Analytics to Support Business Decision in Traditional Industry

25 May 2014
7th R conference (Beijing, China)

DUPONT
PIONEER

Traditional Industry v.s. E-commerce Industry

- NOT big data
- Different time series
- Well defined concentration
- Relationship: Rep Sales play a role
- Without/little online purchasing data
- Machine learning and statistical model
- Market research based on questionnaire
- Time of the year matters (for agriculture)

How Analytics Can Help

➢ Business Opportunities

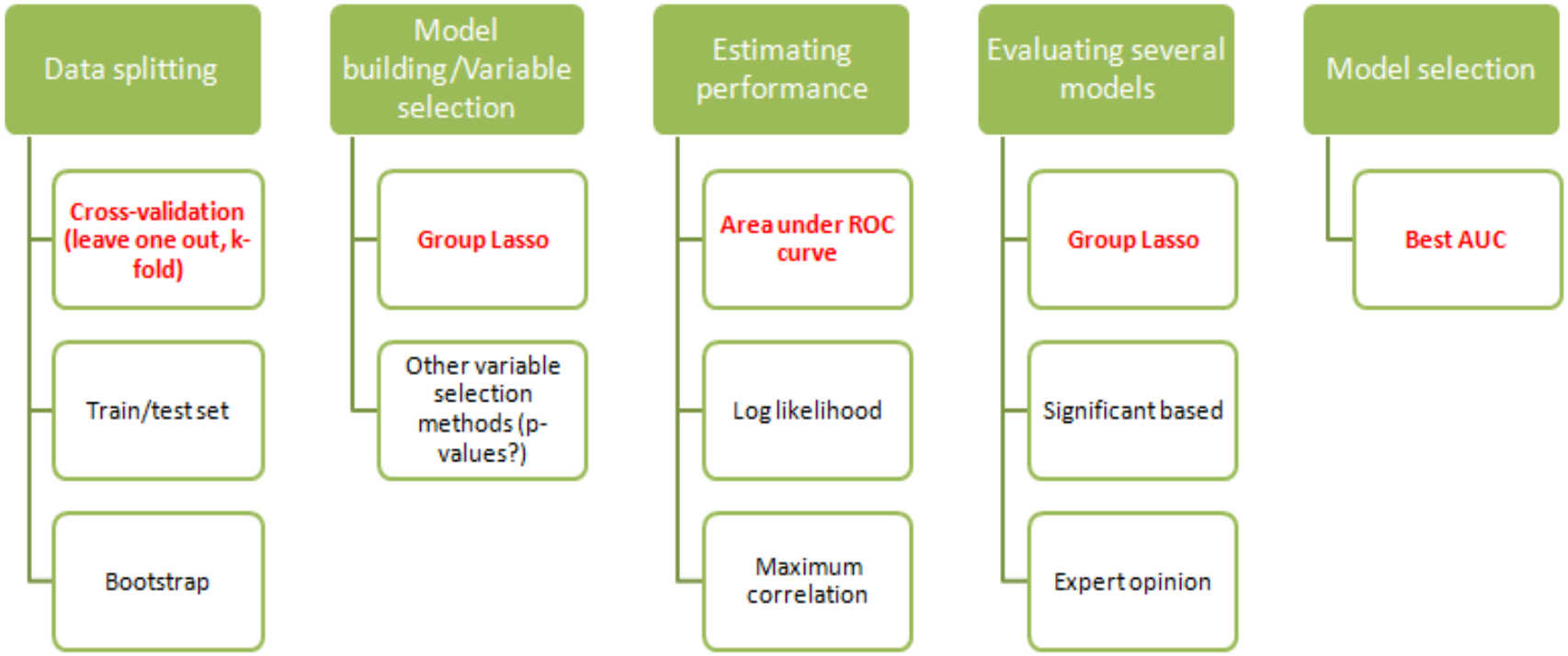
- Cross-selling
- Evaluate marketing campaign
- Target specific population (segmentation)
- R.O.I of different marketing program and services
- Understand customer attrition/conversion/retention
- Predict possible customer reaction before decision implementation
- ……

