Stat 550 Opening Lecture

David W Scott

Rice University

August 22, 2023

Plan

Returning to Houston at the end of this week. Will do a light zoom today and Thursday in the meantime to get up to speed.

Any logistical questions? Display syllabus.

Follow the URL there to find 1st homework assignment.

Introduction to an important topic in modern multivariate statistics. The course will survey topics in data analysis and visualization, multivariate density estimation, nonparametric regression, and applications. The course will provide a comprehensive theoretical introduction to various density estimators, including histograms, frequency polygons, kernel and series methods, nearest neighbor estimators, penalized-likelihood methods, and wavelets. Regression topics will focus on kernel smoothing and local polynomial algorithms. Both asymptotic and finite sample results will be considered and, in particular, modern cross-validation algorithms.

Emphasis will be given to multivariate extensions of univariate density estimators into two, three, four, and five dimensions. Computationally efficient algorithms will be introduced for these cases.

Applications covered include: use of density estimation for interactive exploratory data analysis; spatial data and mapping; clustering and discrimination; density grand tour; nonparametric and modal regression; hazard analysis; bootstrap; projection pursuit; optimal subspace search; and others.

Only an introductory background in probability and statistics is required, although some basic knowledge of classical multivariate statistical methods will be helpful. Students with particular research problems are especially welcomed. The instructor intends to accommodate students from other disciplines who are interested in this topic.

Grading (subject to change)

- 1. Homeworks, 50% (but not very much)
- 2. Participation (important in the 'real' world), 30%
- 3. Joint research paper and/or work in new book section, 20%

Notes on Grading:

- 1. Interaction during class to dig into material and relationships to other courses you've taken.
- 2. Presentation and discussion of selected homework solutions.
- 3. An exciting new feature is the hope that we can divide the class into several groups and do original research on a topic and submit for publication. An alternative might be to write new material not covered in the textbook.
- Textbook: Scott, D.W. (2015), *Multivariate Density Estimation: Theory, Practice, and Visualization, 2nd Edition*, John Wiley & Sons, Hoboken, NJ.

Fundamental Task for the Semester: Estimate an Unknown PDF $f(x)$ with Applications

- What is a parametric pdf estimator?

Fundamental Task for the Semester: Estimate an Unknown PDF $f(x)$ with Applications

- What is a parametric pdf estimator?
- We usually write $\hat{f}(x)$ as $\hat{f}_\theta(x)$

Fundamental Task for the Semester: Estimate an Unknown PDF $f(x)$ with Applications

- What is a parametric pdf estimator?
- We usually write $\hat{f}(x)$ as $\hat{f}_\theta(x)$
- What is a nonparametric pdf estimator?

Fundamental Task for the Semester: Estimate an Unknown PDF $f(x)$ with Applications

- What is a parametric pdf estimator?
- We usually write $\hat{f}(x)$ as $\hat{f}_\theta(x)$
- What is a nonparametric pdf estimator?
- Which kind is $\hat{f}(x) \sim N(\mu, \sigma^2)$?

Fundamental Task for the Semester: Estimate an Unknown PDF $f(x)$ with Applications

- What is a parametric pdf estimator?
- We usually write $\hat{f}(x)$ as $\hat{f}_\theta(x)$
- What is a nonparametric pdf estimator?
- Which kind is $\hat{f}(x) \sim N(\mu, \sigma^2)$?
- MLE's of $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are $\hat{\theta}_1 = \bar{X}$ and $\hat{\theta}_2 = S^2$

Fundamental Task for the Semester: Estimate an Unknown PDF $f(x)$ with Applications

- What is a parametric pdf estimator?
- We usually write $\hat{f}(x)$ as $\hat{f}_\theta(x)$
- What is a nonparametric pdf estimator?
- Which kind is $\hat{f}(x) \sim N(\mu, \sigma^2)$?
- MLE's of $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are $\hat{\theta}_1 = \bar{X}$ and $\hat{\theta}_2 = S^2$
- Which kind is a histogram (with fixed bin width h)?

Fundamental Task for the Semester: Estimate an Unknown PDF $f(x)$ with Applications

- What is a parametric pdf estimator?
- We usually write $\hat{f}(x)$ as $\hat{f}_\theta(x)$
- What is a nonparametric pdf estimator?
- Which kind is $\hat{f}(x) \sim N(\mu, \sigma^2)$?
- MLE's of $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are $\hat{\theta}_1 = \bar{X}$ and $\hat{\theta}_2 = S^2$
- Which kind is a histogram (with fixed bin width h)?
- We also write $\hat{f}(x) \sim \hat{f}_h(x)$.

