Managing the data pull

Brian Caffo, Jeff Leek, Roger Peng

@bcaffo

www.bcaffo.com

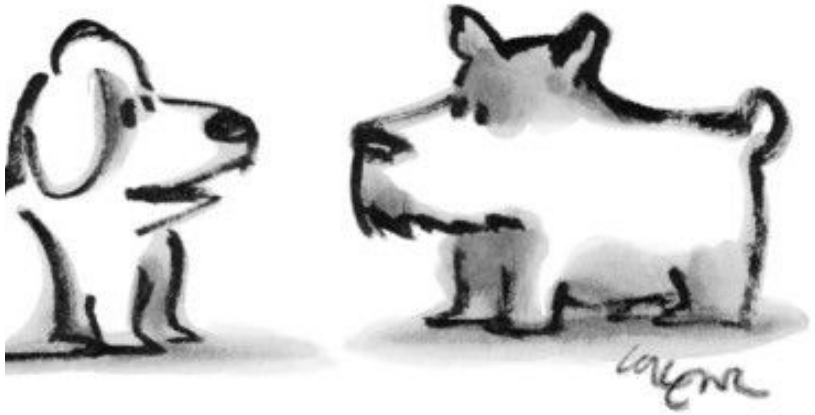
http://www.mlplatform.nl/wp-content/uploads/2013/02/big-numbers1-500x198.jpg

- Almost every data analysis requires at least one of:
  - pulling data from a larger more complex data source
  - merging disparate sorts of data
  - summarizing complex data types (text, speech, images)
  - going from a from a format that is convenient for one purpose to one convenient for analysis (archival -> analytic for example)

- Almost every data analysis requires at least one of:
  - pulling data from a larger more complex data source
  - merging disparate sorts of data
  - summarizing complex data types (text, speech, images)
  - going from a from a format that is convenient for one purpose to one convenient for analysis (archival -> analytic for example)

- As a manager, you likely won't be performing these operations. How do you help manage this process?
  - Lots of resources to help practitioners, fewer resources for managing the practitioners
  - This lecture gives some simple steps for managers

- Summary tables are a great way to catch errors
- In Epi and Biostat the first table summarizing variables is called “Table 1”
- Requiring the regular creation of basic summary tables is a great way to catch errors
  - Get standard deviations along with means, medians and quantiles!
  - Check your units!
  - Compare across reports



"I attribute it to human error. But then I attribute everything to human error."

http://www.condenaststore.com/-sp/I-attribute-it-to-human-error-But-then-I-attribute-everything-to-human-e-New-Yorker-Cartoon-Prints_i8642395_.htm