Not Academic

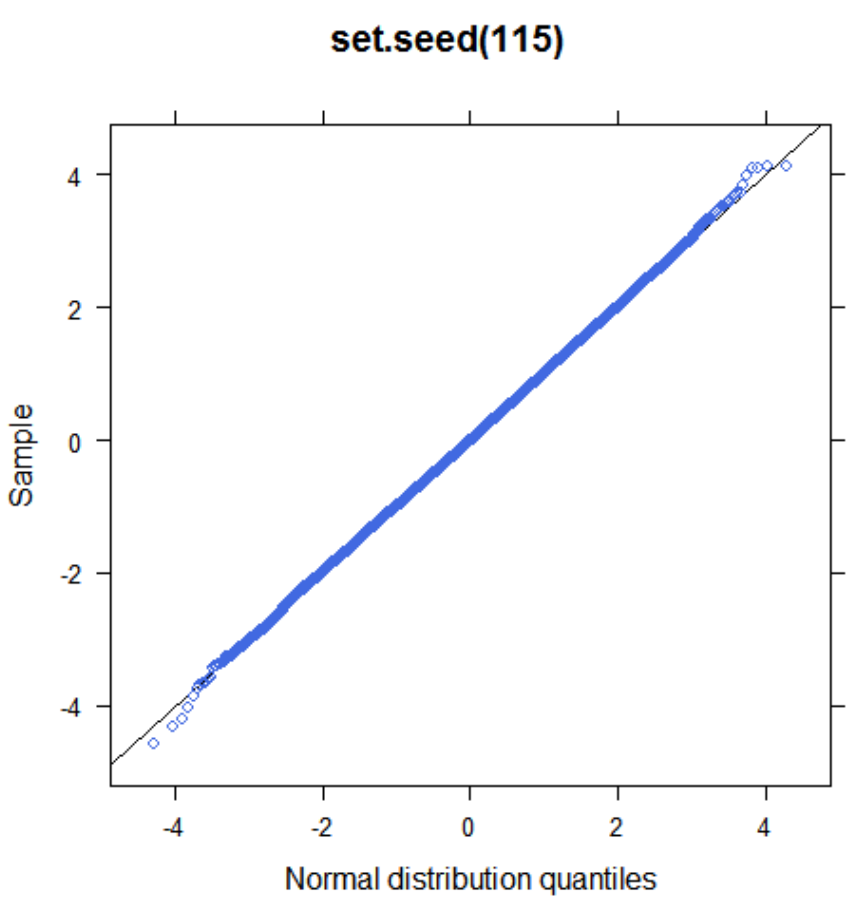
- Dirty data
- Useful v.s. Accurate
- Issues of observational study
  - Hard to define control group
  - Missing value
  - Correlation? Causality?
- Gap between “Implication” and “Implementation”

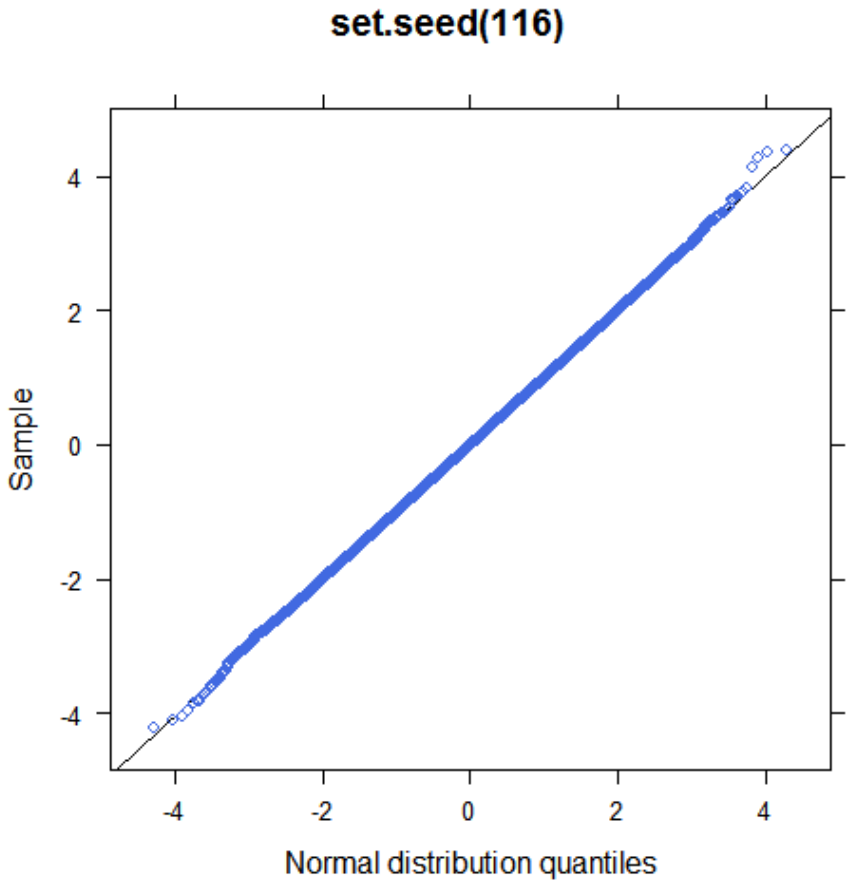
Scoring System for Customer Retention

- Objective: score customers by the likelihood of retention
- Response: 0/1
- Predictors: categorical, continuous (different scales, all possible historical transactional records)
- Data Size: 10000 x 200



50,000 random t variates, 200 d.f.