Fundamental Task for the Semester: Estimate an Unknown PDF $f(x)$ with Applications

- What is a parametric pdf estimator?
- We usually write $\hat{f}(x)$ as $\hat{f}_\theta(x)$
- What is a nonparametric pdf estimator?
- Which kind is $\hat{f}(x) \sim N(\mu, \sigma^2)$?
- MLE's of $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are $\hat{\theta}_1 = \bar{X}$ and $\hat{\theta}_2 = S^2$
- Which kind is a histogram (with fixed bin width h)?
- We also write $\hat{f}(x) \sim \hat{f}_h(x)$.
- If we start histogram bins at the origin, then h is the only (unknown) parameter.

Fundamental Task for the Semester: Estimate an Unknown PDF $f(x)$ with Applications

- What is a parametric pdf estimator?
- We usually write $\hat{f}(x)$ as $\hat{f}_\theta(x)$
- What is a nonparametric pdf estimator?
- Which kind is $\hat{f}(x) \sim N(\mu, \sigma^2)$?
- MLE's of $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are $\hat{\theta}_1 = \bar{X}$ and $\hat{\theta}_2 = S^2$
- Which kind is a histogram (with fixed bin width h)?
- We also write $\hat{f}(x) \sim \hat{f}_h(x)$.
- If we start histogram bins at the origin, then h is the only (unknown) parameter.
- Is there an MLE for h? (Homework: Find $\hat{h}_{MLE}$)

Semester Task (cont'd): Estimate an Unknown PDF $f(x)$

- During this semester, our focus is on continuous (rather than discrete) data. (Discrete pretty easy?)
- Hence, we may assume a random sample of size n, $\{x_1, x_2, \ldots, x_n\}$, has no duplicate values when solving a theoretical (or practical) problem.
- When writing code, ties may occur due to finite precision, eg, $diff(sort(\boldsymbol{x}))$.
- These data should also be continuous, but may have multiple values of 0.
- Question: How does R's *hist()* function handle 0's? Can anyone try 'live'?

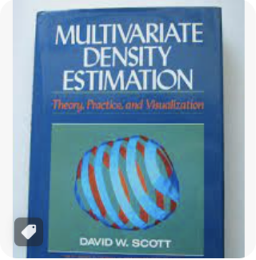
What is Statistics About (at its core)?

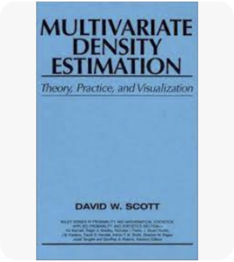
- Have you tried to explain what you do to your parents?
- What are the core elements of statistics?

Multivariate Probability Density Estimation (PDE)

- Author: David W Scott
- Publisher: John Wiley & Sons
- Second Edition 2015
- First Edition 1992 (which had color plates in middle)

- Classical PDE is the histogram (see Chapter 3)





