Playing a Game in Statistical Computing

Martin J. Mohlenkamp

Department of Mathematics

OHIO
UNIVERSITY

December 3, 2014

Abstract

The French Federation of Mathematical Games put out its fifth Competitive Mathematical Game in November 2013 with entries due April 30, 2014. The problem is statistical in nature and has a computing component. The Math department has a shortage of Statisticians, so in Fall 2013 I was tapped to teach MATH 4/5530 Statistical Computing in Spring 2014. As a motivation for the students and myself to learn the material for this course, I used the game and its data set as a theme. In this talk I will describe the game problem, how it was used as a pedagogical device in the class, our entry in the game, and the results.

Outline

Checking an Industrial Process: Scenerio

A company produces 1000 metal cylinders a day for 10 days.
The Safety Authority will take a random sample of 100 of these 10,000 cylinders (uniform, without replacement).
Each cylinder will be cut into 5 layers, and each layer cut into 80 cells.
From these 400 cells, the Safety Authority will sample 10 (uniform, without replacement). In total, the Safety Authority tests 1000 cells.
The percentage Nickel (Ni) and percentage Chromium (Cr) is determined in each cell. For each cell that violates the specification intervals

$$\mathrm{Ni} \in I_1 = [6.94, 9.10] \qquad \text{and} \qquad \mathrm{Cr} \in J_1 = [16.95, 19.10]\,,$$

the Safety Authority fines the company €1,000,000. The number of cells that violate the (stricter) specification intervals

$$\mathrm{Ni} \in I_2 = [7, 9] \qquad \text{and} \qquad \mathrm{Cr} \in J_2 = [17, 19]$$

is counted; if there are more than 50 (5%) then the Safety Authority fines the company an additional €1,000,000.

Checking an Industrial Process: Data

The company first takes their own sample of the percentage Ni and Cr at some locations on some cylinders. The system of splitting a cylinder into layers and cells is the same as the Safety Authority's, but the selection of which cylinders and cells is not. This data was provided to us in a spreadsheet, with 4000 total samples.

variable	meaning	Domain
d	day	$\{1, 2, \ldots, 9, 10\}$
c	cylinder within day	$\{1, 2, \ldots, 999, 1000\}$
x	x coordinate of cell	$\{-5, -4, \ldots, 3, 4\}$
y	y coordinate of cell	$\{-5, -4, \ldots, 3, 4\}$
z	z coordinate of cell	$\{0, 1, 2, 3, 4\}$
Ni	%Ni	$[0, 100]$ (near 8)
Cr	%Cr	$[0, 100]$ (near 18)

To stay on the cylinder, $(x + 1/2)^2 + (y + 1/2)^2 \leq 25$.

Checking an Industrial Process: Question

What is the expected value of the fine?

The full game is at
`http://scmsa.eu/archives/SCM_FFJM_Competitive_Game_2013_2014.pdf`.

MATH 4/5530 Statistical Computing

Catalog Description: Introduction to computational statistics; Monte Carlo methods, bootstrap, data partitioning methods, EM algorithm, probability density estimation, Markov Chain Monte Carlo methods.

Desired Learning Outcomes: Students will be able to:

- Generate distributions by various methods.
- Use computer-intensive methods for estimation and hypotheses testing.
- Conduct data analysis using one or more major statistical models.

- Taught in a computer lab.
- Homework/journal and project based; tests not appropriate.
- 8 undergraduates and 6 graduate students enrolled.

MATH 4/5530 Statistical Computing

The course suffers from being a seemingly disconnected set of topics. We used the game and its data set for 13 of 28 "new content" days to connect the topics together, including

- importing data
- data summaries
- data selection and subsetting
- basic plotting
- linear fitting
- nonlinear least-squares
- generating samples
- Monte Carlo estimation
- bootstrapping
- the jackknife
- advanced plotting