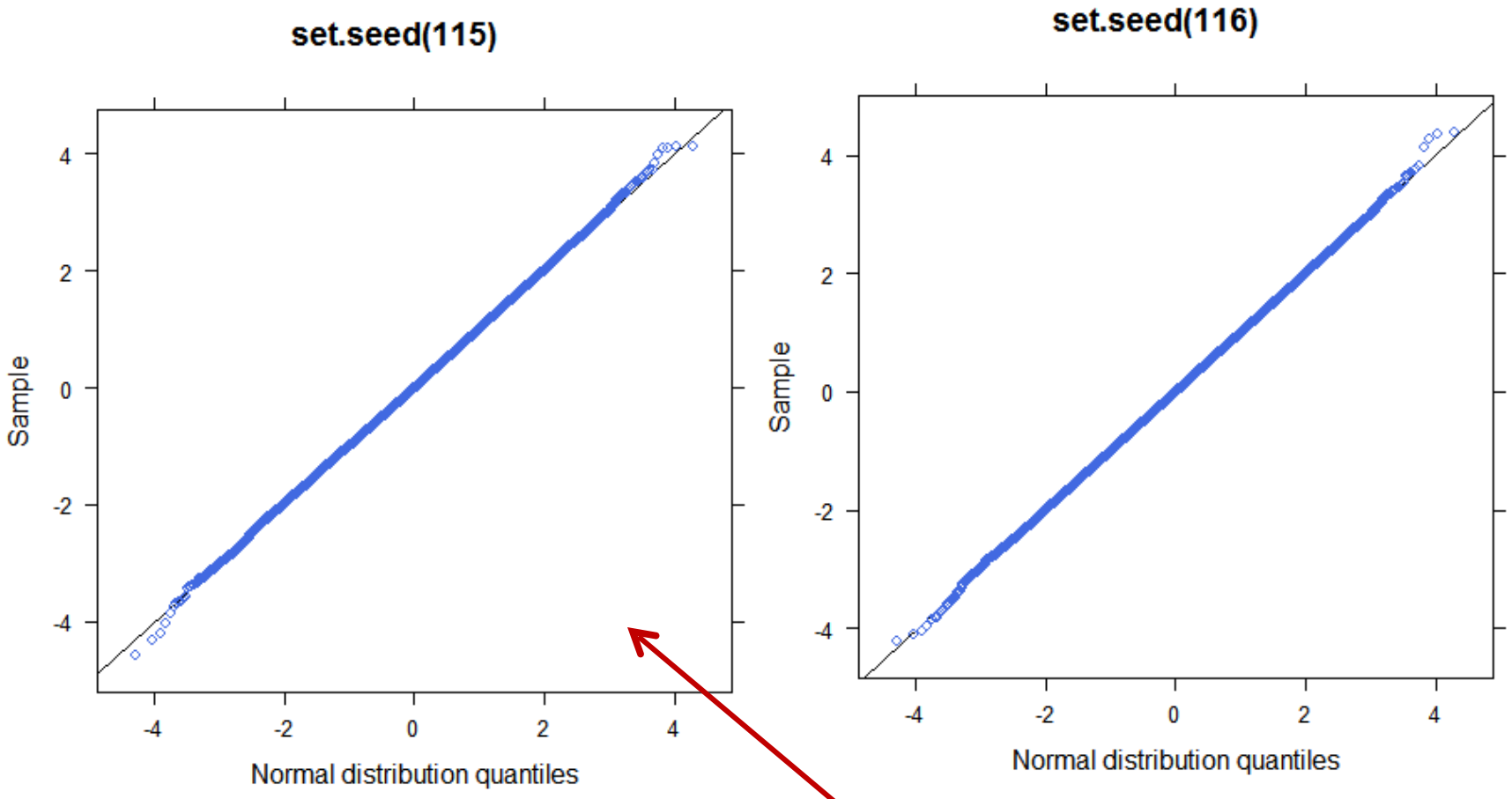




Anderson-Darling normality test
A = 1.1849, p-value = 0.004318

