

UNIVERSITY *of* WASHINGTON

# Data Science UW

## Methods for Data Analysis



More on Hypothesis Testing, The Central Limit Theorem,  
And an introduction to Regression  
Lecture 4  
Steve Elston





Excellent health statistics - smokers are less likely to die of age related illnesses.'

W

# Topics

---

- > Review
- > Outliers
- > Analysis of Variance
- > Central Limit Theorem
- > Introduction to resampling
- > Presentation of data science results



# Review

- > Sampling Methods
- > Law of Large Numbers
- > Hypothesis Testing
  - Normal testing
  - One tailed vs Two tailed
  - P-values
  - T-test (Student's, Welch's)
  - Chi-Squared
  - Fisher's Exact



# Outliers

## > Outlier causes:

- Bad data
  - > Sensor misread, human error, software error
- Non-representative data
  - > Real data that can be argued to be out of our interest. E.g. a sample of annual salaries that includes Warren Buffet.
- Must provide a legitimate argument to consider as outlier.
- Or, an **interesting aspect of the dataset** previously overlooked?



# Finding Outliers – Statistical methods

- > Alpha trimmed mean – aka truncated mean
    - Trim percentage (alpha) of outliers
    - Upper, lower or balanced trimming
    - Iterative method
    - Biased estimator
    - Windsor mean – replace outlier values with trim point values
- Tukey et. al. 1947

Example with two-sided alpha = 1/3

$$X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9 + X_{10} + X_{11} + X_{12}$$

---

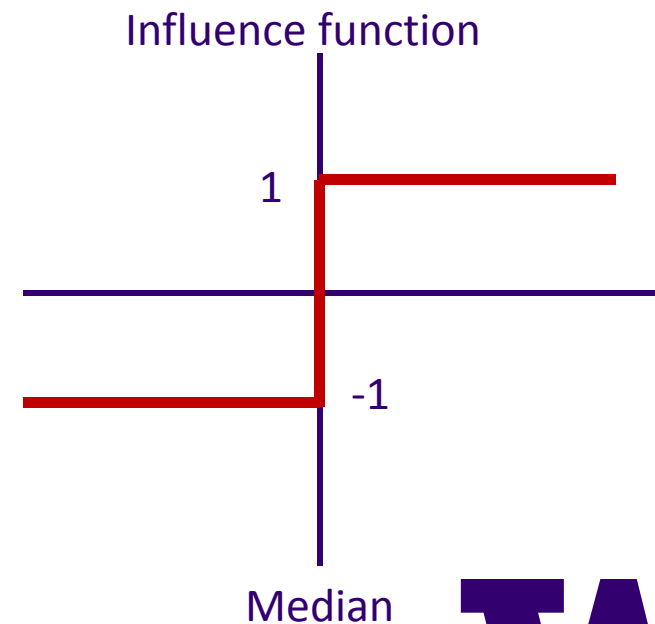
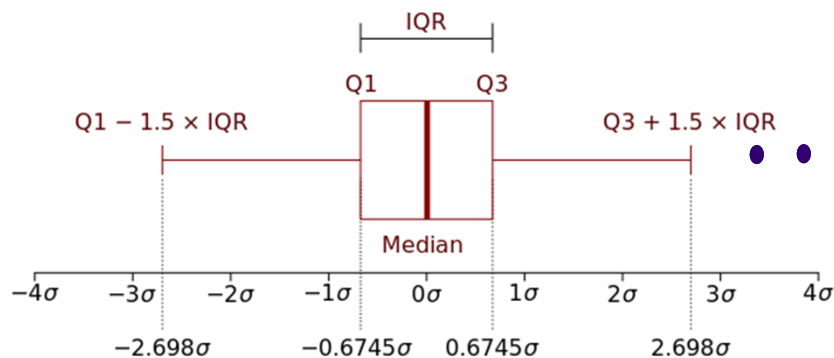
$$12 - 4$$