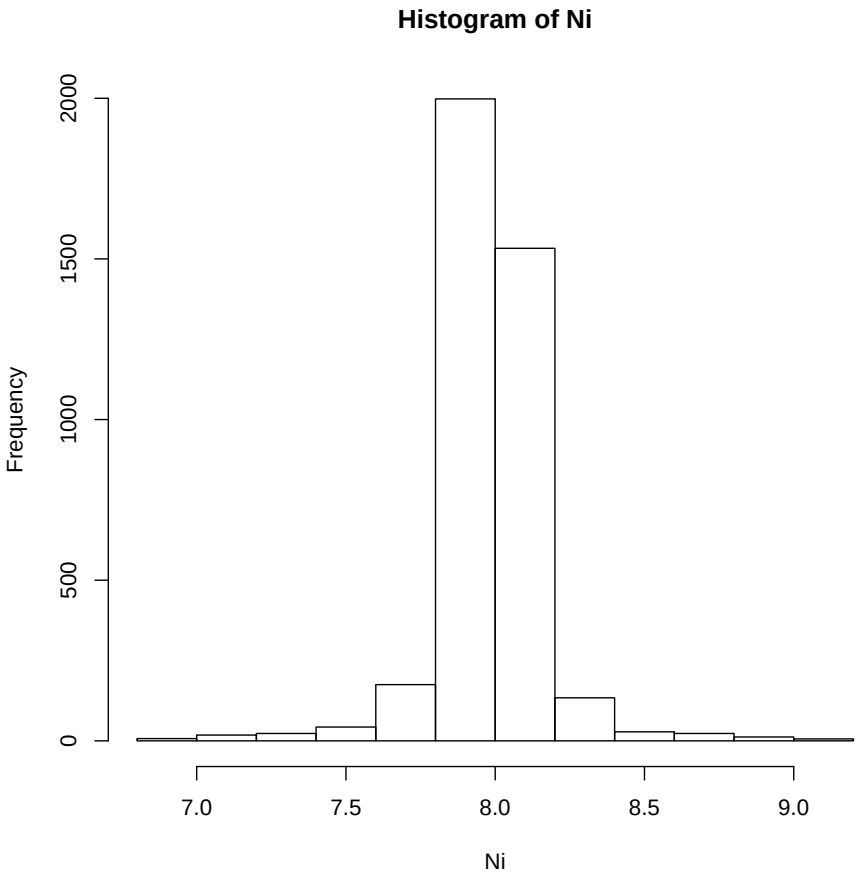
MATH 4/5530 Statistical Computing

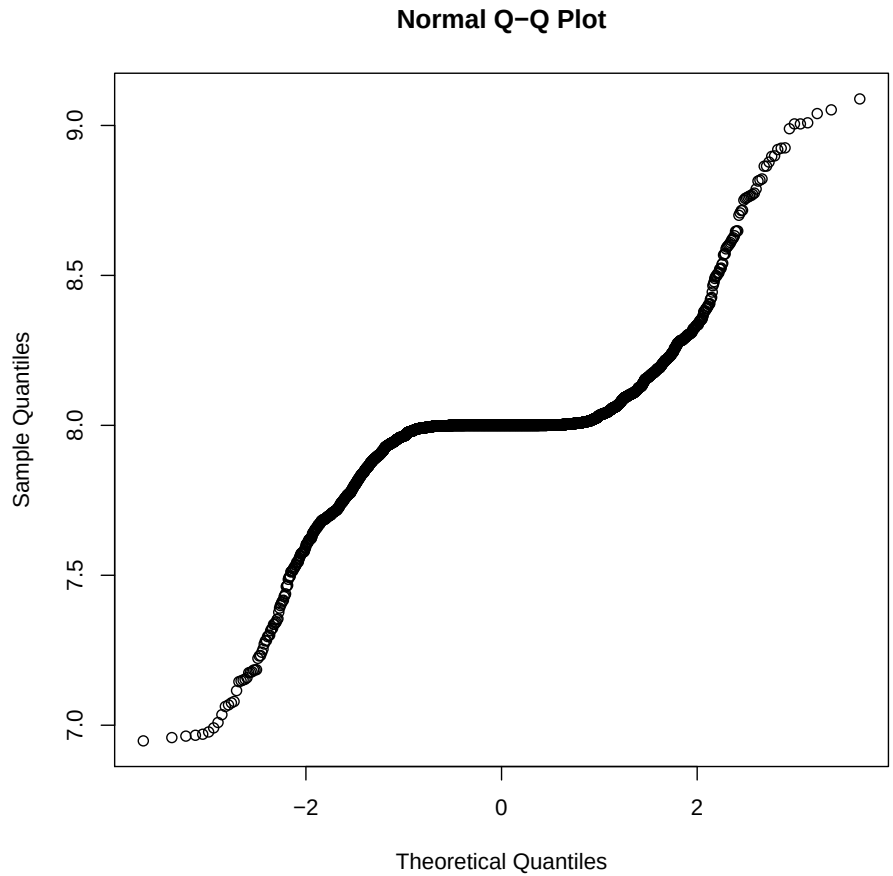
Other features of this class:

- 0% lecture.
- Textbook-free.
- Programming in R.
- Worked collaboratively in the Sagemath cloud `cloud.sagemath.com`.
- Included journals, projects, presentations, and reports.
- Full info at
  `http://www.ohio.edu/people/mohlenka/20142/4530-5530/`

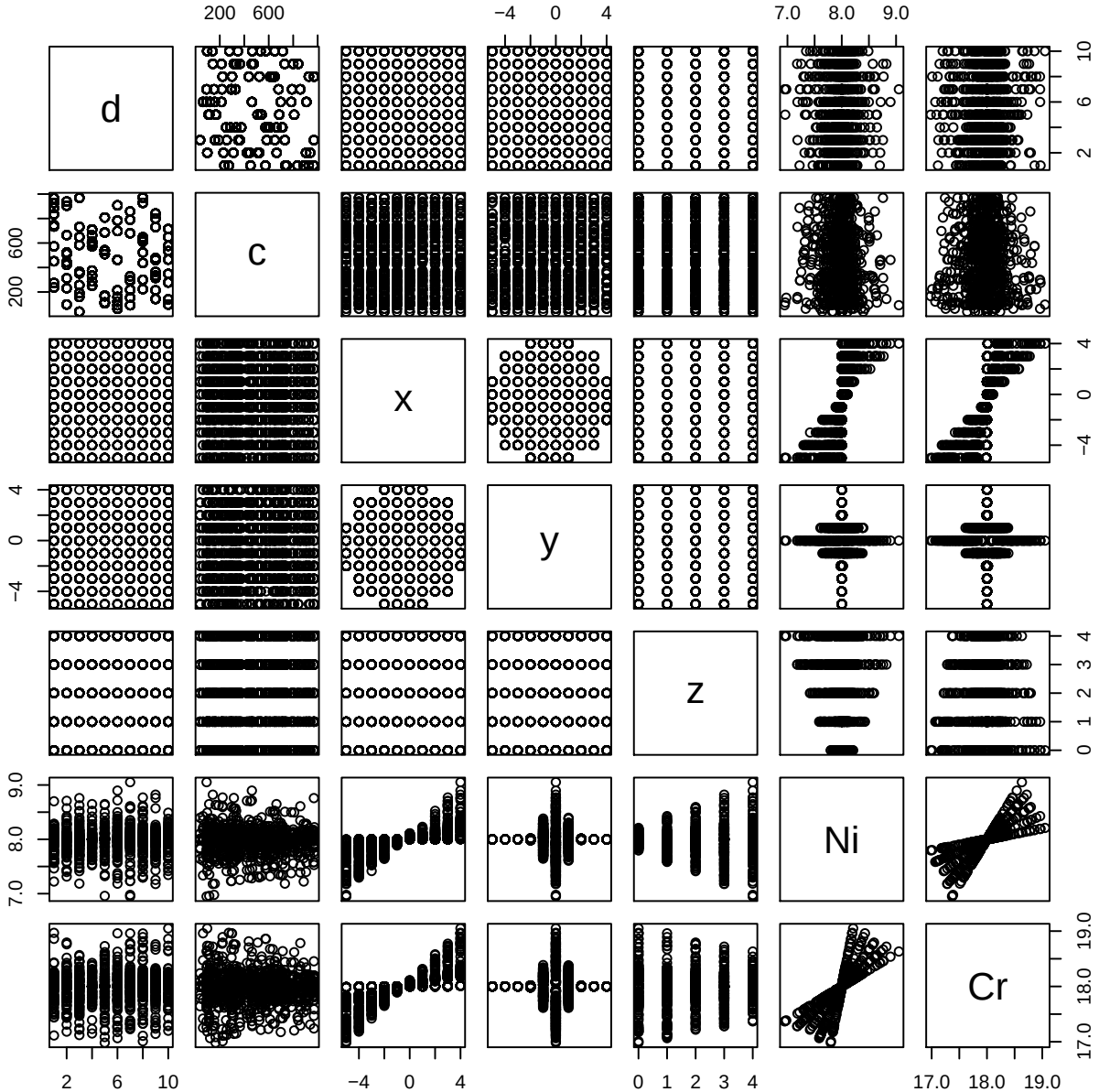
Masters student S. Elaine Hale continued with the game for her final project and she and I submitted a solution.

