

# TG3: Setting the stage with beginning data analyses

Marianne Huebner, Saskia Le Cessie,  
Werner Vach, Maria Blettner,  
Danielle Bodicoat

*“The initial examination of data is a valuable state of most statistical investigations, not only for scrutinizing and summarizing data, but also for model formulations.”*

*--Chatfield. JRSSA 1985*

*“In practice one has only to look at the literature to see that the methods are still generally undervalued, often neglected, and sometimes actively regarded with disfavor.”*

*--Chatfield. JRSSA 1985*

# It's a topic of interest:

[Workflow for statistical analysis and report writing](#)

viewed **17020** times

[How to efficiently manage a statistical analysis project?](#)

viewed **7159** times

[How do you combine “Revision Control” with “Workflow” for R?](#)

viewed **3074** times

# It takes time

*80% of data analysis is spent on the process of cleaning and preparing the data.*

Dasu and Johnson 2003

# It's time well spent

*Even with best intentions during data collection:  
data integrity checks find error rates 2-5% in the  
“best” datasets*

Feedback from practicing statisticians  
from various institutions

# Spreadsheets can be problematic

ID	Sex	Date of Surgery	Height (cm)	Weight (kg)	Diagnosis
1	male	1/1/2011	163	68	1
2	M	15/1/99	167	80	2,1
3	F	2/1/09	166	unknown	2
4	M	2/15/11	172cm	82	2
4		8/19/12		85	2
5	MALE	March 1, 2013	180	67	2
6	m	3/15/2008	164	62	2 (dx 5/2/11)
7	m	4-1-2013	165 ???	66	1
8	female	April, 2005	166	n.a.	1
9	F	2007-01-25	62	65kg	diabetes
			Average=166		

# Spreadsheet – corrected

id	sex	datesurgery	height	weight	diagnosis1	diagnosis2
1	male	2011-01-01	163	68	1	
2	male	1999-01-15	167	80	2	1
3	female	2009-01-02	166	NA	2	
4	male	2011-02-15	172	82	2	
4	male	2012-08-19	172	85	2	
5	male	2013-03-01	180	67	2	
6	male	2008-03-15	164	62	2	
7	male	2013-04-01	165	66	1	
8	female	2005-04-15	166	NA	1	
9	female	2007-01-25	162	65	3	



# Structuring datasets

1. Each variable forms a column.
2. Each observation forms a row.

Things go wrong when:

- column headers are values, not variable names
- multiple variables are stored in one column
- variables are stored in both rows and columns
- a subject is stored in multiple tables

***“Despite the amount of time it takes, there has been surprisingly little research on how to clean data well. Part of the challenge is the breadth of activities it encompasses: from outlier checking, to date parsing, to missing value imputation.”***

***H. Wickham, Tidy Data 2014***

# It's not enough

*“Data preparation is not just a first step, but must be repeated many times over the course of analysis as new problems come to light or new data is collected. “*

H. Wickham

→ Other Topic Groups

# Data quality

- Do the date sequences make sense (birth before surgery)?
- Are data consistent between variables? (date of surgery and date of discharge vs length of stay)
- What is the proportion of missing values for each variable (e.g. Echocardiogram, 30% missing at one month follow-up, 70% missing at one year follow-up)
- What is meant by time frames of follow-up, e.g “one month”, “one year”?

# RedCap data checks

- Missing values
- Missing values required fields only
- Field validation (incorrect data type)
- Field validation (out of range)
- Outliers for numerical fields