Anderson-Darling normality test
A = 0.1731, p-value = 0.9281

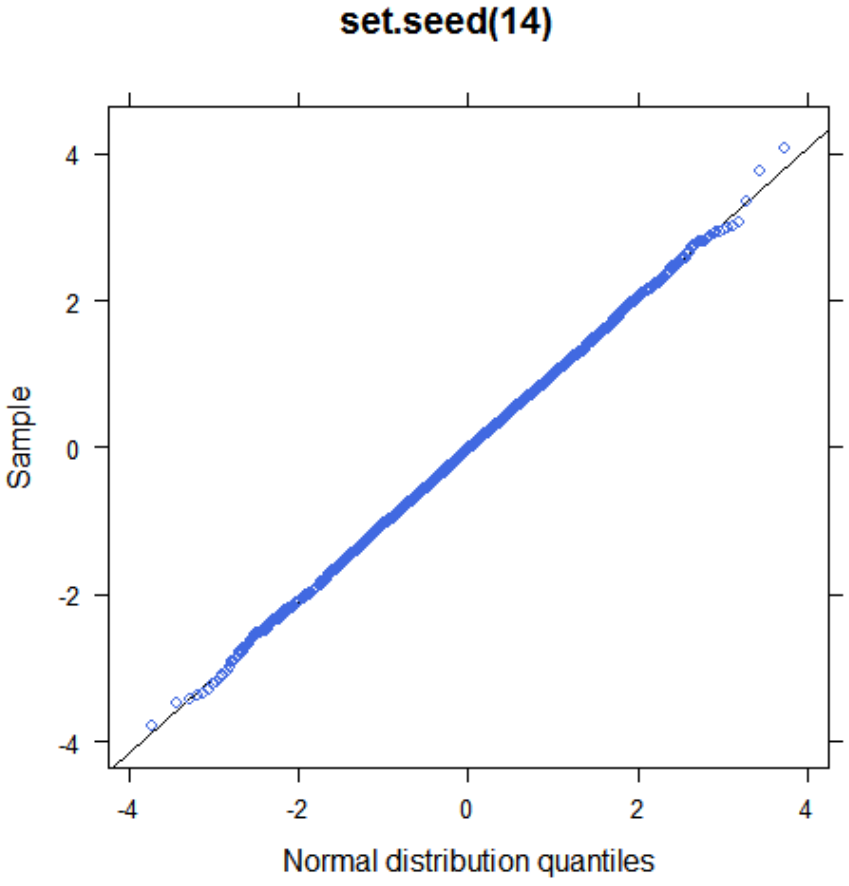
50,000 random t variates, 200 d.f.