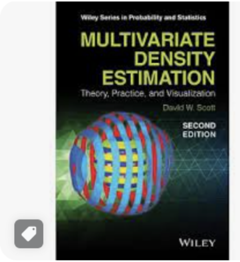
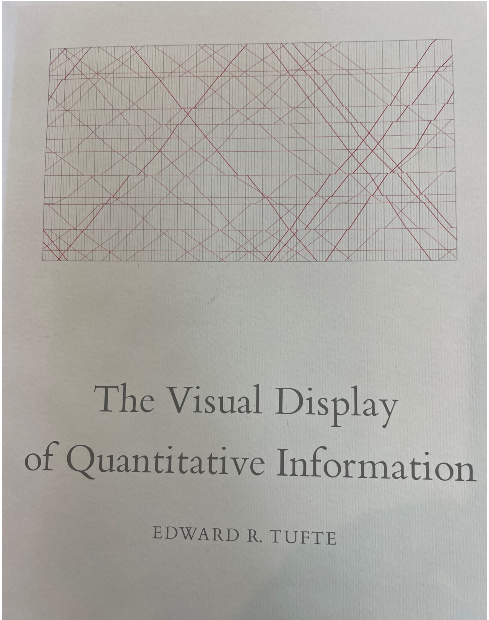
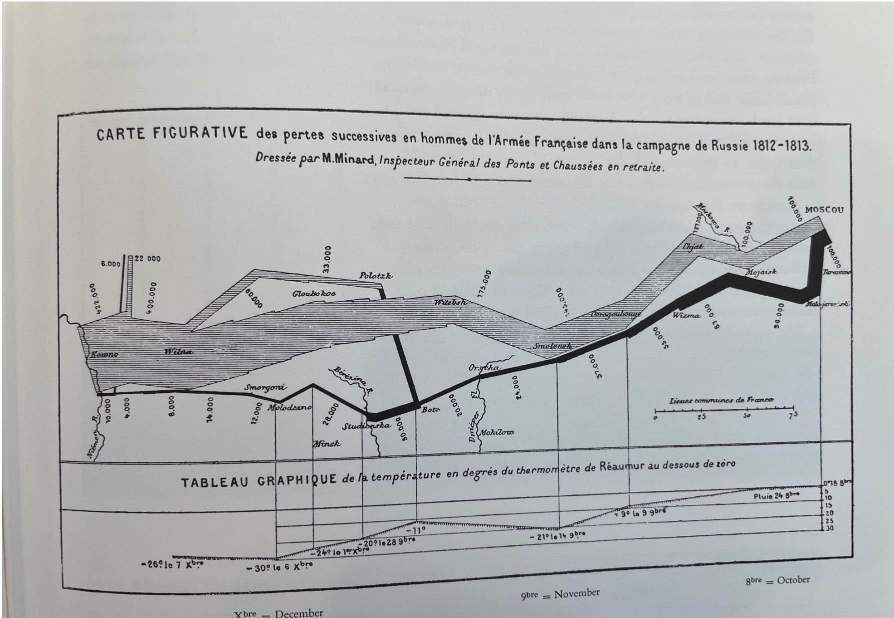


Table of Contents (MDE)

Edward Tufte's Lovely Books



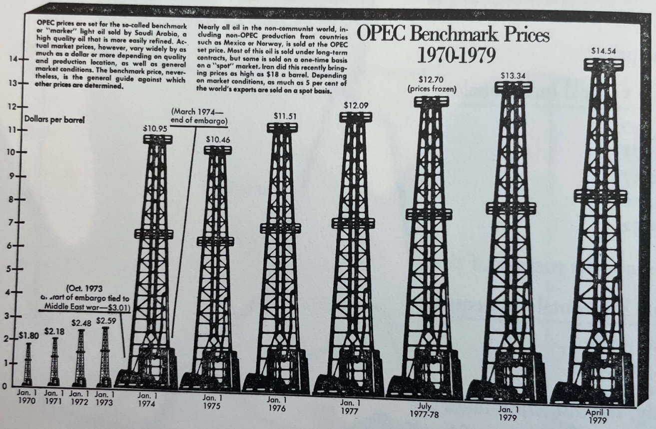
Repro: M. Minard's Napoleon's Russian Campaign 1812-13



Tufte's Examples *à la Lying With Statistics, Darrell Huff*

Graphics that convey an incorrect visual impression, eg.,

And an increase of 708 percent is shown as 6,700 percent, for a Lie Factor of 9.5:



Washingt
A-18.

All these accounts of oil prices made a second error, by showing the price of oil in inflated (current) dollars. The 1972 dollar was

Tufte's Most Famous (?) Idea: Data-to-Ink Ratio

Graphics with lots of 'ink' but little data are suspicious.

Data Density and Size of Data Matrix, Statistical Graphics in Selected Publications, Circa 1979–1980

	Data Density (Numbers per square inch)			Size of Data Matrix		
	median	minimum	maximum	median	minimum	maximum
Nature	48	3	362	177	15	3780
Journal of the Royal Statistical Society, B	27	4	115	200	10	1460
Science	21	5	44	109	26	316
Wall Street Journal	19	3	154	135	28	788
Fortune	18	5	31	96	42	156
The Times (London)	18	2	122	50	14	440
Journal of the American Statistical Association	17	4	167	150	46	1600
Asahi	13	2	113	29	15	472
New England Journal of Medicine	12	3	923	84	8	3600

The Economist	9	1	51	36	3	192
Le Monde	8	1	17	66	11	312
Psychological Bulletin	8	1	74	46	8	420
Journal of the American Medical Association	7	1	39	53	14	735
New York Times	7	1	13	35	6	580
Business Week	6	2	12	32	14	96
Newsweek	6	1	13	23	2	96
Annuaire Statistique de la France	6	1	25	96	12	540
Scientific American	5	1	69	46	14	652
Statistical Abstract of the United States	5	2	23	38	8	164
American Political Science Review	2	1	10	16	9	40
Pravda	0.2	0.1	1	5	4	20