**W**

# Finding Outliers – Statistical methods

## > Median

- Median is a robust estimator
- Use interquartile range to detect outliers
- Biased estimator

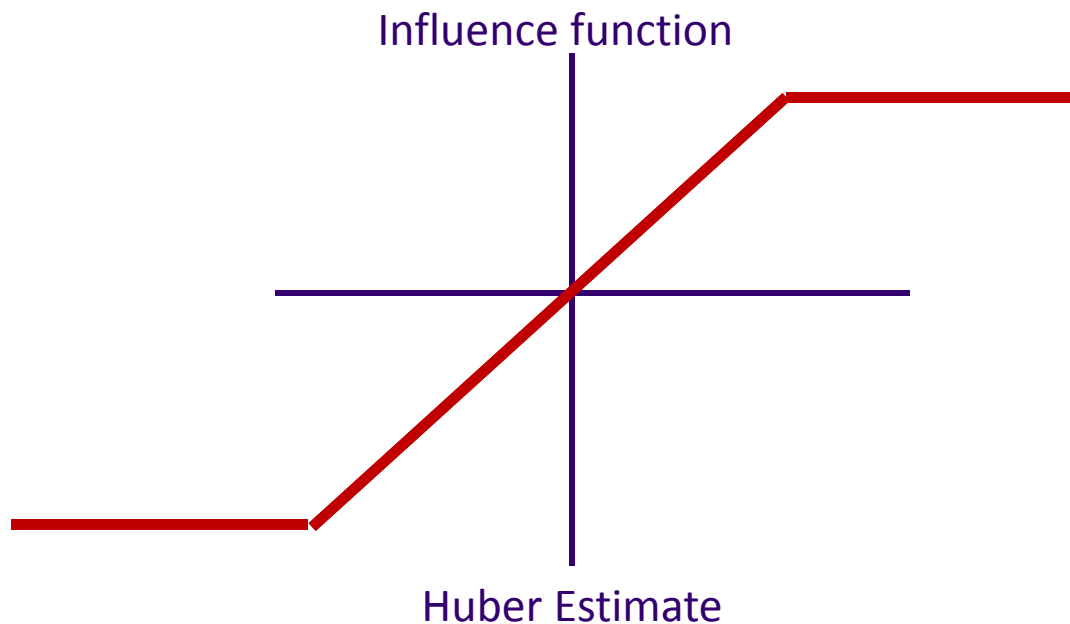


W

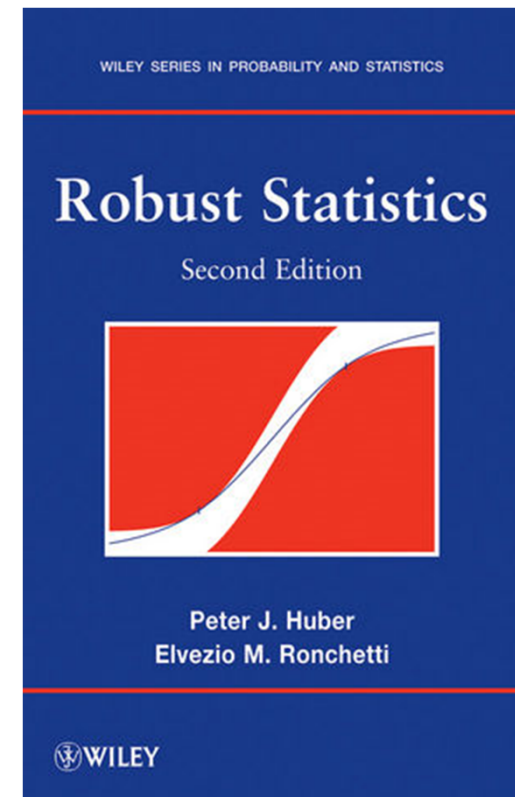
# Finding Outliers – Statistical methods

## > Huber estimator

- Attempt to reduce bias
- Limit influence of outliers
- Use piecewise influence function



First Edition 1981



W



# Finding Outliers – Statistical methods

## > Resampling

- Find points with exceptional 'influence' on the estimate
- Special case of Jackknife method
- Computationally intensive



# Validating Outliers

---

Is an outlier an error or a valuable case?