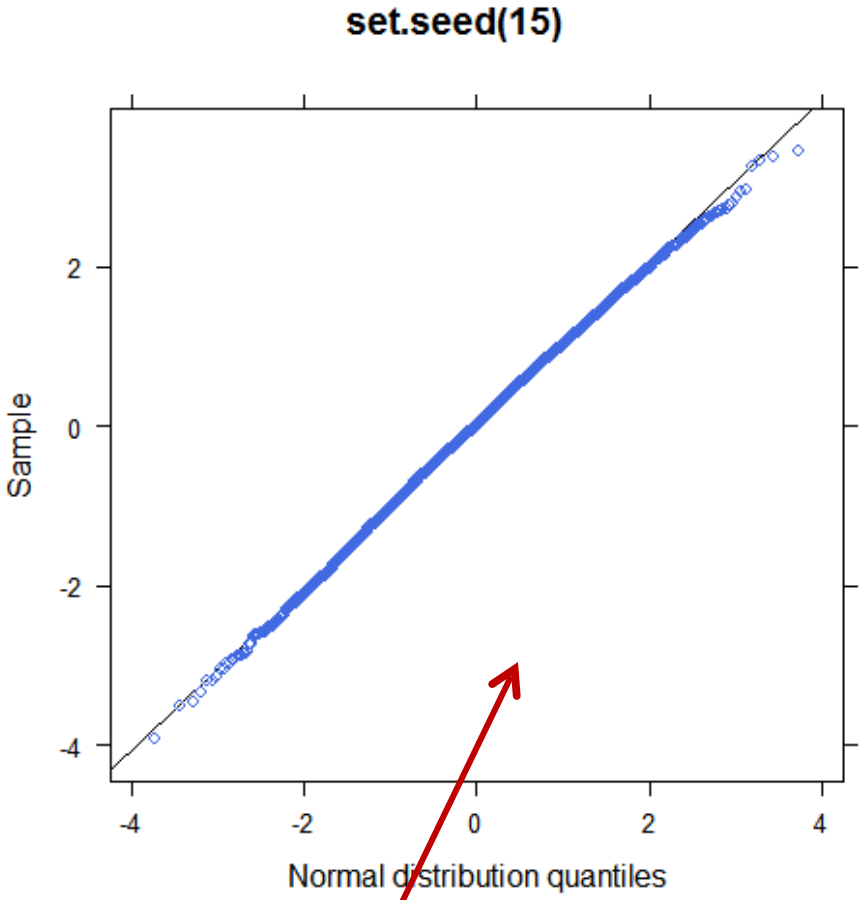


Anderson-Darling normality test
A = 1.1849, p-value = 0.004318

Anderson-Darling normality test
A = 0.1731, p-value = 0.9281

5,000 random t variates, 200 d.f.