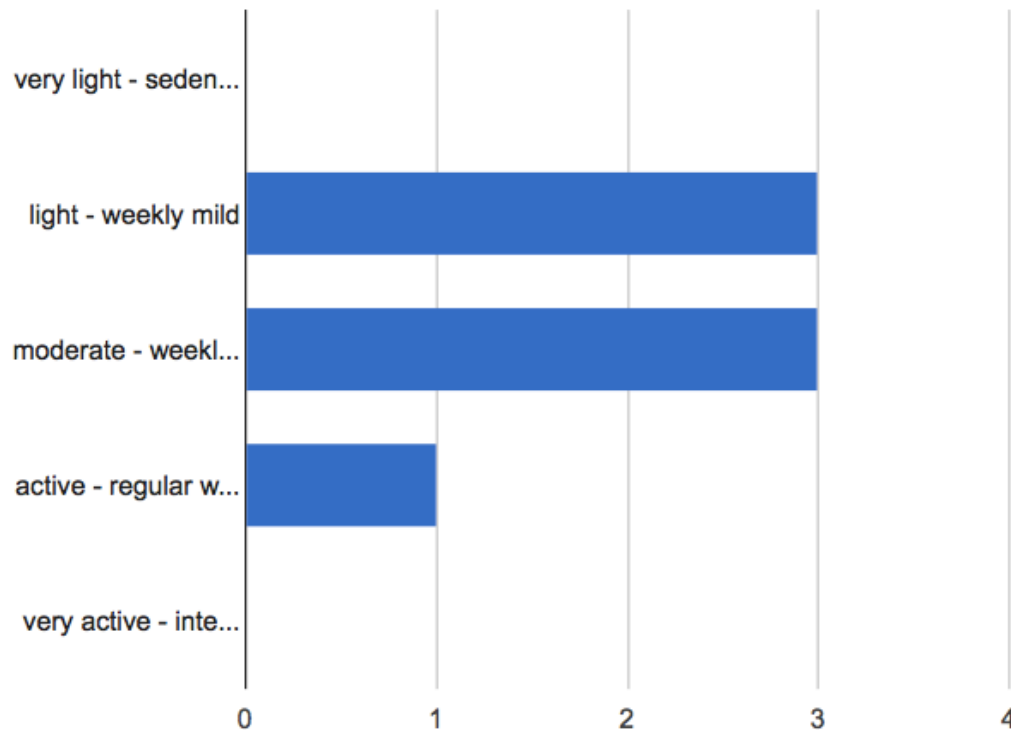
The REDCap Consortium is composed of 1,106 active institutional partners from CTSA, GCRC, RCMI and other institutions in 83 countries.

# REDCap data summaries

What is your activity level?: [Refresh Plot](#) | [View as Bar Chart](#)

Total (N)	Missing	Unique
7	0 (0%)	3

**Counts/frequency:** very light - sedentary (0, 0%), light - weekly mild (3, 42.9%), moderate - weekly intense (3, 42.9%), active - regular workouts (1, 14.3%), very active - intense sports (0, 0%)



# Reproducible research



Reinhart, Rogoff: Growth in a time of debt. 2010

Herndon, Ash, Pollin: A critique of Reinhart and Rogoff.  
2013

- Selected exclusion of years/countries
- Unconventional weighting
- Coding error (averaging of wrong cells)
- Averaging a variable with missing data.

# R markdown: data, code, report

## Inference for means (t-interval or t-test)

The airflow rate, FEV1, is the ratio of a person's forced expiratory volume to the vital capacity, VC (max. volume of air a person can exhale after taking a deep breath). If the enzyme has an effect, it will be to reduce the FEV1/VC ratio. The norm is 0.80 in persons with no lung dysfunction.

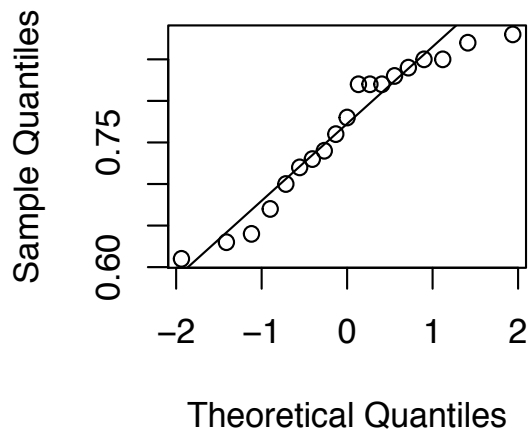
```
ratio <- c(0.61, 0.7, 0.63, 0.76, 0.67, 0.72, 0.64, 0.82, 0.88, 0.82, 0.78,  
          0.84, 0.83, 0.82, 0.74, 0.85, 0.73, 0.85, 0.87)
```

Summary statistics

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	0.610	0.710	0.780	0.766	0.835	0.880

Are the data symmetric or approximately normal?