- > Investigate multiple relationships in dataset to validate outlier
- > Think what interesting or important relationship the 'outlier' might represent.



# Treating Outliers

- > Censor
- > Interpolate new value
- > Use substitute values



# Hypothesis Testing Summary (so far)

- > If data is normal,
  - If you know population mean and variance,
    - > Use standard normal 'z-test'.
  - If you just know population mean,
    - > Use t-test (unpaired data).
    - > Use Welch's t-test (paired data).
- > For categorical comparison tests,
  - If the sample/subgroup size is large enough,
    - > Use Chi-squared test
  - If the sample/subgroup size is small,
    - > Use Fisher's Exact test.
- > How do we know the data is normal?



# Testing Between Multiple Groups

- > What if we had multiple groups and we wanted to compare their means?
- > Why can't we just do multiple two-sample t-tests for all pairs?
  - Results in increased probability of accepting a false hypothesis.
  - E.g., if we had 7 groups, there would be  $(7 \text{ Choose } 2) = 21$  pairs to test. If our alpha cutoff is 5%, then we are likely to accept about 1 false hypothesis ( $21 * 0.05$ ).



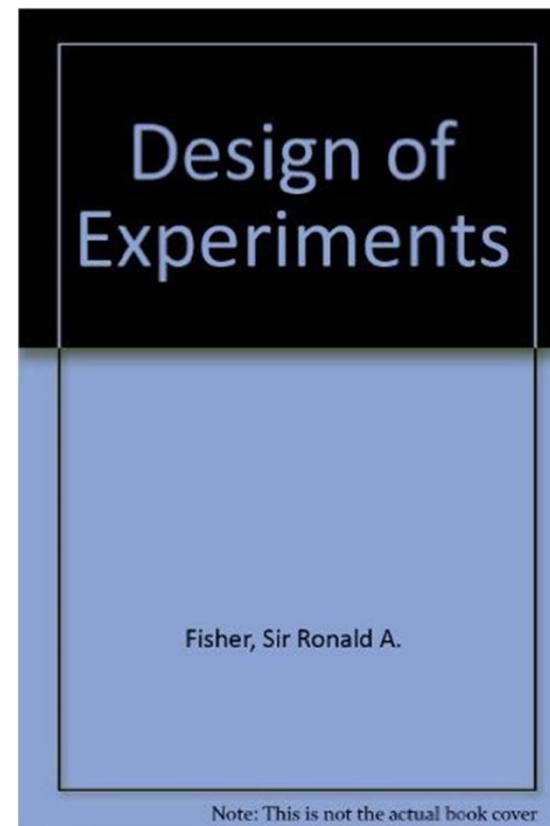
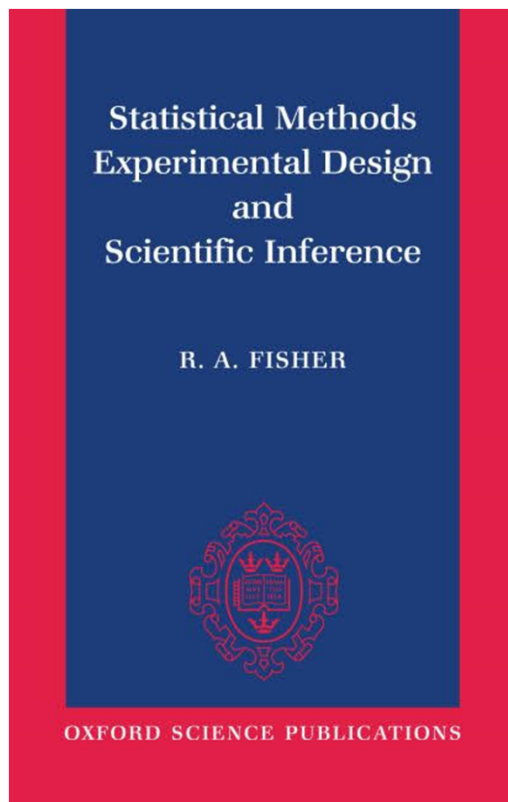
# Testing Between Multiple Groups