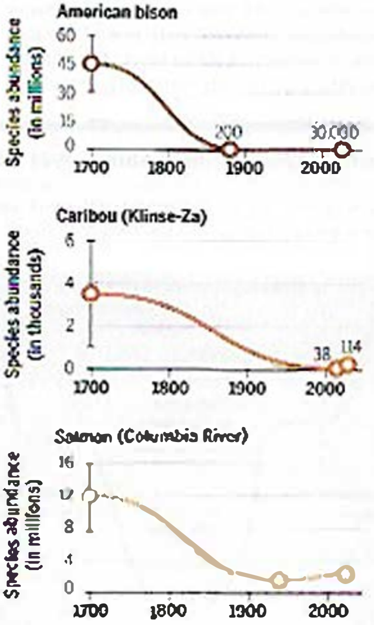
Let's Explore *Science* Magazine (3rd in Tufte's Table)

- On the good side of the data-to-ink scale are the weekly publications
- *Nature*, a British journal since 1869
- *Science*, published by AAAS since 1880

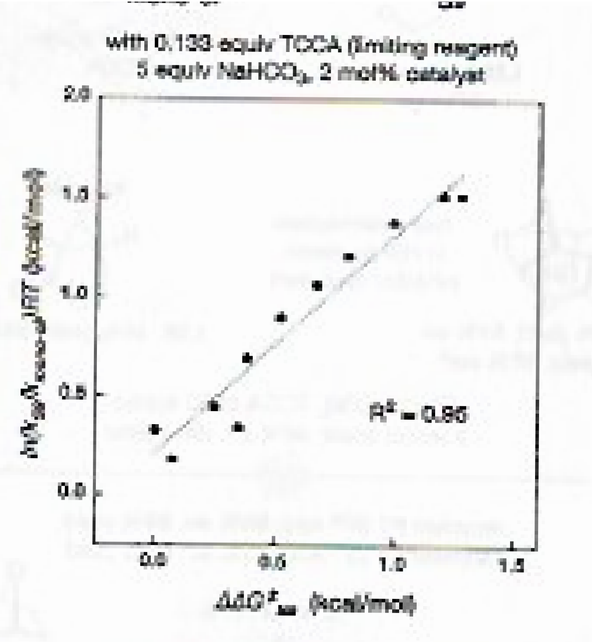
- Let's look at a sampling of the statistical techniques and graphics in the 5/19/2023 issue


Hybrid crops that never lose their vigor p. 684
The emergence of kissing in ancient societies p. 688
A wearable device restores thermal sensation in amputees p. 731
Science
$15
19 MAY 2023
science.org
AAAS
SHRINKING LAKES
Global water storage in large lakes is decreasing
pp. 693 & 743

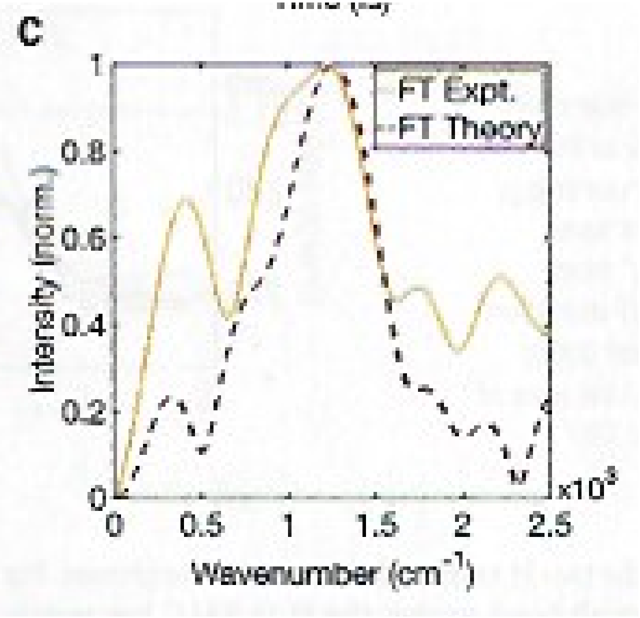
Culturally meaningful N. American species decline