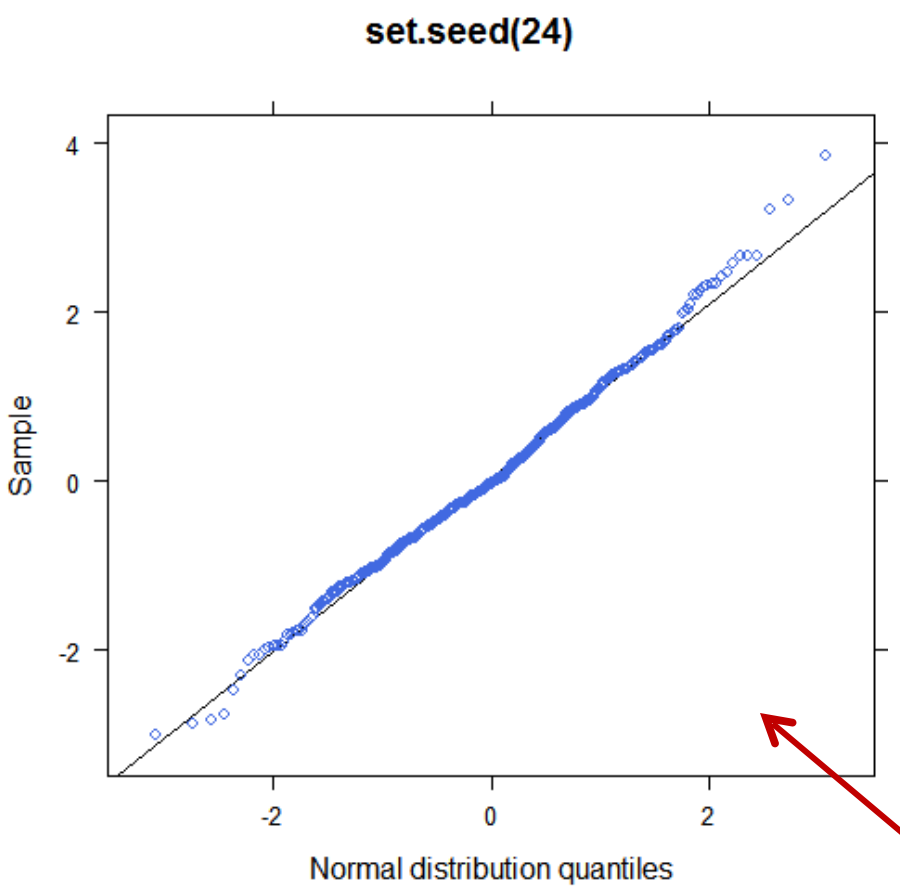


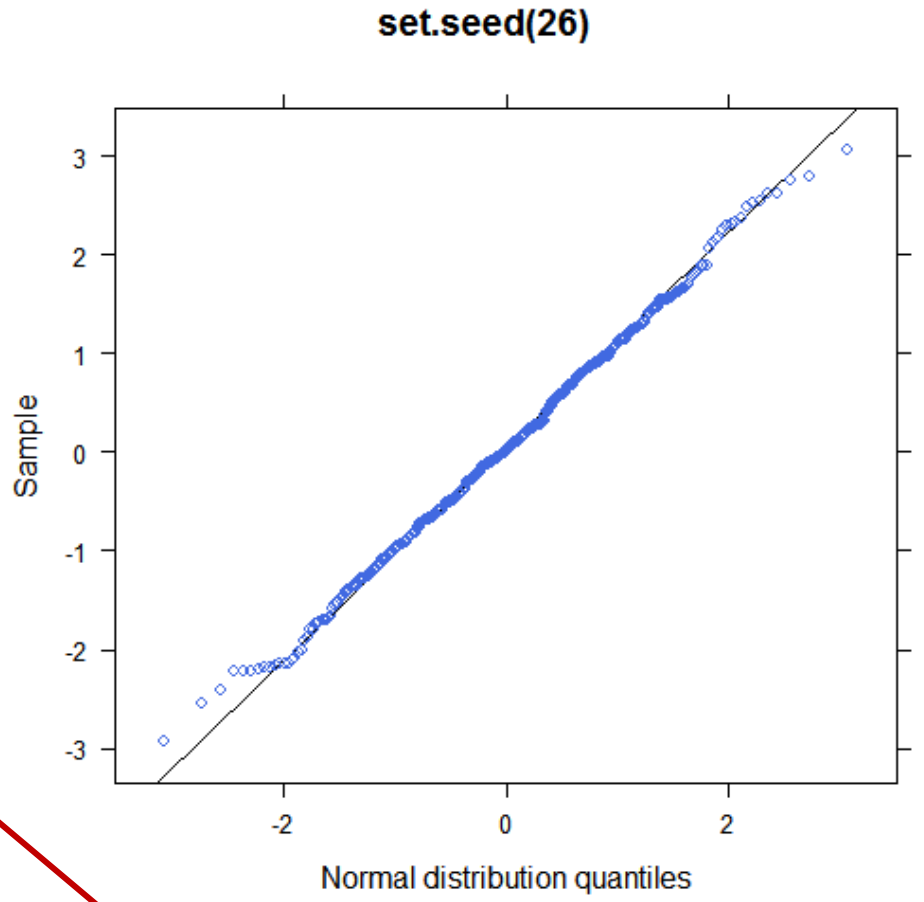


Anderson-Darling normality test
A = 0.2676, p-value = 0.6856

Anderson-Darling normality test
A = 1.0479, p-value = 0.009378

500 random t variates, 200 d.f.