### Normal Q-Q Plot





# Report content

- Statistical report is more extensive than what will be in the manuscript
- Read in raw data
- Steps of processing data
- Numerical data summaries
- Graphical explorations, e.g density plots, boxplots, plots over time, plots of association of variables, overlaid density plots from different categories

-> Feedback from you?

# Reproducible research

- **Data:** raw, processed
- **Figures:** exploratory, final
- **Code:** raw script, final script
- **Text:** readme files, documents, markdown/  
knitr/sweave file
  
- Making data and code available: **Markdown, Knitr, Sweave, Github**

# Baseline characteristics

- Data summaries for each variable and/or group
  - Location measures
  - Small or large variation
  - Conceptually or statistically motivated groupings
  - Zero inflated
  - Missingness
- Explore missing data
  - Table with number of missing for each variable
  - Comparing missing and non-missing cases
  - Always assume missingness hides a meaningful value for analysis (R. Little, T Raghunathan)

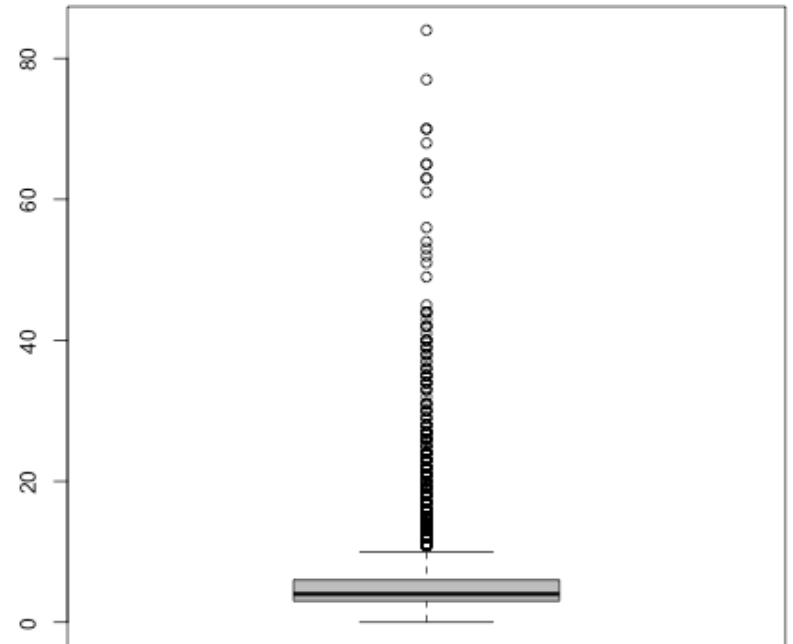
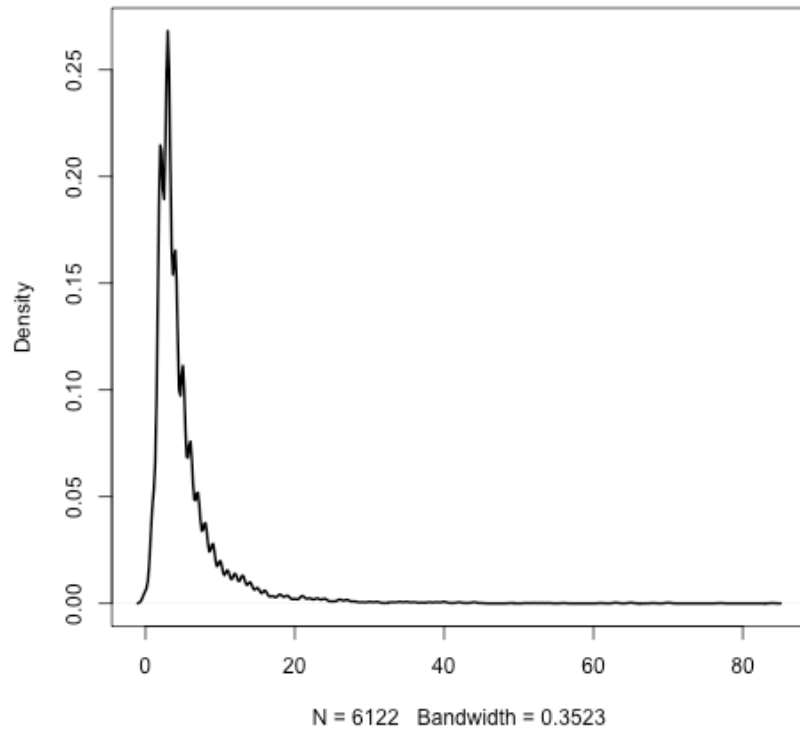
# Exploring distribution of variables