A First Look at the Data: no fines, not normal



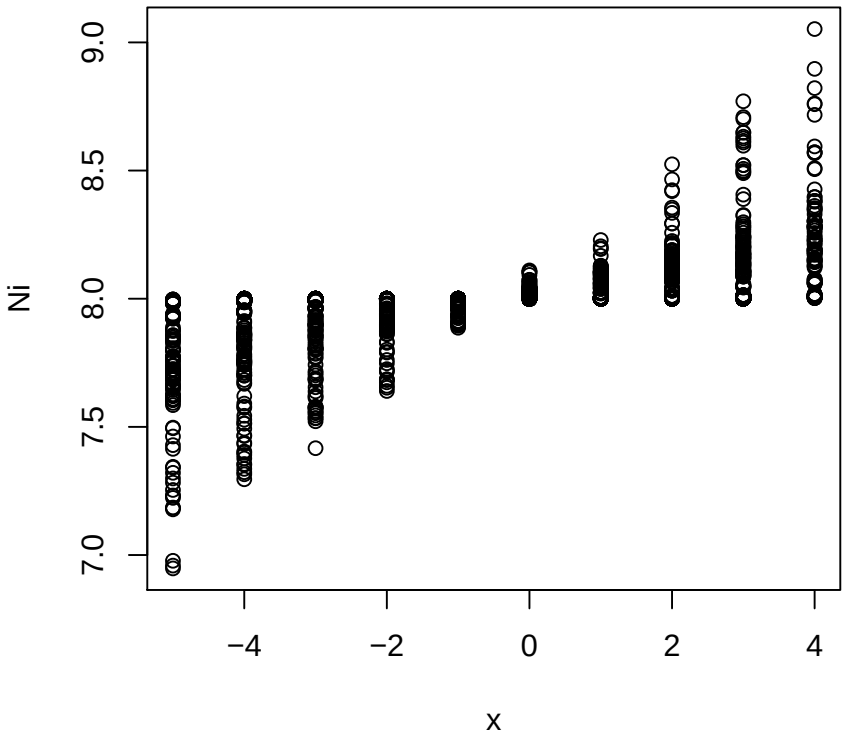


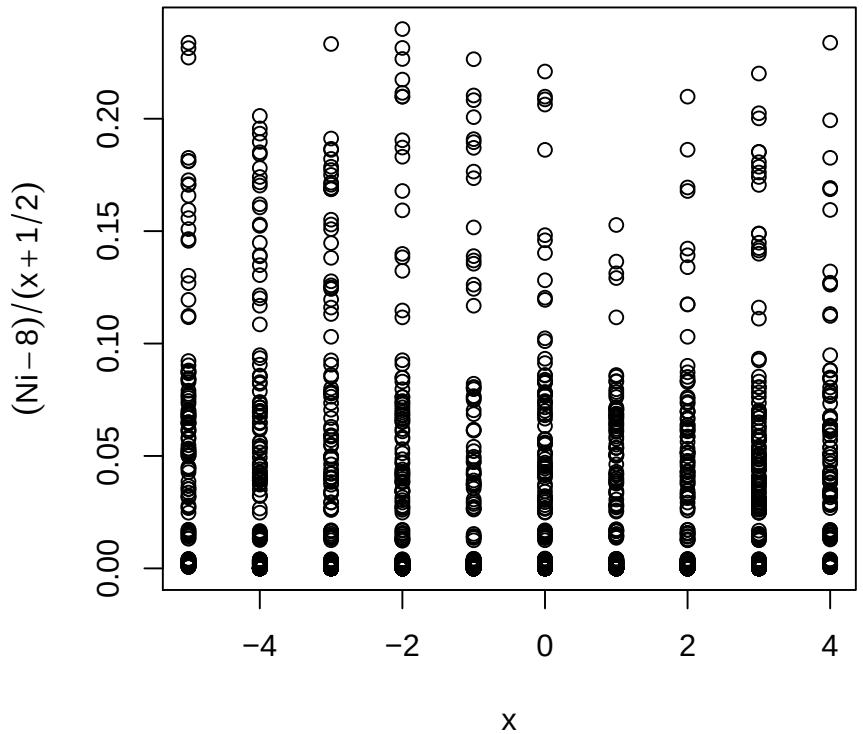
Initial Data Analysis: Pairs()