- > Null Hypothesis:
  - All groups are just samples from the same population.
- > Alternative Hypothesis:
  - At least one group has a statistically different mean.
- > This type of analysis is called “ANalysis Of VAriance”, or ANOVA.
  - We make data independence and normality assumptions first.



# ANOVA

- > Laplace, 1827
- > Fisher, 1922, 1925, 1935 – F statistic



**W**

# ANOVA Calculations

Basic ANOVA calculations:

$I$  = number of treatments

$n$  = number of data

$SS$  = sum of squares

$$SS_{\text{Total}} = SS_{\text{Treatment}} + SS_{\text{Error}}$$

$$DF_{\text{Total}} = DF_{\text{Treatment}} + DF_{\text{Error}} = (I - 1) + (n - I) = n - 1$$

$$F \text{ statistic with } I - 1 \text{ DF} = \frac{\text{Variance between treatments}}{\text{Variance within treatments}}$$
$$= \frac{\frac{SST}{DF_{\text{Treatment}}}}{\frac{SSE}{DF_{\text{Error}}}}$$





# Performing ANOVA

- > ANOVA table lays out calculation
- > F statistic determines significance (P value)
- > Significance (P value) is key

Example;

One-way balanced ANOVA for 5 treatments, 30 samples

**ANOVA**

		Sum of Squares	df	Mean Square	F	Sig.
Self-Concept	Between Groups	1914.087	4	478.522	1.297	.298
	Within Groups	9223.687	25	368.947		
	Total	11137.8	29			

**W**

# Performing Multiple Hypothesis Tests

- > For non-ANOVA methods, remember that performing many hypothesis tests increases our risk of incorrectly rejecting a null-hypothesis.
- > To compensate for this we decrease the p-value cutoff.
- > The most common way of doing this is with the Bonferroni Correction.

$$p' = \frac{p}{(\# \text{ of Hypotheses})}$$

- > This correction is argued to be too strong and other approximations for a new-p can be used instead.
  - Tukey's Range Test
- > This is VERY important in genetics/bioinformatics.

**W**

# Additional Hypothesis Testing

## > Tests may lack power!

- Need sufficient sample size
- Size of the effect must large enough
- ‘Reasonable’ significance level

$$\text{Power} = P(\text{reject } H_0 | H_1 \text{ is true})$$

## > Parametric test types:

- Mean comparison
- Variance comparison
- More distribution comparisons

## > R Example



# Central Limit Theorem

- > Sample a population many times, the distribution of means of all samples are normally distributed, regardless of the population distribution.

$\bar{X}$ =sample mean.