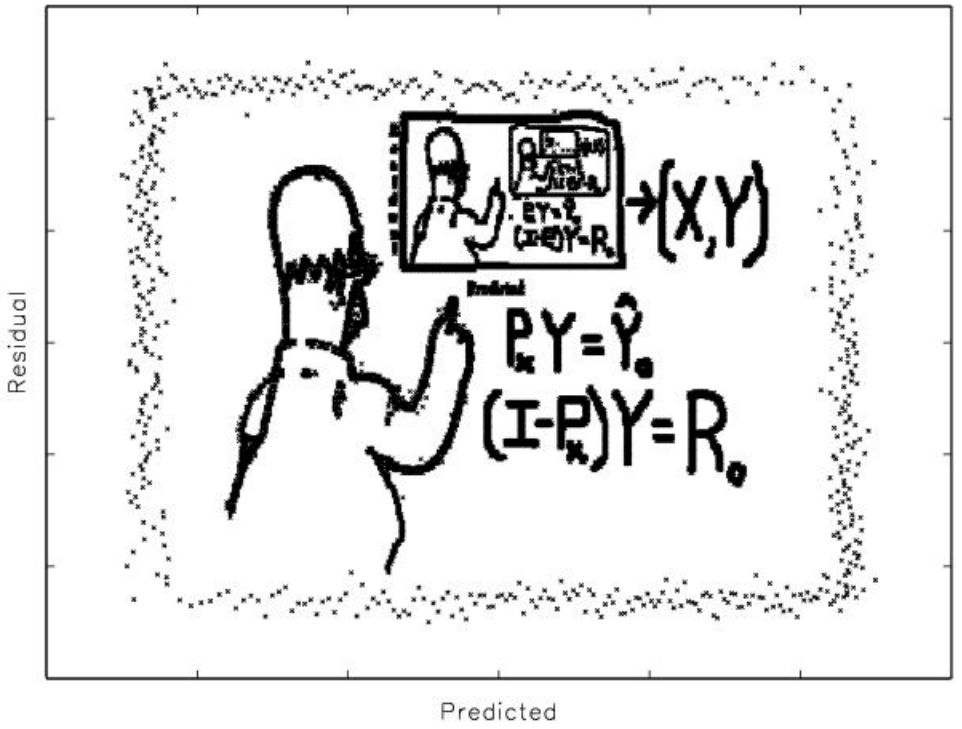
	Volume (cm^3, mean (SD[†]))			
	Gray matter		White matter	
	Regular analysis	Permuted analysis	Regular analysis	Permuted analysis
Peak tibia lead association volume	9.24 (1.14)	0.01 (0.00)	37.33 (3.80)	0.02 (0.01)
Cognitive domain association volume				
Visuo-construction	109.36 (12.17)	1.21 (0.15)	72.61 (7.67)	0.62 (0.09)
Verbal memory and learning	0.01 (0.00)	0.08 (0.02)	0.78 (0.15)	0.01 (0.00)
Visual memory	58.92 (7.65)	0.19 (0.03)	41.42 (4.88)	0.10 (0.01)
Executive function	86.93 (9.49)	0.66 (0.13)	96.45 (9.91)	0.20 (0.05)
Eye-hand coordination	24.68 (3.08)	5.66 (0.65)	54.08 (5.69)	2.77 (0.38)
Processing speed	32.74 (3.86)	22.96 (3.89)	97.26 (10.13)	5.47 (0.80)

http://aje.oxfordjournals.org/content/167/4/429/T1.expansion.html

Regression diagnostics

http://www4.stat.ncsu.edu/~stefanski/NSF_Supported/Hidden_Images/stat_res_plots.html

```
-2.2334 -0.12309 0.11661 -0.27489 -0.18969
-2.2075 0.31311 -0.080886 0.35159 -0.96198
-2.1816 0.016508 0.03525 0.25422 -0.71549
-2.3055 -0.024318 0.03394 -0.0025414 -0.43821
-2.2796 -0.018672 -0.0023774 -0.07975 -0.33685
-2.2666 -0.35624 0.11294 0.057874 -0.2638
-2.2537 0.11336 -0.072337 0.31323 -0.74411
-2.2277 -0.14432 0.20022 -0.14629 -0.36512
-2.2148 0.04702 -0.29785 0.1121 -0.27182
-2.2018 -0.11159 0.19897 -0.0092971 -0.51818
-2.1759 0.031665 0.38068 -0.094308 -0.75903
-2.1629 0.20907 0.028912 -0.34154 -0.34405
-2.2508 0.3781 -0.30421 0.15836 -0.61263
-2.2249 -0.25707 -0.25772 -0.3242 0.36802
-2.1341 0.019149 0.11334 -0.23196 -0.34893
-2.1212 0.20901 -0.18044 0.21326 -0.63029
-2.1082 -0.24877 -0.48139 0.41956 -0.085211
-2.248 -0.12026 -0.20718 -0.2184 0.096614
-2.2221 0.13112 -0.065497 -0.080288 -0.40604
-2.2091 0.26119 0.030953 -0.0038927 -0.69406
-2.1961 -0.17863 0.023455 -0.25395 -0.050767
-2.1702 -0.086929 -0.081813 0.28292 -0.51474
```