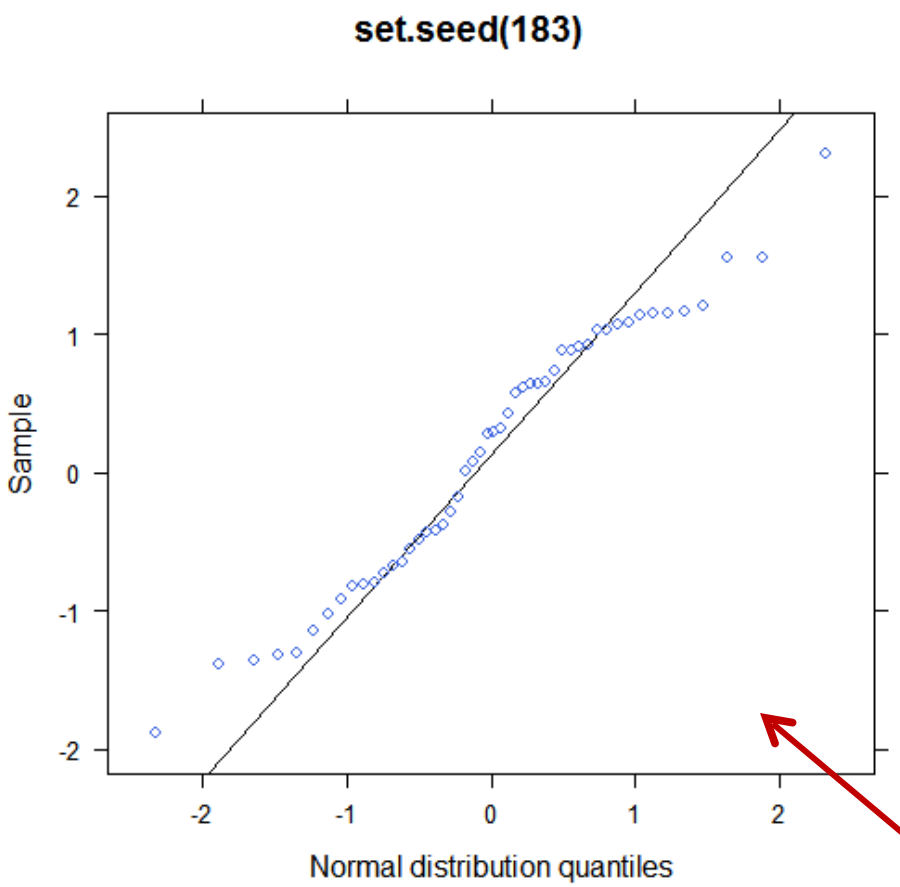


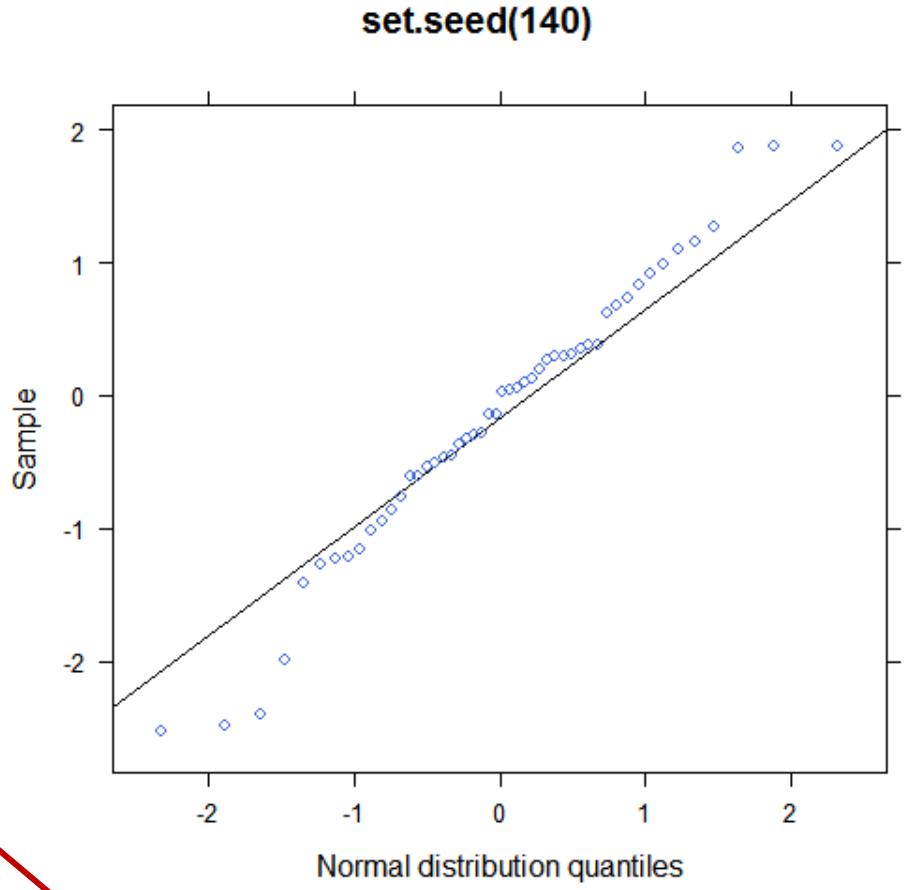


Anderson-Darling normality test
A = 0.7488, p-value = 0.05094

Anderson-Darling normality test
A = 0.2427, p-value = 0.7672

50 random t variates, 200 d.f.