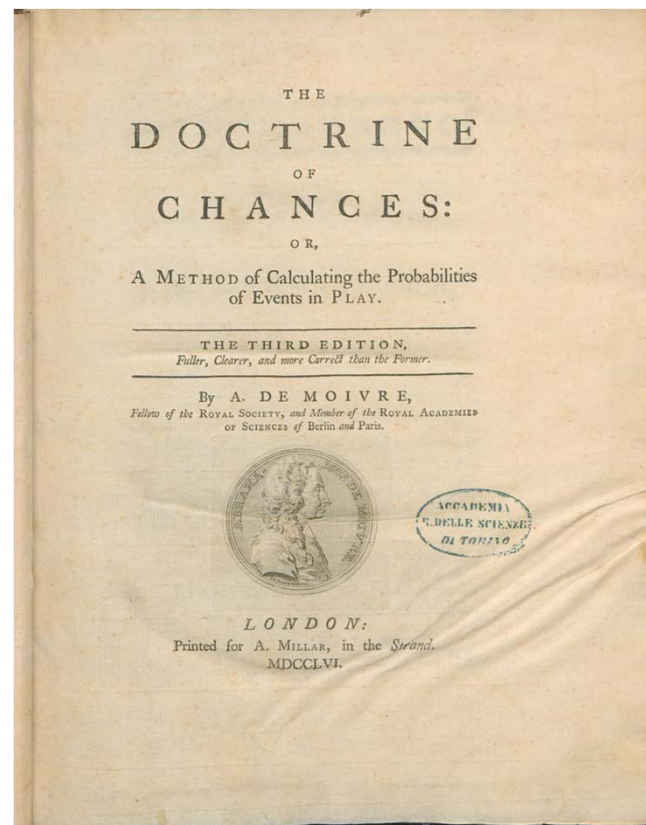
$$\bar{X} \sim N\left(\text{mean}, \frac{\text{st. dev}}{\sqrt{n}}\right)$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$



# Central Limit Theorem

- > de Moivre, 1738 – prof of special case for Bernoulli trials
- > Laplace 1776, 1785, 1820
- > Chebyshev, 1887 – rigorous proof



W

# Central Limit Theorem

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- > We can use this central limit theorem to generate confidence intervals on expressing the population mean.
- > We know the sample mean, sample variance, and number of samples.
- > Then we know how our estimate of the population mean is distributed (from above formula).
- > We can then generate 90%, 95%, ... confidence intervals around our sample mean.



# Confidence Intervals

- > Confidence intervals are a way to express uncertainty in *population* parameters, as estimated by the sample.
- > E.g. If we create a 95% confidence interval for the population mean, say  $\hat{\mu} = \bar{X} = 10 \pm 5$ 
  - Then we say that the true population mean,  $\mu$ , has a 95% chance of being between 5 and 15.

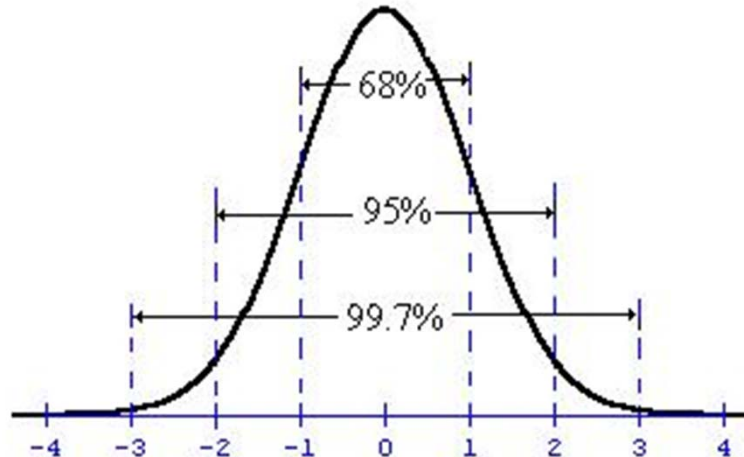
It is **not** correct to say:

- ~~“95% of the sample values are in this range.”~~
- ~~“There is a 95% chance that the mean of another sample will be in this range.”~~

W

# Confidence Intervals

- > To create confidence intervals for population means, we use the central limit theorem and create confidence intervals based on the normal distribution.
  - Repeatedly sample from the population.
  - Calculate the mean for each sample.
  - Use the average of the sample means as the population estimate and create a C.I. based on the s.d. of the sample means.
  - R demo



**W**



# How to work with limited sample size?

Sample size is always limited

- > Point estimates are only computed once
- > How reliable are point estimates?



# Resampling Methods

What are resampling methods?