Ni dependence on x

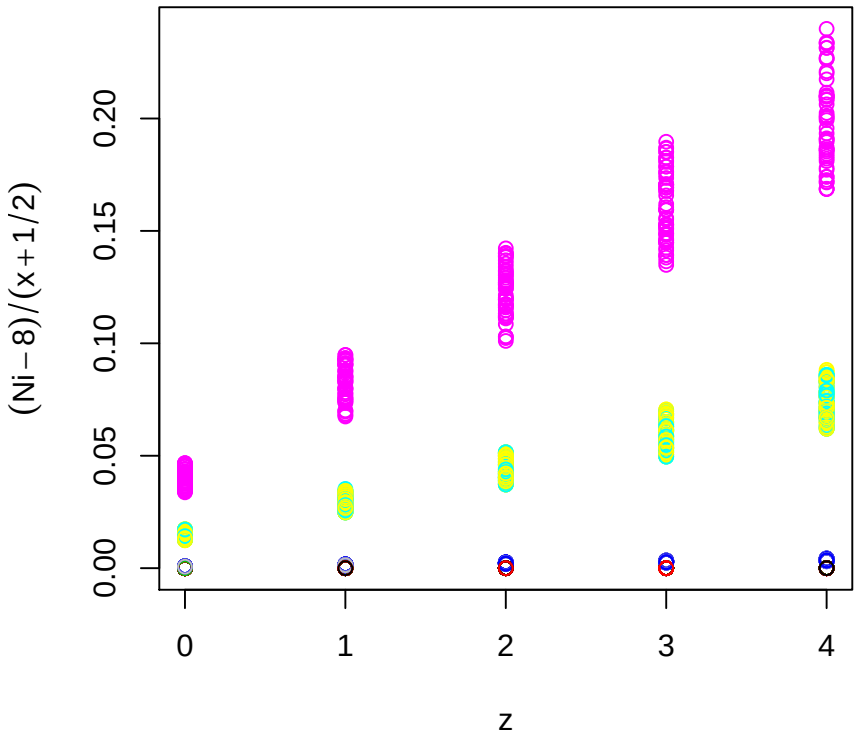
We see a linear relationship. The lines seem to cross at $(8, -1/2)$. To remove the dependence on x, we subtracted 8 and divided by $x + 1/2$.

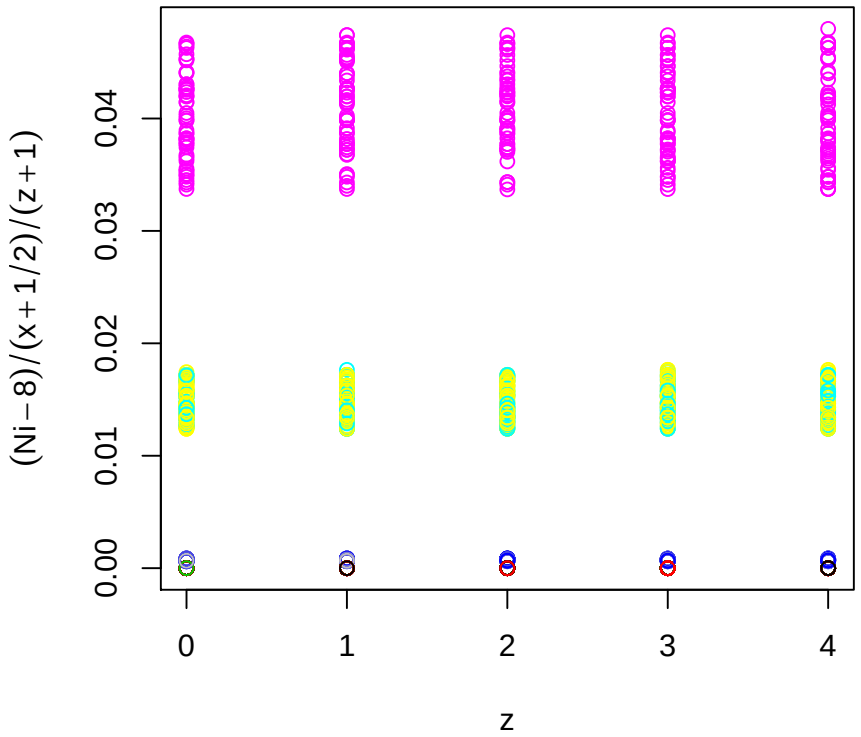




Ni dependence on z

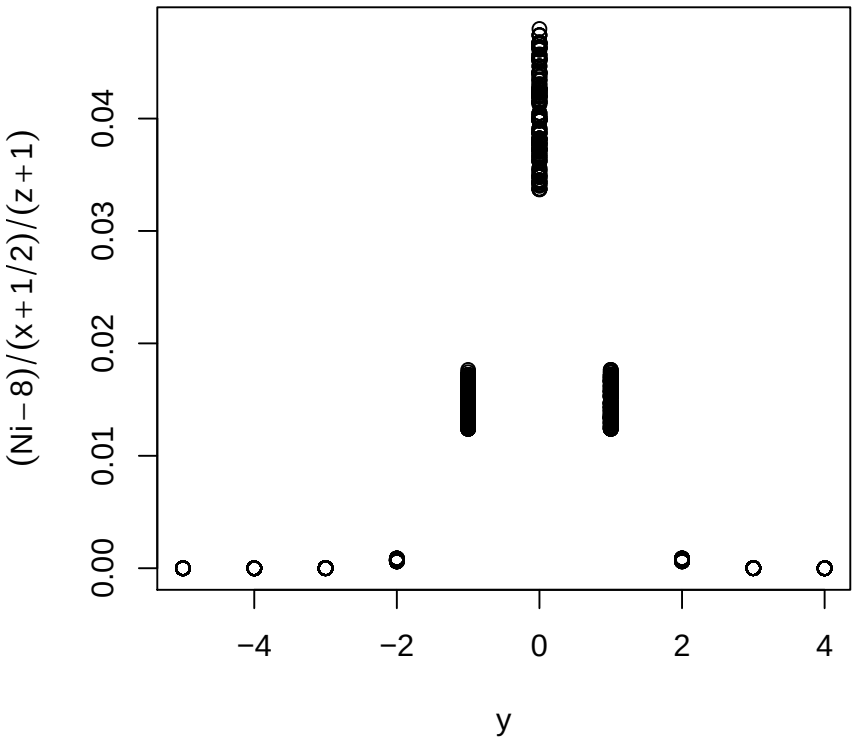
We plot z versus $(Ni - 8)/(x + 1/2)$ and can see a linear relationship. When fitting a line to this data restricting to $y = 0$, we noticed that the intercept and slope are similar, which suggests dependence proportional to $z + 1$. To remove the dependence on z, we divided by $(z + 1)$.