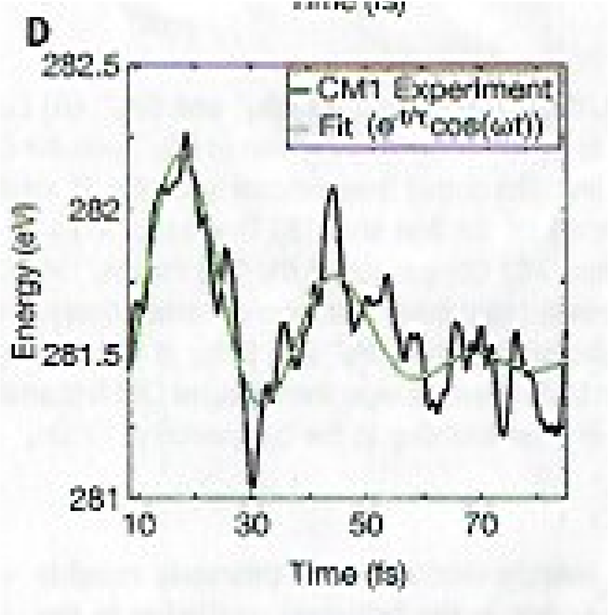
Intermolecular competition study



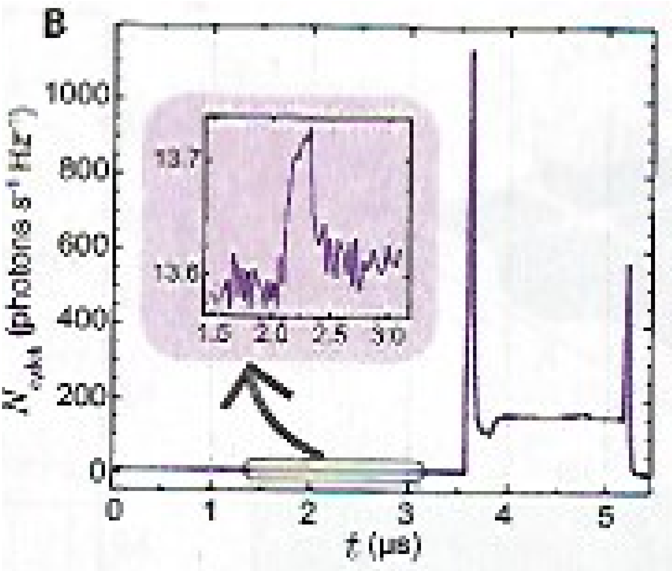
Experimental FT (solid) vs Theoretical Model (dotted)



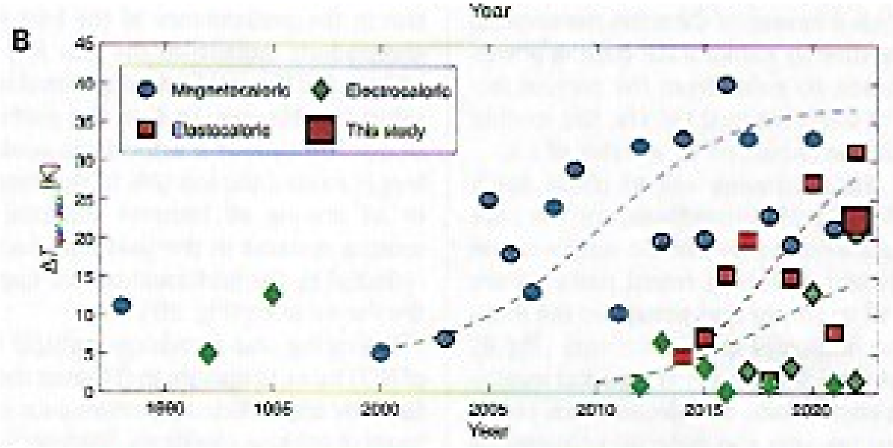
Another Experimental vs Theoretical Curve Comparison