- > Resampling methods allow computation of statistics from limited data
- > Compute statistic from multiple subsamples of dataset
- > Minimal distribution assumptions
- > Computationally intensive



# Resampling Methods

## Common resampling methods

- > Permutation methods
- > Bootstrap: resample with equivalent size and replacement
- > Jackknife: leave one out resampling
- > Cross validation: resample into folds without replacement



# Bootstrap Methods

- > Efrom, 1979
- > Re-compute statistic many times with sample with replacement
- > Randomly subsample (e.g. Bernoulli sample) data with replacement
- > Subsamples have the same size as original sample
- > Works with any statistic ... in principle

Example compute bootstrap mean

$$\text{Meanboot} = \left( \sum \text{mean}(\text{sample}_i) \right) / \text{nsample}$$

$$\text{sample}_i = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_1 + X_5$$



# Jackknife Methods

- > Quenouille, 1949, 1956; Tukey, 1958
- > Re-compute statistic many times with sample with replacement
- > Randomly leave one (or n) out sampling
- > Only use with statistics with continuous derivatives

Example compute jackknife mean

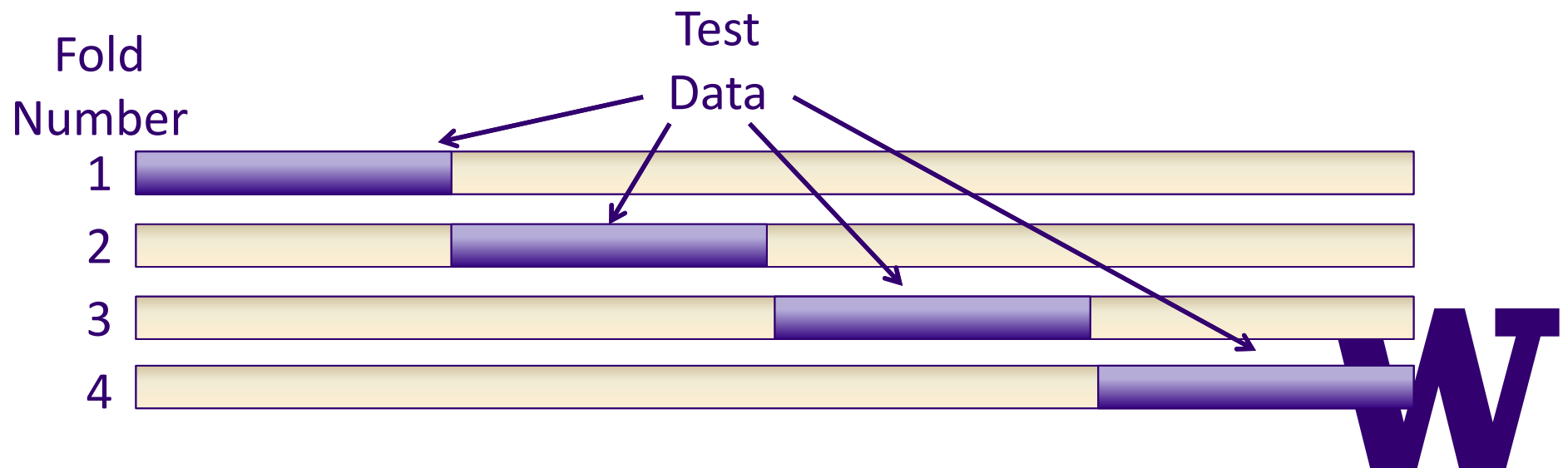
$$\text{Meanjackknife} = \left( \sum \text{mean}(\text{sample}_i) \right) / \text{nsample}$$

$$\text{sample}_i = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_{10}$$



# Cross Validation Methods

- > Mosteller and Tukey, 1968
- > Divide dataset into N subsamples
- > N – 1 Folds train model
- > One Fold evaluate model
- > Nest cross validation to compare models



# Resampling Pitfalls

There is no free lunch

- > If sample is biased, resample statistic is biased
- > Sample variance and CIs are no better than sample allows



# Presentation and story telling

---

Important part of data science

- > Data science must have **impact**
- > Results **only** have impact if they are understood
- > Need to **'tell the story'**
- > **Draw clear conclusion**
- > Evidence supports conclusion

