03-11 19:00:00       1492 EMPL_0917    20     8    20
## # ... with 602 more rows, and 10 more variables: Cr <dbl>, Si <dbl>,
## #   Pb <dbl>, Ph <dbl>, Ca <dbl>, Zn <dbl>, Mg <dbl>, Mo <dbl>,
## #   Sn <dbl>, Na <dbl>
```

oil readings dataset (long version)

```
## # A tibble: 7,956 × 6
##     Ident                Date WorkingAge   TakenBy element   ppm
##    <fctr>              <dttm>      <dbl>    <fctr>   <chr> <dbl>
## 1  448576 1999-05-10 19:00:00        243 EMPL_0917      Fe    13
## 2  448576 1999-07-26 19:00:00        569 EMPL_0917      Fe    18
## 3  448576 1999-09-29 19:00:00        830 EMPL_9375      Fe    26
## 4  448576 1999-10-08 19:00:00        862 EMPL_0917      Fe    15
## 5  448576 1999-11-02 19:00:00        946 EMPL_9375      Fe    14
## 6  448576 1999-12-09 19:00:00       1088 EMPL_0917      Fe    18
## 7  448576 1999-12-27 19:00:00       1157 EMPL_9375      Fe    24
## 8  448576 2000-01-14 19:00:00       1238 EMPL_9375      Fe    27
## 9  448576 2000-02-15 19:00:00       1376 EMPL_9375      Fe    16
## 10 448576 2000-03-11 19:00:00       1492 EMPL_0917      Fe    20
## # ... with 7,946 more rows
```

the main questions

- where did the data come from?

the main questions

- where did the data come from?
  - were the units chosed randomly from a population?

the main questions

- where did the data come from?
  - were the units chosed randomly from a population?
  - were the units randomly assigned into groups?

the main questions

- where did the data come from?
  - were the units chosed randomly from a population?
  - were the units randomly assigned into groups?
- what are the (joint) *distributions* of the data?

random sample, experiment, observational data

Sometimes the data come from a *random sample* from a larger *population*, in which case statements about the sample can apply to the population using laws of probability.

random sample, experiment, observational data

Sometimes the data come from a *random sample* from a larger *population*, in which case statements about the sample can apply to the population using laws of probability.

(Not a focus of this course.)

random sample, experiment, observational data

Sometimes the data come from a *random sample* from a larger *population*, in which case statements about the sample can apply to the population using laws of probability.

(Not a focus of this course.)

Sometimes data come from an *experiment* where units are randomly assigned to different *levels* of one or more *factors*, in which cause cause-and-effect can be inferred using laws of probability.

random sample, experiment, observational data

Sometimes the data come from a *random sample* from a larger *population*, in which case statements about the sample can apply to the population using laws of probability.

(Not a focus of this course.)

Sometimes data come from an *experiment* where units are randomly assigned to different *levels* of one or more *factors*, in which cause cause-and-effect can be inferred using laws of probability.

Often the data are just some records of what happened. Grander inferences might be made, but only on a subject-matter basis.

distribution (informally)

- A *distribution* is a

distribution (informally)

- A *distribution* is a
  - Complete description of...

distribution (informally)

- A *distribution* is a
  - Complete description of...
  - ...the possible values of one or more variables...

distribution (informally)

- A *distribution* is a
  - Complete description of...
  - ...the possible values of one or more variables...
  - ...and the relative frequency of those values.

distribution (informally)

- A *distribution* is a
  - Complete description of...
  - ...the possible values of one or more variables...
  - ...and the relative frequency of those values.
- A dataset contains **empirical** information about distribution(s) that can be assessed

distribution (informally)

- A *distribution* is a
  - Complete description of...
  - ...the possible values of one or more variables...
  - ...and the relative frequency of those values.
- A dataset contains **empirical** information about distribution(s) that can be assessed
  - numerically

distribution (informally)

- A *distribution* is a
  - Complete description of...
  - ...the possible values of one or more variables...
  - ...and the relative frequency of those values.
- A dataset contains **empirical** information about distribution(s) that can be assessed
  - numerically
  - graphically

distribution (informally)

- A *distribution* is a
  - Complete description of...
  - ...the possible values of one or more variables...
  - ...and the relative frequency of those values.
- A dataset contains **empirical** information about distribution(s) that can be assessed
  - numerically
  - graphically

distribution (informally)

- A *distribution* is a
  - Complete description of...
  - ...the possible values of one or more variables...
  - ...and the relative frequency of those values.
- A dataset contains **empirical** information about distribution(s) that can be assessed
  - numerically
  - graphically

through a process called *exploratory data analysis*

a taxonomy of variables

- Numerical or categorical?

a taxonomy of variables

- Numerical or categorical?
  - Numerical: length, ppm, time-to-event, etc.

a taxonomy of variables

- Numerical or categorical?
  - Numerical: length, ppm, time-to-event, etc.
  - Categorical: yes/no, colour, etc.

a taxonomy of variables

- Numerical or categorical?
  - Numerical: length, ppm, time-to-event, etc.
  - Categorical: yes/no, colour, etc.
  - Lots of grey areas even in this classification!

a taxonomy of variables

- Numerical or categorical?
  - Numerical: length, ppm, time-to-event, etc.
  - Categorical: yes/no, colour, etc.
  - Lots of grey areas even in this classification!
    - Categories can have an inherent order

a taxonomy of variables

- Numerical or categorical?
  - Numerical: length, ppm, time-to-event, etc.
  - Categorical: yes/no, colour, etc.
  - Lots of grey areas even in this classification!
    - Categories can have an inherent order
    - "Likert scale" (strongly disagree coded as 1 and so on...)

a taxonomy of variables

- Numerical or categorical?
  - Numerical: length, ppm, time-to-event, etc.
  - Categorical: yes/no, colour, etc.
  - Lots of grey areas even in this classification!
    - Categories can have an inherent order
    - "Likert scale" (strongly disagree coded as 1 and so on...)
- Numerical variables could be discrete (counting something) or continuously measured.