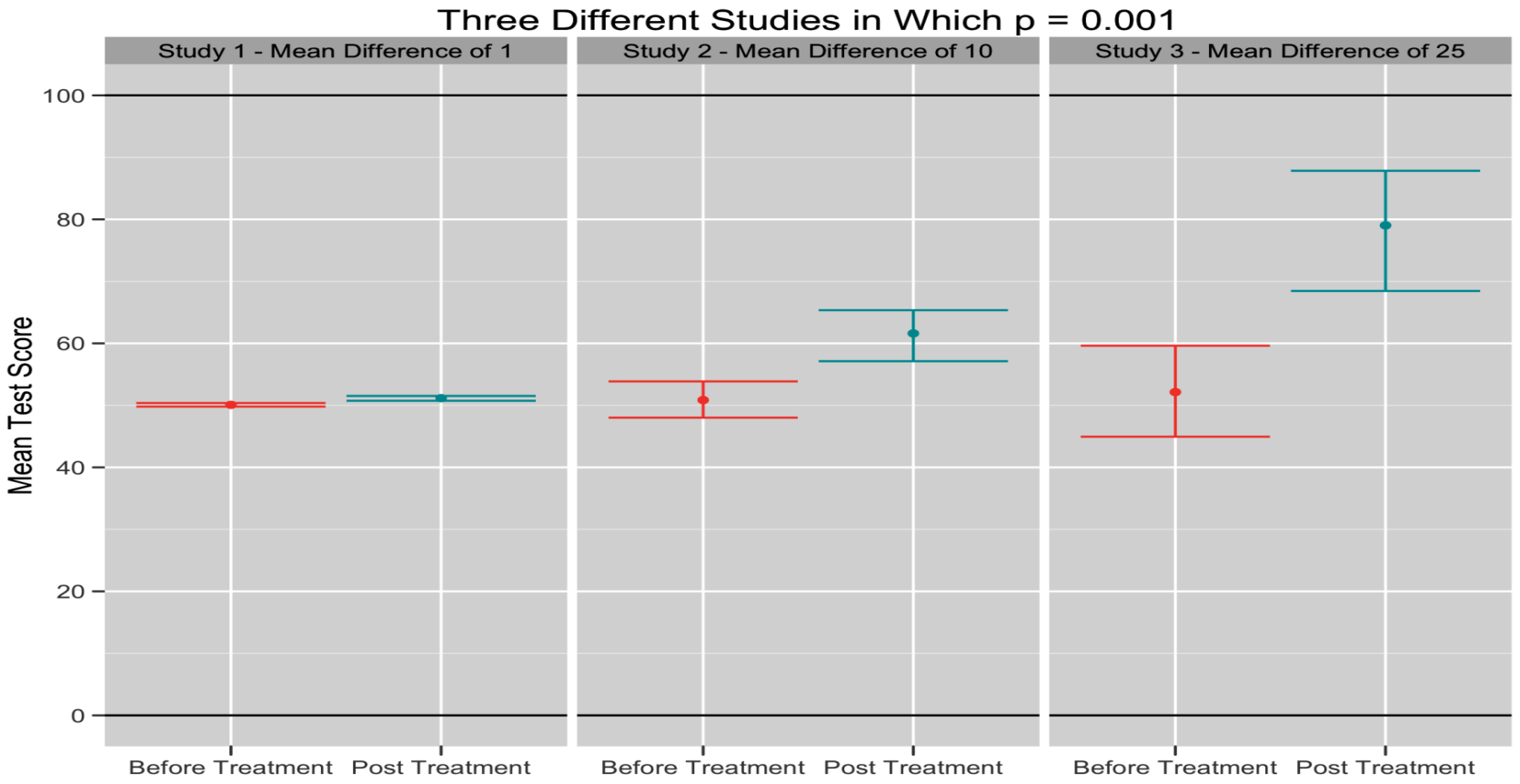




Anderson-Darling normality test
A = 0.7495, p-value = 0.0477

Anderson-Darling normality test
A = 0.2845, p-value = 0.6156

Equal p-values are not all equal



Study 1: Non-meaningful difference, precisely measured.
Study 2: Moderate difference, moderate accuracy.
Study 3: Important difference, poorly measured.
P-values collapse two-dimensional data (mean/sd) down to one dimension

http://www.johnmyleswhite.com/notebook/2012/05/14/criticism-3-of-nhst-essential-information-is-lost-when-transforming-2d-data-into-a-1d-measure/

A Simulation Study

- 800 farms (i.e. n=800) and 120 survey questions (i.e. G=120) in each dataset
- i^{th} farm is generated from a $Bernoulli(1,\ p_i)$
- $ln(\frac{p_i}{1-p_i}) = \beta_0 + \sum_{g=1}^{G} \mathbf{x_{i,g}^T}\beta_{\mathbf{g}}$, where β_0 is the intercept, $\mathbf{x_{i,g}}$ is a three dimentional indication vector for question answer and $\beta_{\mathbf{g}}$
- $\gamma = 0.1, 0.25, 0.5, 1$ and 2 and $\beta_0 = -\frac{40}{3}\gamma$

$$\beta_{\mathbf{g}} = (1, 0, -1) \times \gamma, g = 1, ..., 40,$$

$$\beta_{\mathbf{g}} = (1, 0, 0) \times \gamma, g = 41, ..., 80.$$

$$\beta_{\mathbf{g}} = (0, 0, 0) \times \gamma, g = 81, ..., 120.$$

Simulation Study Result

Simulation study result with various values of coefficient γ. Reported are mean and standard deviation of AUC for both methods and mean difference

Coefficient γ	Group Lasso (mean±sd)	Logistic Regression (mean±sd)
0.1	0.57 ± 0.03	0.54 ± 0.06
0.25	0.71 ± 0.02	0.64 ± 0.04
0.5	0.91 ± 0.03	0.78 ± 0.03
1	0.92 ± 0.01	0.82 ± 0.02
2	0.95 ± 0.01	0.84 ± 0.02

Challenges

- Communication
- Approve or change
- 360 degree data
- “Small p” and “Big n”
- Actionable questions
- Short term v.s long term
- Development to implementation

Iowa State


Snedecor
Hall
Statistics Department
Statistical Laboratory
Center for Survey Statistics and
Methodology

DuPont Pioneer

Pioneer employs statisticians!

DuPont Pioneer

谢谢！
Thank you!