- What do we expect the distribution to look like?
- Do these expectations hold?
- Check variation and outliers
- Do a few observations have a large influence?
- What is to be considered in later analyses?

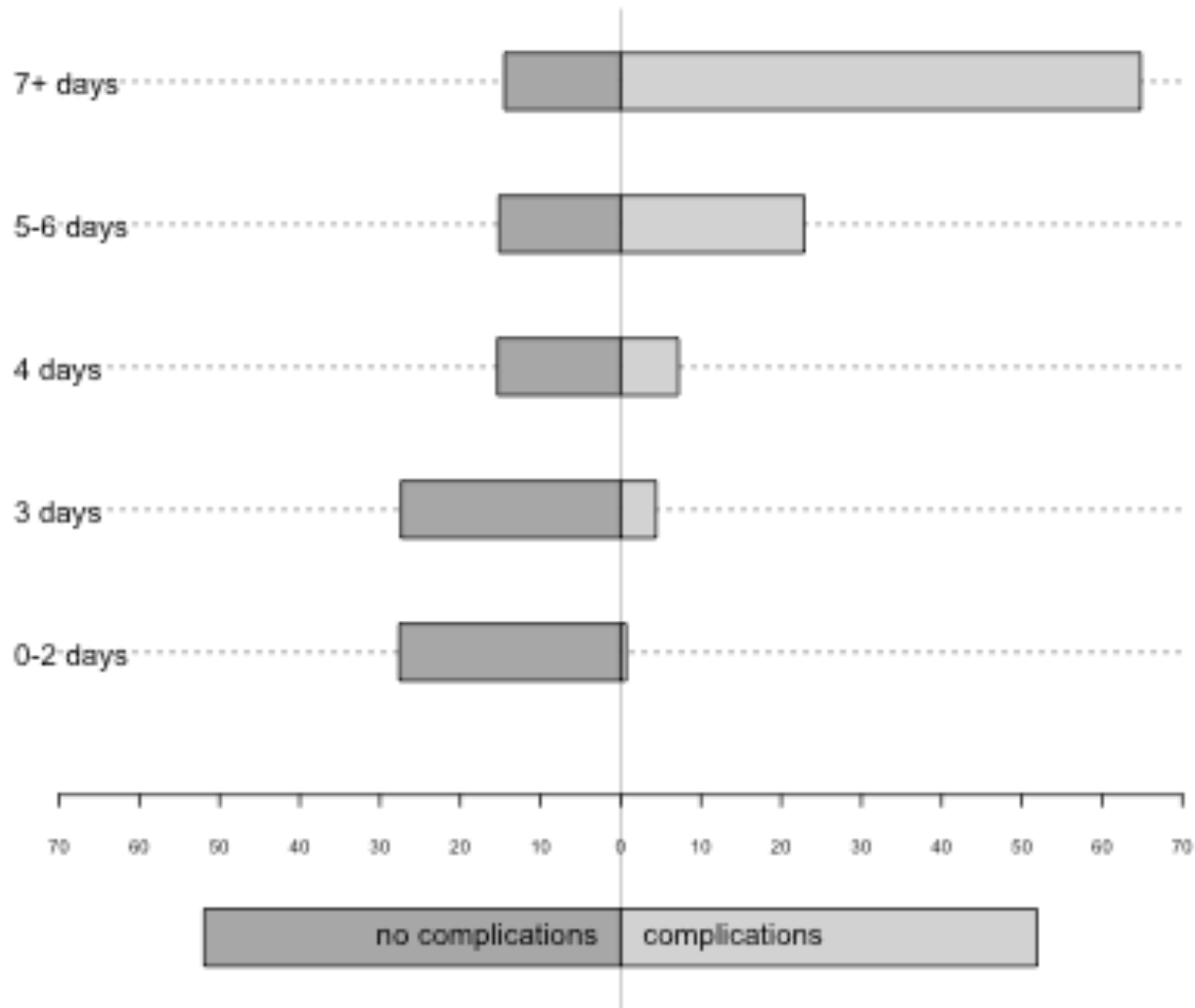
# Length of Hospital Stay [days]