**Presenting results is hard!**

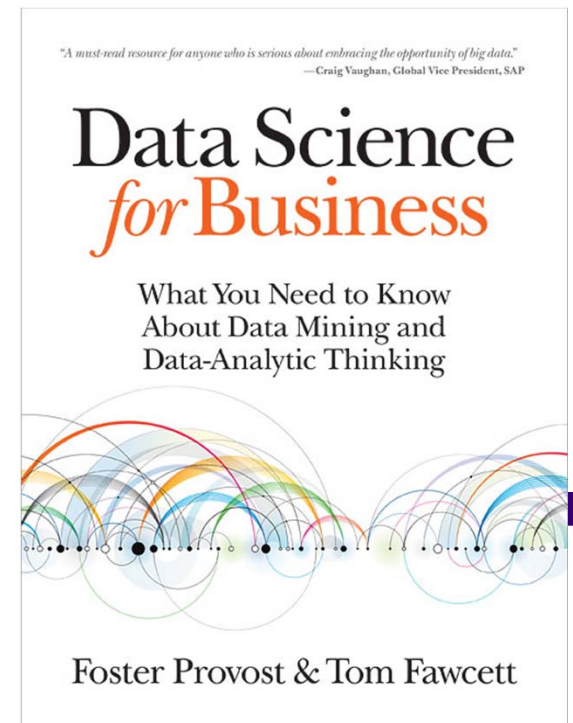
**W**



# Data analytic thinking

Thinking about problems using objective analysis of data

- > Define problem in terms of the business impact
- > Review available data sources
- > Explore the data
- > Try various models
- > Actionable results generate value
- > Support recommendations with data and analysis
- > Define metrics of success



# Tips for story telling

## Make the story clear

- > Occam's Razor
- > You will only hold attention for a short time
- > Don't distract your audience
- > Start with your conclusion
- > Support your conclusion with evidence
- > Few words = **greater impact!**

