BIOSTAT M280/BIOMATH 280/STAT M230: Statistical Computing

Tue/Thu 1:00pm-2:50pm, CHS 41-235

Instructor: Dr. Hua Zhou, `huazhou@ucla.edu`

1 Lecture 1: Jan 5

Today

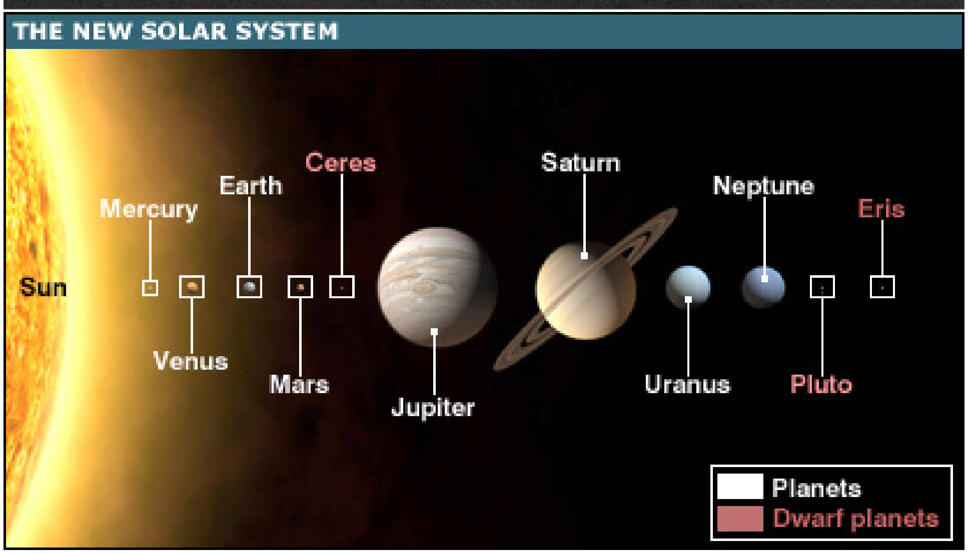
- Introduction and course logistics
- Computer storage and arithmetic
- Homework (not graded): fill out a short survey at `https://www.surveymonkey.com/r/G8BCVVM`
- Homework (not graded): Read the Introduction and Foundations of *Advanced R* by Hadley Wickham `http://adv-r.had.co.nz`. Install R. Try examples while you read the book. Do the quizzes.

What is statistics?

- People collect data in order to answer certain questions. (Bio)statisticians's job is to help make sense of data.
- Statistics, the science of *data analysis*, is the applied mathematics in the 21st century.
- Read papers *The future of data analysis* by John Tukey (`http://hua-zhou.github.io/teaching/biostatm280-2016winter/readings/Tukey61FutureDataAnalysis.pdf`) and *50 years of data sicence* by David Donoho (`http://hua-zhou.github.`

io/teaching/biostatm280-2016winter/readings/Donoho15FiftyYearsDataScience.pdf).

How Gauss became famous?



- 1801, *Dr. Carl Friedrich Gauss*, 24; proved Fundamental Theorem of Algebra; wrote the book *Disquisitiones Arithmetic*, which is still being studied today.

- 1801, Jan 1-Feb 11 (41 days), astronomer Piazzi observed Ceres (a dwarf planet), which was then lost behind sun.

- 1801, Aug–Sep, futile search by top astronomers; Laplace claimed it unsolvable.

- 1801, Oct–Nov, Gauss did calculations by *method of least squares*.
- 1801, Dec 31, astronomer von Zach relocated Ceres according to Gauss' calculation.
- 1802, *Summarische Übersicht der Bestimmung der Bahnen der beiden neuen Hauptplaneten angewandten Methoden*, considered the origin of linear algebra.
- 1807, Professor of Astronomy and Director of Göttingen Observatory in remainder of his life.
- 1809, *Theoria motus corporum coelestium in sectionibus conicis solum ambientium* (Theory of motion of the celestial bodies moving in conic sections around the Sun); birth of the Gaussian (normal) distribution, as an attempt to rationalize the method of least squares.
- 1810, Laplace consolidated the importance of Gaussian distribution by proving the central limit theorem.
- 1829, Gauss-Markov Theorem. Under Gaussian error assumption (actually only uncorrelated and homoscedastic needed), least square solution is the best linear unbiased estimate (BLUE), i.e., it has the smallest variance and thus MSE among all linear unbiased estimators. Note other estimators such as the James-Stein estimator may have smaller MSE, but they are *nonlinear*.