- N=6123 from electronic health records
- Years 2010-2013
- Median (1<sup>st</sup>, 3<sup>rd</sup> quartile): 4 (3,6)
- Range: 0-531 days
- Largest five LOS: 68, 70, 70, 77, 84, 531
- > Error 531 days
- Mean (sd): 5.4 (5.7) (without the 531 LOS)

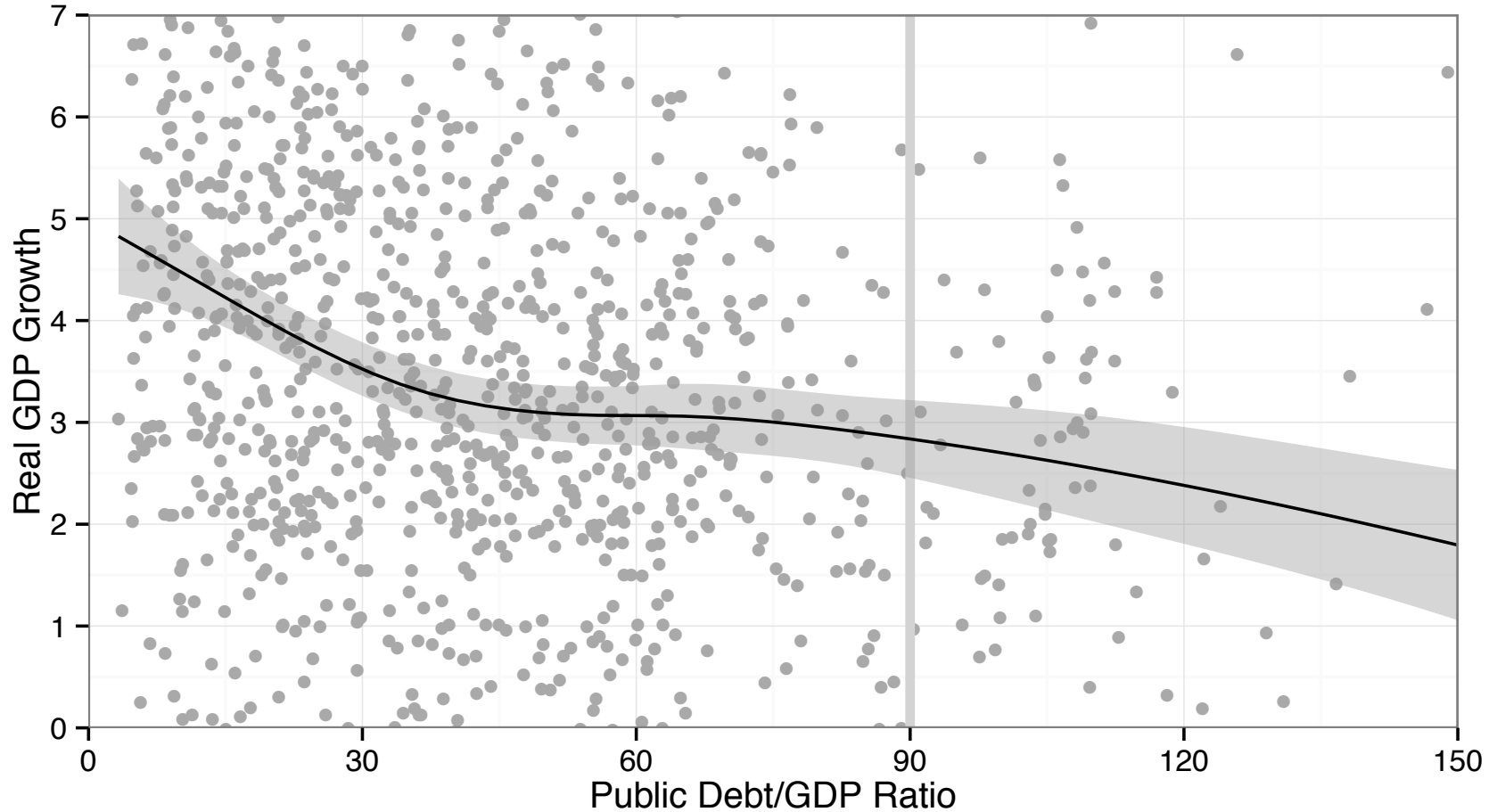
# Length of Stay [days]