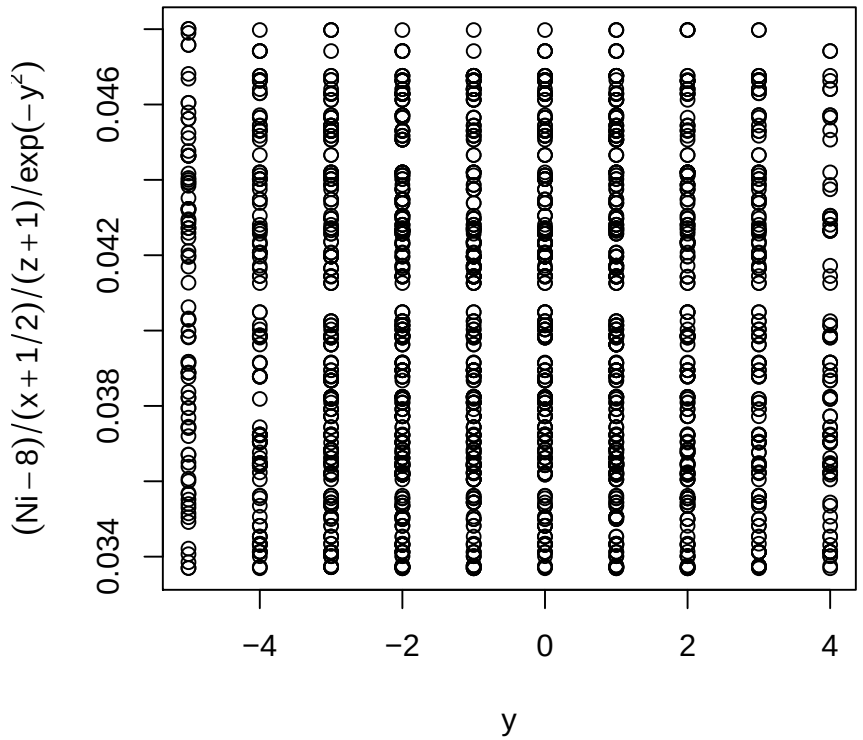




Ni dependence on y

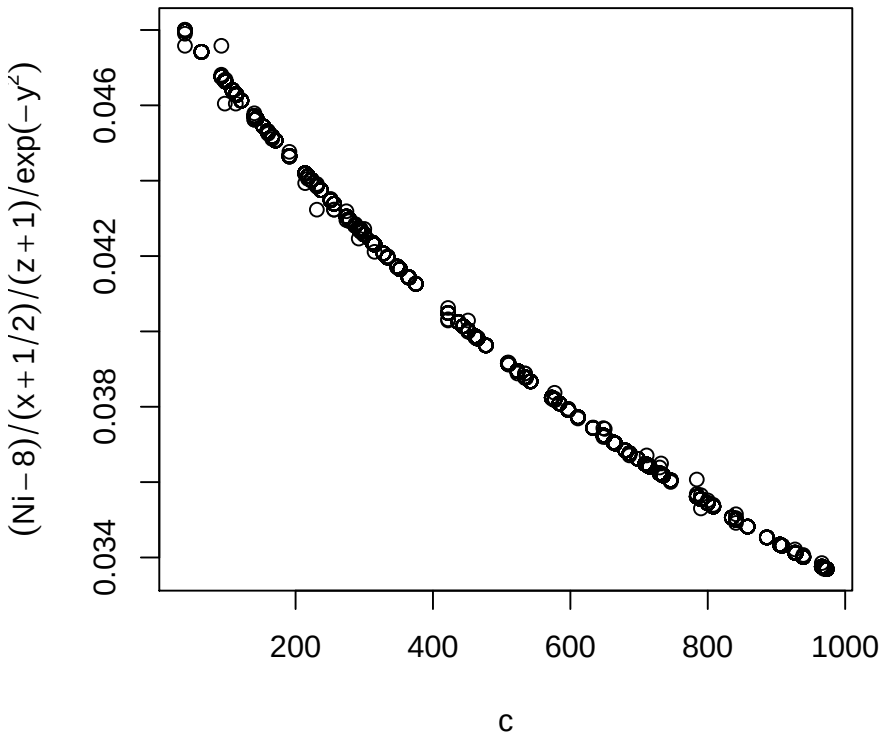
We plot y versus $(Ni - 8)/(x + 1/2)/(z + 1)$. We tried several functional forms using the nls() function in R and found $\exp(-y^2)$ to be an excellent fit. To remove the dependence on y, we divided by $\exp(-y^2)$.