For more details of the story

- `http://www.keplersdiscovery.com/Asteroid.html`
- Teets and Whitehead (1999)

ARTICLES

The Discovery of Ceres: How Gauss Became Famous

DONALD TEETS
KAREN WHITEHEAD
South Dakota School of Mines and Technology
Rapid City, SD 57701

"The Duke of Brunswick has discovered more in his country than a planet: a super-terrestrial spirit in a human body."

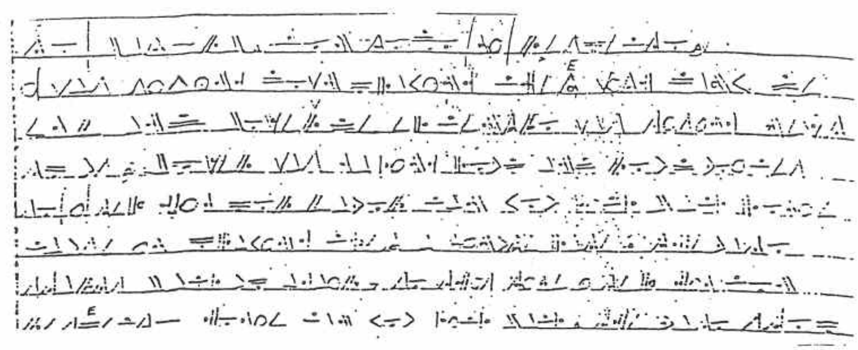
These words, attributed to Laplace in 1801, refer to the accomplishment of Carl Friedrich Gauss in computing the orbit of the newly discovered planetoid *Ceres Ferdinandea* from extremely limited data. Indeed, although Gauss had already achieved some fame among mathematicians, it was his work on the Ceres orbit that "made Gauss a European celebrity—this a consequence of the popular appeal which astronomy has always enjoyed..." **[2]**. The story of Gauss's work on this problem is a good one and is often told in biographical sketches of Gauss (e.g., **[2]**, **[3]**, **[6]**), but the mathematical details of how he solved the problem are invariably omitted from such historical works. We are left to wonder, how did he do it? *Just how did Gauss*

- Stigler (1986) gives a more comprehensive account from statisticians's view.

Gauss' story

- Motivated by a real problem.
- Heuristic solution: method of least squares.
- Solution readily verifiable: Ceres was re-discovered!
- *Algorithmic development*: linear algebra, Gaussian elimination, FFT (fast Fourier transform).
- Theoretical justification: Gaussian distribution, Gauss-Markov theorem.

A sampler by Marc Coram



ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS
NOBLER IN THE MIND TO SUFFER THE SLINGS AND ARROWS OF OUTRAGEOUS
FORTUNE OR TO TAKE ARMS AGAINST A SEA OF TROUBLES AND BY OPPOSING END

100 ER ENOHDLAE OHDLO UOZEOUNORU O UOZEO HD OITO HEOQSET IUROFHE HENO ITORUZAEN
200 ES ELOHRNDE OHRNO UOVEOULOSU O UOVEO HR OITO HEOQAET IUSOPHE HELO ITOSUVDEL
300 ES ELOHANDE OHANO UOVEOULOSU O UOVEO HA OITO HEOQRET IUSOFHE HELO ITOSUVDEL
400 ES ELOHINME OHINO UOVEOULOSU O UOVEO HI OATO HEOQRET AUSOWHE HELO ATOSUVKEL
500 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
600 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
900 ES ELOHANME OHANO UODEOULOSU O UODEO HA OITO HEOQRET IUSOWHE HELO ITOSUDMEL
1000 IS ILOHANMI OHANO RODIORLOSR O RODIO HA OETO HIOQUIT ERSOWHI HILO ETOSRDHIL
1100 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1200 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1300 ISTILOHARMITOHAROT ODIO LOS TOT ODIOTHATOENOTHIOQUINTE SOWHITHILOTENOS DMIL
1400 ISTILOHAMRITOHAMOT OFIO LOS TOT OFIOTHATOENOTHIOQUINTE SOWHITHILOTENOS DMIL
1600 ESTEL HAMRET HAM TO CE OL SOT TO CE THAT IN THE QUENTIOS WHETHEL TIN SOCREL
1700 ESTEL HAMRET HAM TO BE OL SOT TO BE THAT IN THE QUENTIOS WHETHEL TIN SOBREL
1800 ESTER HAMLET HAM TO BE OR SOT TO BE THAT IN THE QUENTIOS WHETHER TIN SOBLER
1900 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
2000 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER

to bat-rb. con todo mi respeto. i was sitting down playing chess with danny de emf and boxer de el centro was sitting next to us. boxer was making loud and loud voices so i tell him por favor can you kick back homie cause im playing chess a minute later the vato starts back up again so this time i tell him con respecto homie can you kick back. the vato stop for a minute and he starts up again so i tell him check this out shut the f**k up cause im tired of your voice and if you got a problem with it we can go to celda and handle it. i really felt disrespected thats why i told him. anyways after i tell him that the next thing I know that vato slashes me and leaves. dy the time i figure im hit i try to get away but the c.o. is walking in my direction and he gets me right dy a celda. so i go to the hole. when im in the hole my home boys hit doxer so now "b" is also in the hole. while im in the hole im getting schoold wrong and