# Length of Stay [% cases]



# Cutoff points?

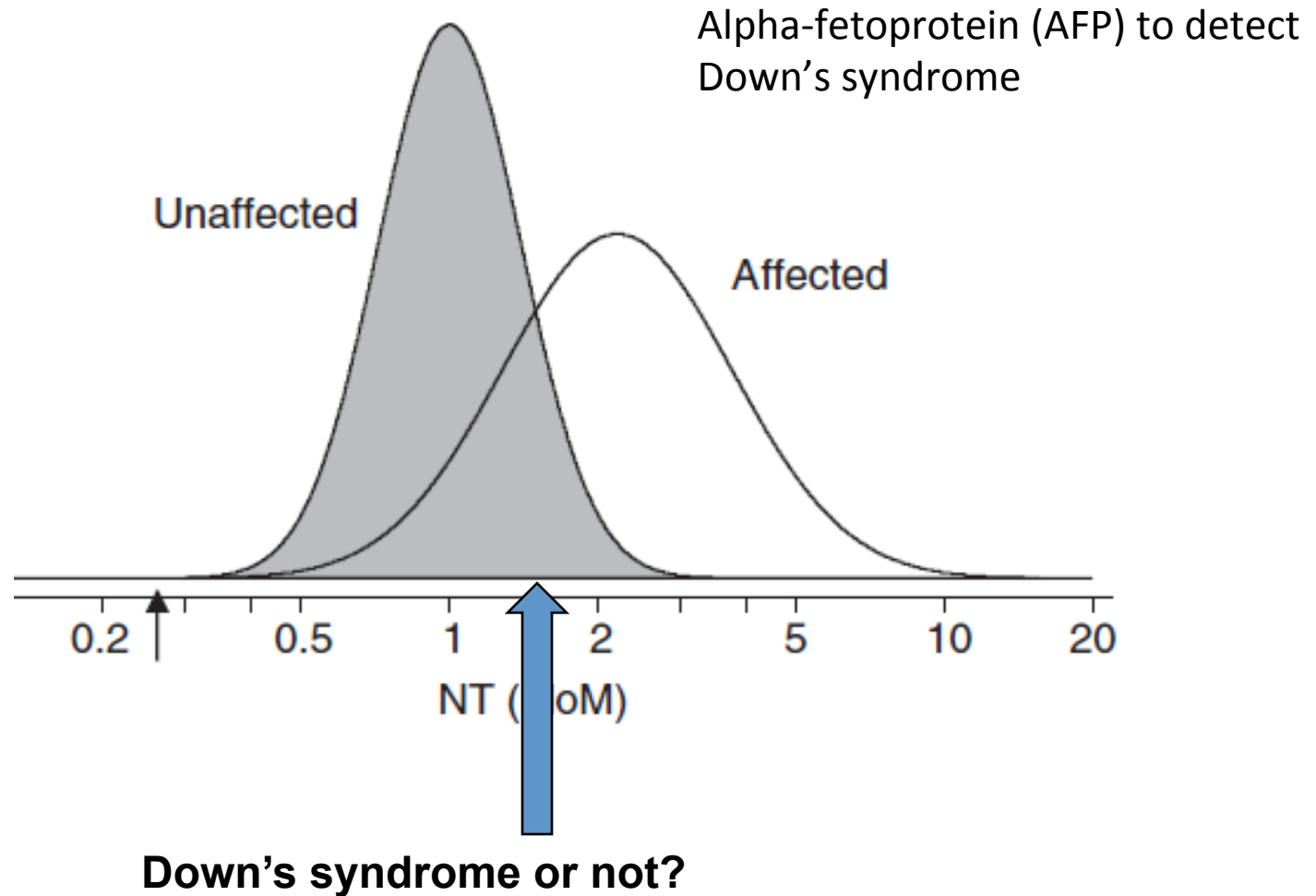


Original groupings were 0-30, 30-60, 60-90, 90+

Herndon, Ash, Pollin. A Critique of Reinhart and Rogoff. 2013.



# Categorization of continuous variables



# Time-to-event analyses

- How consistent/reliable is the follow-up?
  - All subject were contacted or only incidental recording of an event?
  - Detailed records for a time period but sporadic after?
  - All subjects or convenient subjects (e.g readmissions)
- Start time
  - Time since diagnosis
  - Time since surgery
  - Time since entry into study

# Correlated events

1. Several measurements per subject
2. Time dependent covariates (one endpoint per subject, but a covariate changes over time)
  - Crossover treatments
  - Lab tests
3. Multiple events per subjects
  - Repeated infections
  - Rehospitalizations
  - Recurrence of tumors

# Data set-up for sequential events

Choices in creating the dataset. Which model is being fit?

id	tstart	tstop	status	event	strata	duration
1	0	221	1	0	1	221
2	0	193	0	1	0	193
2	193	1100	0	1	1	907
2	1100	1130	1	0	1	30

Therneau and Grambsch: Modeling Survival Data

Therneau and Crowson: Time dependent variables. Vignette (online)

Putter, Fiocco, Geskus: Competing risks and multistate models