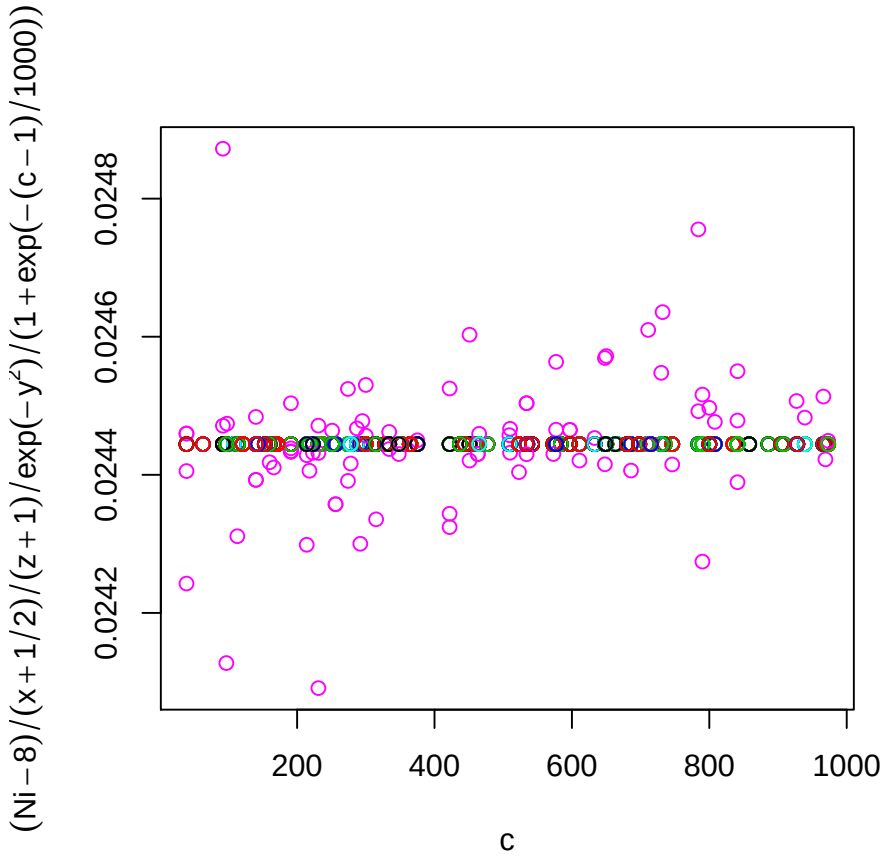




Ni dependence on c

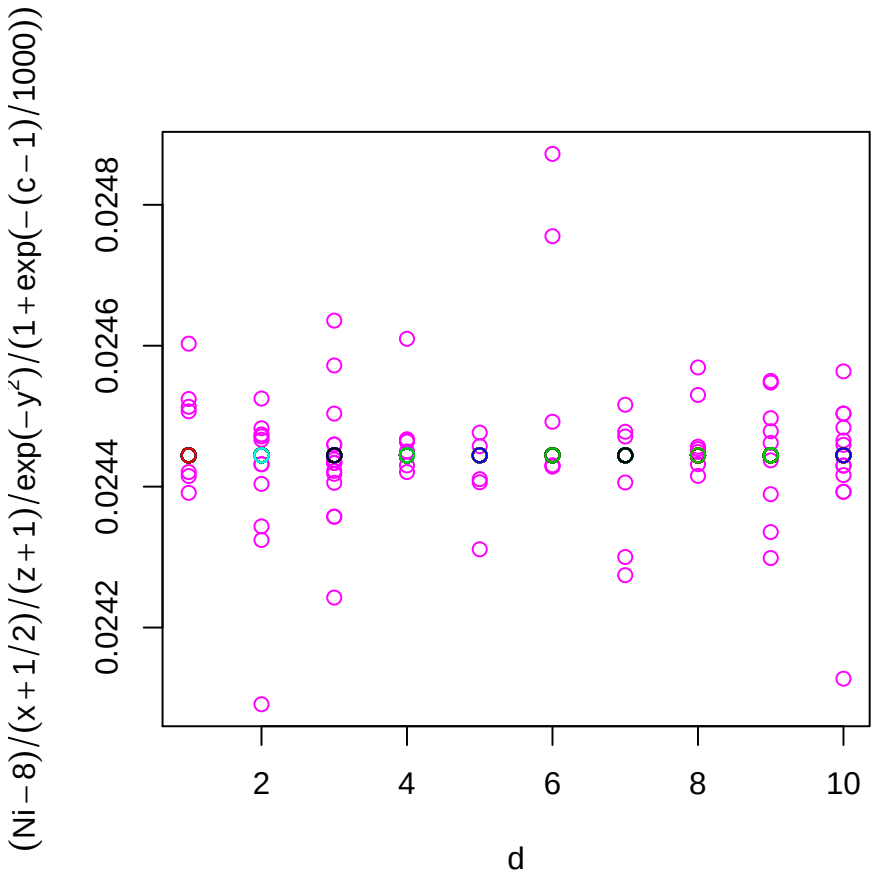
We plot c versus $(Ni - 8)/(x + 1/2)/(z + 1)/\exp(-y^2)$, and can see a slow non-linear decrease away from $c = 1$. Trying several functional forms using the nls() function in R, we found $1 + \exp(-(c - 1)/1000)$ to be an excellent fit. To remove the dependence on c, we divided by it.