- A consulting project by Marc Coram (then a graduate student in statistics at Stanford); customer is a professor in political science, who wants to understand a cryptic message circulated in a state prison.
- Marc modeled letter sequence by a Markov chain (26×26 transition matrix) and estimated transition probabilities from *War and Peace*.

- Now each mapping σ yields a likelihood $f(\sigma)$ of the symbol sequence.

- Find the σ that maximizes f. Sample space is at least $26! = 4.0329 \times 10^{26}$. Combinatorial optimization – hard!

- *Metropolis sampling*: At each iteration, generate a new σ' by random transposition of two letters; accept σ' with probability $\min\left\{\frac{f(\sigma')}{f(\sigma)}, 1\right\}$.

Marc Coram's story

- Motivated by a real problem.

- Solution readily verifiable: we can read it!

- *Algorithm*: Metropolis is one of top 10 algorithms in the 20th century.

- Read Diaconis (2009) for more details.

What is this course about?

- Not a course on "packages and languages for data analysis". It does not answer questions such as "How to fit a linear mixed model in SAS, SPSS or R?"

- Not a programming course, although programming is *extremely* important and we do homework in R.

- This course is about "numerical methods in statistics". Our focus is on *algorithms*.

> The form of a mathematical expression and the way the expression should be evaluated in actual practice may be quite different.
>
> James Gentle

For a common numerical task in statistics, say the least squares solution $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$, we need to know which methods/algorithms are out there and what are their advantages and disadvantages. You will *fail* this course if you use

```
solve(t(X) %*% X) %*% t(X) %*% y
```

Using `lm(y ~ X)` is correct but is not the purpose of this course. We want to understand what is going on when calling the `lm()` function.

Science, 287 # 5454, Feb 4, 2000, p799

Algorithms for the Ages

"Great algorithms are the poetry of computation," says Francis Sullivan of the Institute for Defense Analyses' Center for Computing Sciences in Bowie, Maryland. He and Jack Dongarra of the University of Tennessee and Oak Ridge National Laboratory have put together a sampling that might have made Robert Frost beam with pride—had the poet been a computer jock. Their list of 10 algorithms having "the greatest influence on the development and practice of science and engineering in the 20th century" appears in the January/February issue of *Computing in Science & Engineering*. If you use a computer, some of these algorithms are no doubt crunching your data as you read this. The drum roll, please:

1946: The Metropolis Algorithm for Monte Carlo. Through the use of random processes, this algorithm offers an efficient way to stumble toward answers to problems that are too complicated to solve exactly.

1947: Simplex Method for Linear Programming. An elegant solution to a common problem in planning and decision-making.

1950: Krylov Subspace Iteration Method. A technique for rapidly solving the linear equations that abound in scientific computation.

1951: The Decompositional Approach to Matrix Computations. A suite of techniques for numerical linear algebra.

1957: The Fortran Optimizing Compiler. Turns high-level code into efficient computer-readable code.

1959: QR Algorithm for Computing Eigenvalues. Another crucial matrix operation made swift and practical.

1962: Quicksort Algorithms for Sorting. For the efficient handling of large databases.

1965: Fast Fourier Transform. Perhaps the most ubiquitous algorithm in use today, it breaks down waveforms (like sound) into periodic components.

1977: Integer Relation Detection. A fast method for spotting simple equations satisfied by collections of seemingly unrelated numbers.

1987: Fast Multipole Method. A breakthrough in dealing with the complexity of n-body calculations, applied in problems ranging from celestial mechanics to protein folding.

Syllabus

Check course website frequently for updates and announcements.
`http://hua-zhou.github.io/teaching/biostatm280-2016winter`
Lecture notes will be updated and posted after each lecture.

References

Diaconis, P. (2009). The Markov chain Monte Carlo revolution. *Bull. Amer. Math. Soc. (N.S.)*, 46(2):179–205.

Stigler, S. M. (1986). *The History of Statistics*. The Belknap Press of Harvard University Press, Cambridge, MA. The measurement of uncertainty before 1900.

Teets, D. and Whitehead, K. (1999). The discovery of Ceres: how Gauss became famous. *Math. Mag.*, 72(2):83–93.