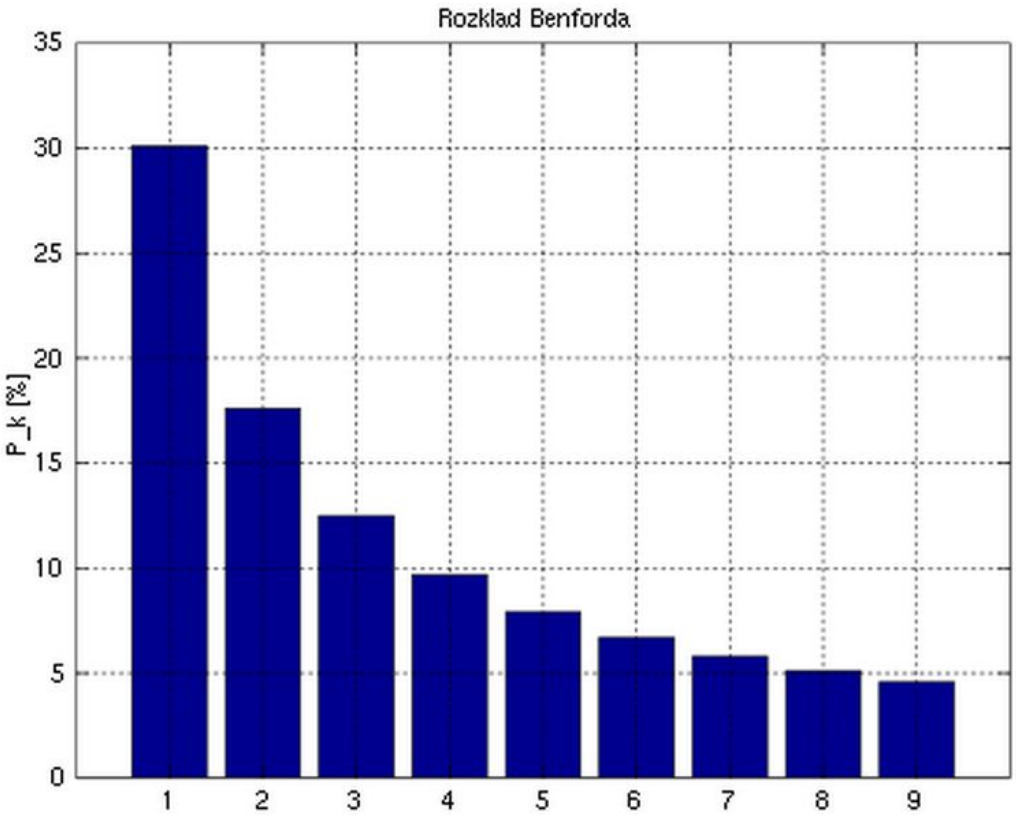


- **Residual** - the difference between the response and the fitted value; residual plots shouldn't have systematic patterns
- **Hat values** - consider how variable a data row is among the space of predictors
- **DF fits, DF betas, Cook's distance** - how much do fitted values and coefficients change when a point is not included in the fit?
- **PRESS residuals, leave one out residuals** - how much do predictions change when a point is left out of an analysis

“**Benford's law**, also called the **First-Digit Law**, is a phenomenological law about the frequency distribution of leading digits in many (but not all) real-life sets of numerical data.”

https://en.wikipedia.org/wiki/Benford%27s_law

Try it out



"Benford-physical" by Drnathanfurious at en.wikipedia - Transferred from en.wikipedia to Commons by User:Tam0031 using CommonsHelper.. Licensed under Public Domain via Commons - https://commons.wikimedia.org/wiki/File:Benford-physical.svg#/media/File:Benford-physical.svg

- You can’t check every data point
- However, you can query and check some
- Use statistical sampling logic to estimate the proportion of bad data in your sample