# Data set-up for unordered events

Choices in creating the dataset. Which model is being fit?

id	tstart	tstop	event type	duration
1	0	221	1	221
2	0	193	2	193
2	193	366	2	173
2	366	1200	1	834

# Modern Challenges

- New technologies: complex data, high dimensional data, big data
- Combining data from various sources: electronic health records, laboratory, pharmacy, operation notes
- Feeding data summaries to mobile apps

ERP Compliance (Elective Surgery)

Filter LOS Trends: All Diagnosis: All

Time Period: Last 3 Months

Surgeon

ERP Orderset (%)

Last 3 Months  
YTD

LOS (Days)

Last 3 Months  
YTD CRS  
Average

Pre-Op

Intra-Op

Post-Op

G C I F D N

	96% 93%		88%	79%	63%	83%	83%	82%
	85% 90%		76%	74%	66%	79%	82%	80%
	98% 97%		94%	95%	86%	91%	83%	91%
	84% 86%		75%	78%	68%	72%	76%	84%
	92% 95%		82%	79%	76%	84%	87%	92%
	84% 90%		72%	69%	60%	73%	69%	67%
	100% 96%		86%	92%	59%	76%	88%	100%
	94% 95%		83%	75%	81%	85%	76%	88%
<b>CRS</b>	92% 93%	4.67 4.4	82%	79%	71%	81%	80%	84%

# Modern Challenges

- Reading data from various sources: Images, web, API (Application Programming Interface, e.g twitter, facebook), GIS
- Merging data from various sources (SAS, SPSS, R, Minitab, Excel)

MOOC: *Getting and Cleaning Data*. Coursera, Jeff Leek, John Hopkins University