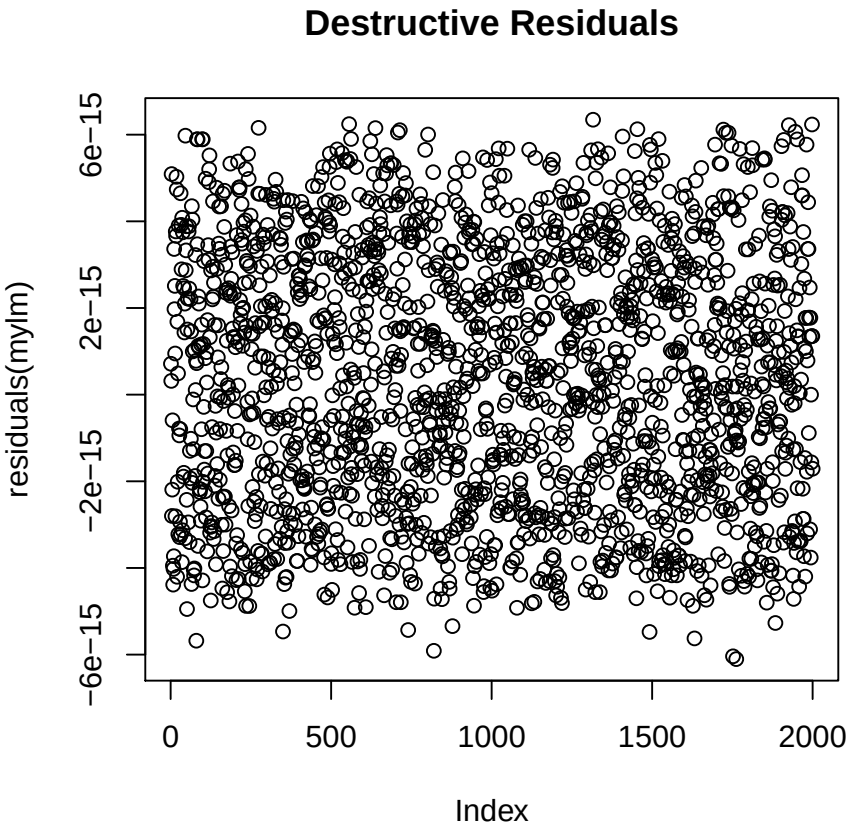
Ni dependence on d

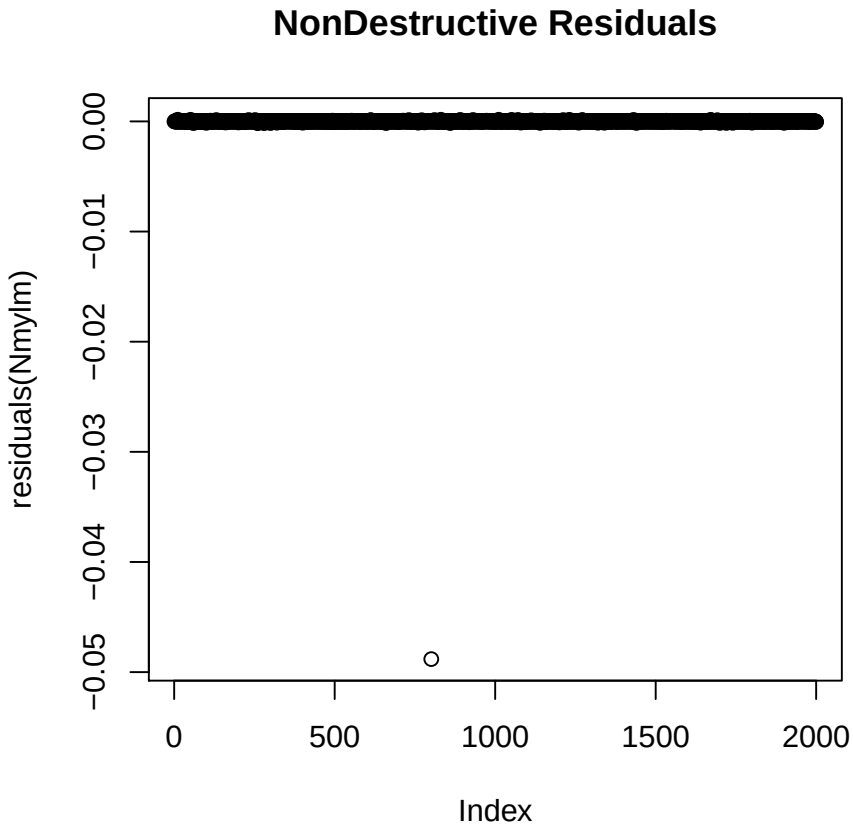
We plot d versus
$(Ni - 8)/(x + 1/2)/(z + 1)/\exp(-y^2)/(1 + \exp(-(c - 1)/1000))$
and can see no apparent dependence.



Ni Amplitude

We used the lm() function in R to compare $Ni - 8$ and $(x + 1/2)(z + 1)\exp(-y^2)(1 + \exp(-(c - 1)/1000))$.
From this, we found an amplitude of $11/450$.
Except for one outlier, the residuals are less than 7×10^{-15}.





Prediction Functions for Nickel and Chromium

Based on our dependencies and amplitude, we found the following prediction formula for Ni:

$$\mathrm{Ni}(d, c, x, y, z) = 8 + \frac{11}{450}\left(x + \frac{1}{2}\right) e^{-y^2}(z+1)\left(1 + e^{-\frac{c-1}{1000}}\right).$$

Following the same method as for Ni, we get the following prediction formula for Cr:

$$\mathrm{Cr}(d, c, x, y, z) = 18 + \frac{11}{900}\left(x + \frac{1}{2}\right) e^{-y^2}(10 - z)\left(1 + e^{-\frac{c-1}{1000}}\right).$$

Expected Fine

Since we know the value of Ni and Cr in every cell, we can compute

$c \in$	number of violations of I_1 or J_1	I_2 or J_2
$[1, 76]$	2	4
$[77, 96]$	1	4
$[97, 200]$	0	4
$[201, 1000]$	0	0

Selecting a single cell, the probability it violates I_1 or J_1 is

$$p_1 = 0.00043$$

and the probability it violates I_2 or J_2 is

$$p_2 = 0.00201\,.$$

Semi-Analytic Prediction of the Expected Fine

If we make a simplifying assumption:

> *The Safety Authority samples 1000 cells uniformly and with replacement from among the* 400×1000 *cells in one day's production.*

then the expected fine is €1,000,000 times

$$1000p_1 + \left(1 - \sum_{k=0}^{50} B(k; 1000, p_2)\right)$$

where $B(k; 1000, p_2)$ is the probability of k successes in a Binomial distribution with 1000 trials and probability of success p_2.

The second term is negligible so the fine is €1,000,000p_1 = €430,000.