**W**

# Don't obfuscate your message!

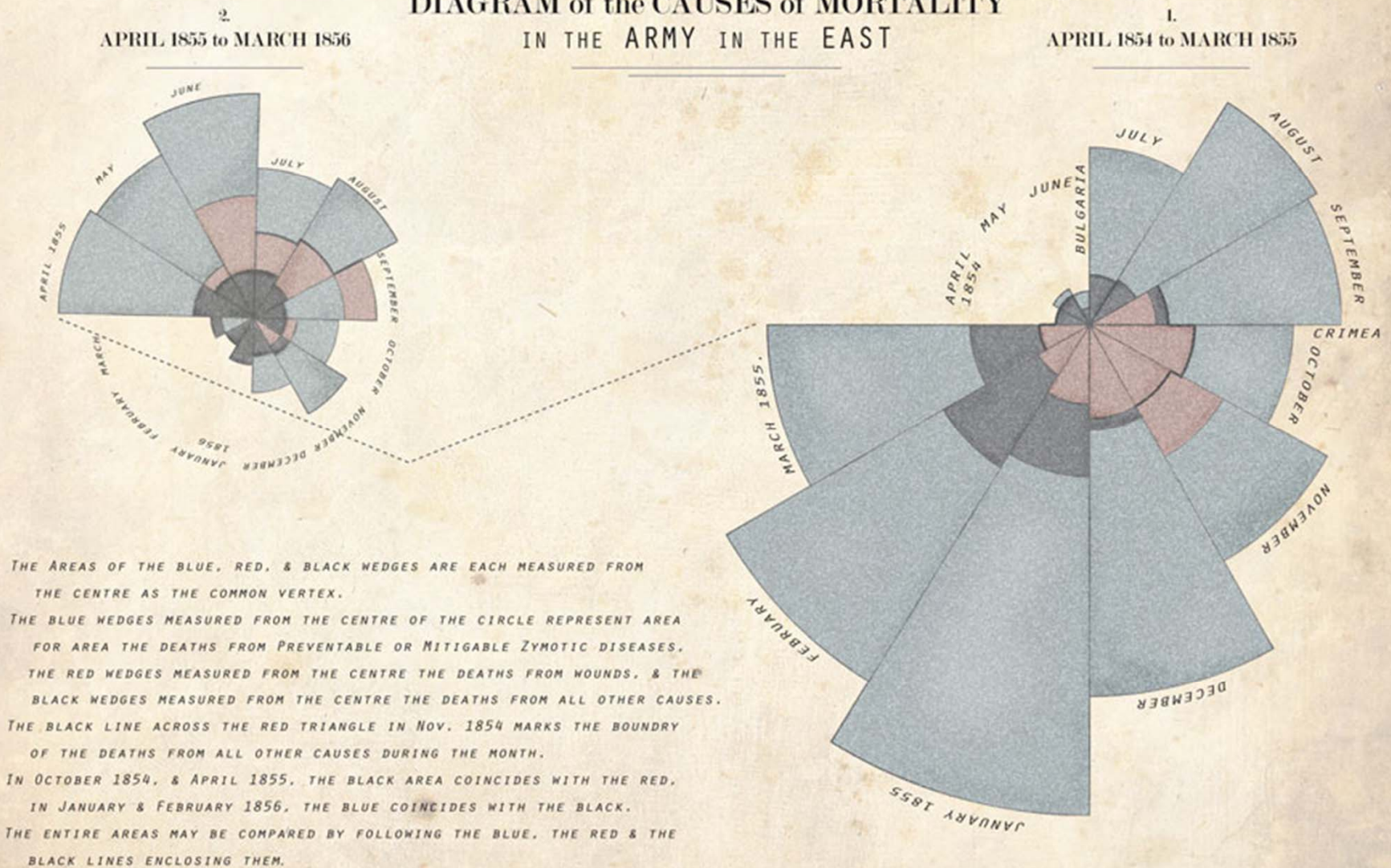
Short and simple has business impact

- > Minimize discussion of methodology and technical detail
- > Clear charts
  - Label axis
  - Minimize over-plotting
  - Simplify
- > Short simple tables
  - Label rows and columns
  - Highlight key point
  - Minimal rows and columns



# Historical example

## DIAGRAM of the CAUSES of MORTALITY IN THE ARMY IN THE EAST







# Presenting Data Science Results

## Suggested reading

- > [www.unomaha.edu/mahbubulmajumder/data-science/fall-2014/lectures/28-presenting-result/28-presenting-result.html#/](http://www.unomaha.edu/mahbubulmajumder/data-science/fall-2014/lectures/28-presenting-result/28-presenting-result.html#/)
- > <http://dupress.com/articles/telling-a-story-with-data/>
- > <http://www.kaushik.net/avinash/data-presentation-tips-focus-think-simplify-visualize/>



# Demo of testing statistical function

Ran out of time before



# Homework and final project

---

- > Grading is based on results and clear and complete presentation
  - Quality, completeness and clarity, not volume, count!
- > Presentation must explain specific conclusions
  - Specific conclusions have impact
- > Support conclusions with charts and tables
  - Narrative must call out the evidence
  - Presentation of evidence to maximize impact
- > **Simplify your presentation!**





# Assignment

## > Complete Homework 4:

- Apply ANOVA to the auto price data:
  - > Compare the price (log price) of autos for several multi-valued categorical variables – number of doors, body style, drive wheels, number of cylinders, engine type
  - > Graphically explore the differences – Hint, make sure you have enough data for each category.
  - > Use standard ANOVA and Tukey ANOVA in R
  - > Use the bootstrap distribution CIs of the (differences of) means – Hint write a function for to perform this calculation for any number of categories (levels) pairs
- You must submit:
  - > One R-script.
  - > Document discussing and supporting your conclusions

## > Read Statistical Thinking for Programmers Chapters 6 and 7.