Monte Carlo Prediction of the Expected Fine

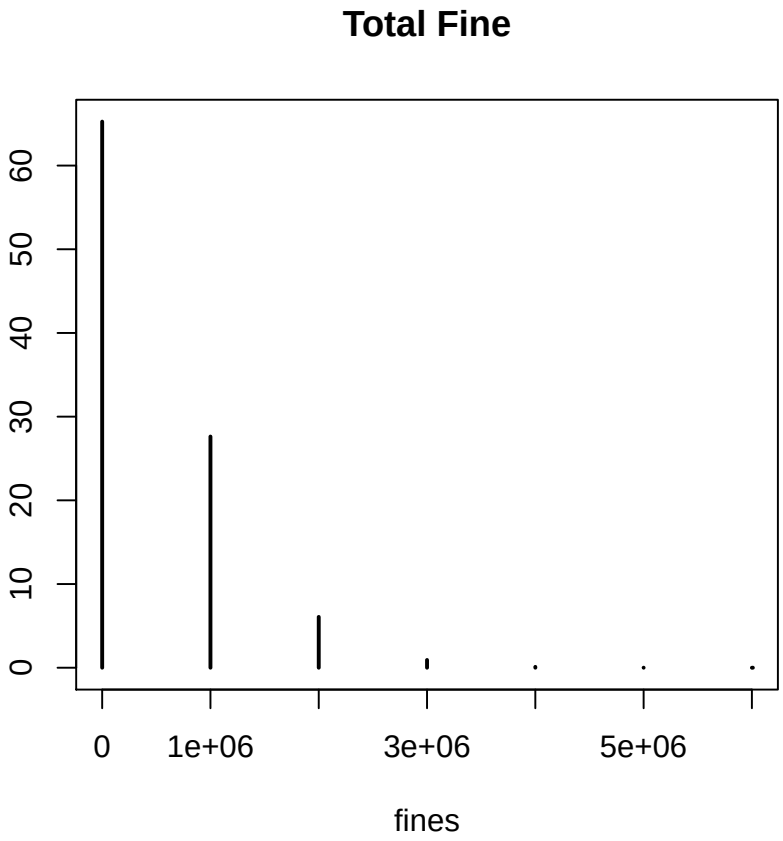
Reproduce the sampling and fining method of the Safety Authority 50,000 times and find the average to get an expected fine of about €429,732.

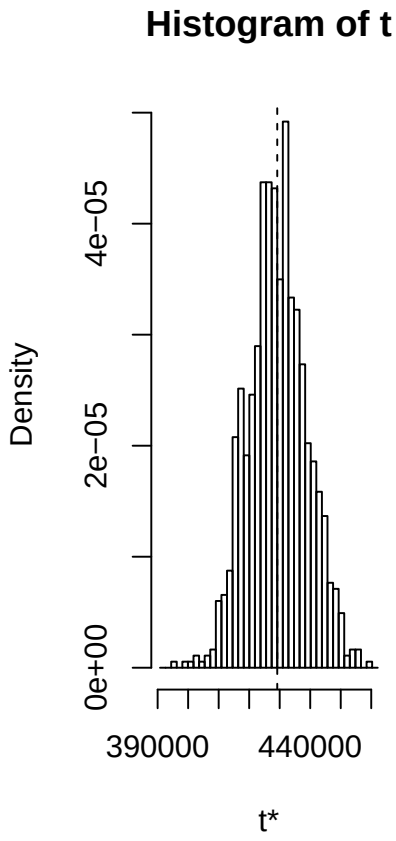
1. Take 100 samples uniformly without replacement from $1, 2, \ldots, 10000$ and interpret the results as $c + 1000(d - 1)$ to determine d and c.
2. For each c, if $200 < c$ do nothing; otherwise take 10 samples uniformly without replacement from the array of length 400:

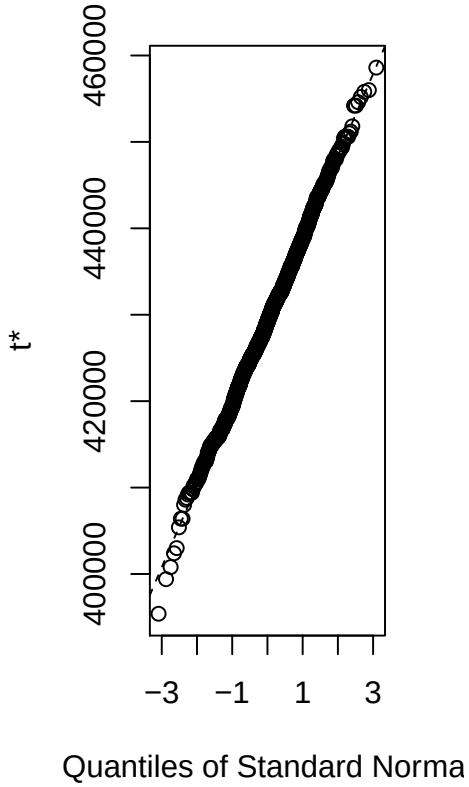
   if $1 \leq c \leq 76$ use $[2, 2, 1, 1, 0, \ldots, 0]$ or
   if $76 < c \leq 96$ use $[2, 1, 1, 1, 0, \ldots, 0]$ or
   if $96 < c \leq 200$ use $[1, 1, 1, 1, 0, \ldots 0]$.

   Count and accumulate the number of 2's and the number of nonzeros.
3. For each 2, add a fine of €1,000,000.
   If more than 50 nonzeros were obtained, add a fine of €1,000,000.

Bootstrapping







- Resample repeatedly from the Monte Carlo data.
- Form a histogram to find confidence intervals, etc.
- We obtained 95% confidence interval [€427,879, €431,588].

Results

- In the game, S. Elaine Hale and I won first place in the "group" category. Our entry and other information is available at `http://www.ohio.edu/people/mohlenka/20142/4530-5530/game/`
- In the class, I had fun and learned a lot.
- Many, but not all, students seemed to enjoy the class and learn a lot.

The game this year is *Uncertainties in GPS Positioning* and is available at `http://scmsa.eu/archives/SCM_FFJM_Competitive_Game_2014_2015.pdf`