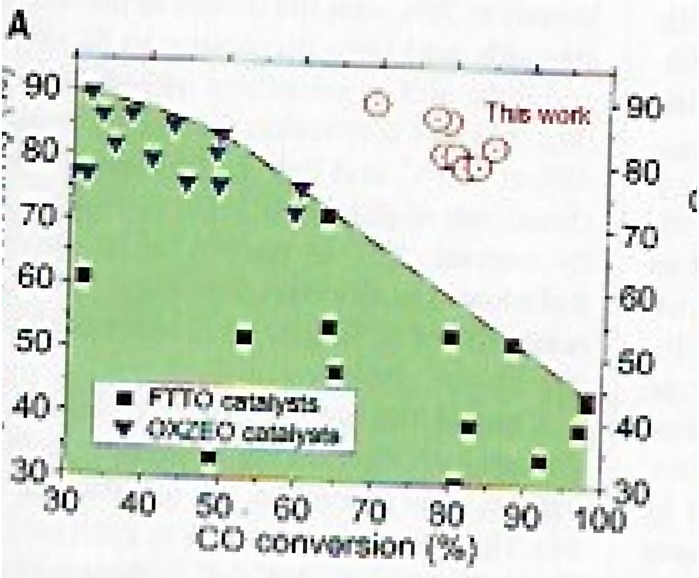
Bump Hunting in Power Spectrum



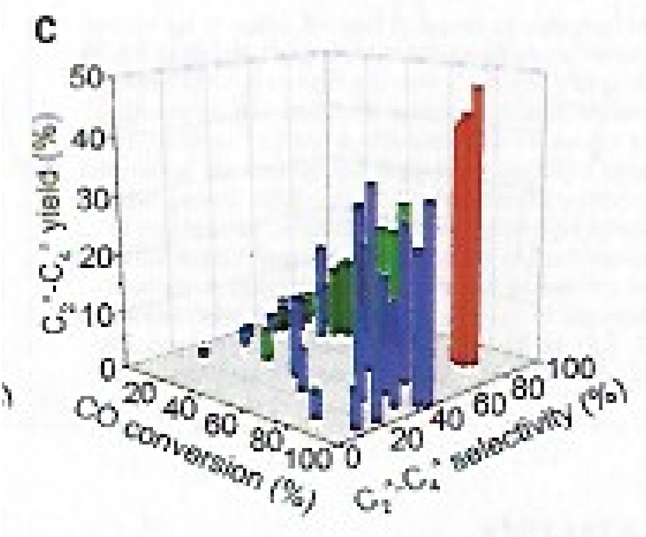
Maximum Power Generation w/ 3 Types



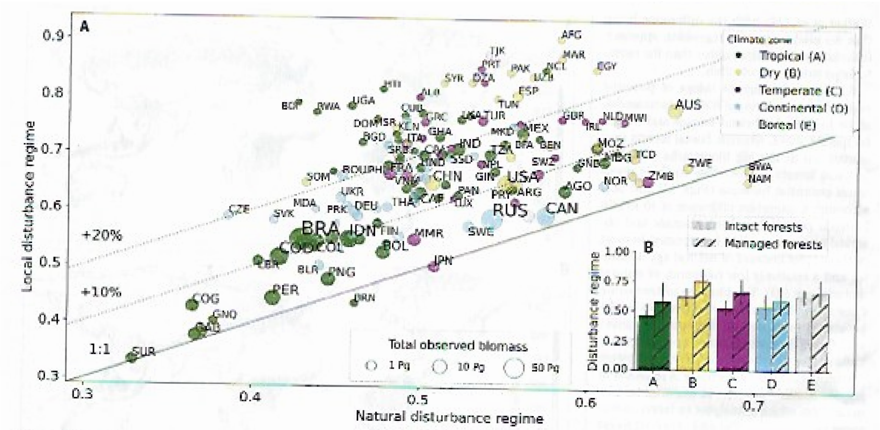
Catalytic Conversion of Syngas to Light Olefins (envelope)



Comparison of 3 Catalyts Performance



Forest Management of World Countries



Random Ideas

- How important is a random sample?

Random Ideas

- How important is a random sample?
- Data science often skirts over this in favor of 'large n' or using all the available data?

Random Ideas

- How important is a random sample?
- Data science often skirts over this in favor of 'large n' or using all the available data?
- What is the AI result? Bias?

Random Ideas

- How important is a random sample?
- Data science often skirts over this in favor of 'large n' or using all the available data?
- What is the AI result? Bias?
- Can the output of ChatGPT (Generative Pre-trained Transformer) be tilted to reduce bias? What do you think?

Random Ideas

- How important is a random sample?
- Data science often skirts over this in favor of 'large n' or using all the available data?
- What is the AI result? Bias?
- Can the output of ChatGPT (Generative Pre-trained Transformer) be tilted to reduce bias? What do you think?
- What can you do with a random sample?

Random Ideas

- How important is a random sample?
- Data science often skirts over this in favor of 'large n' or using all the available data?
- What is the AI result? Bias?
- Can the output of ChatGPT (Generative Pre-trained Transformer) be tilted to reduce bias? What do you think?
- What can you do with a random sample?
  - visualization that reflects reality: stem-and-leaf, box plot, histogram...

Random Ideas

- How important is a random sample?
- Data science often skirts over this in favor of 'large n' or using all the available data?
- What is the AI result? Bias?
- Can the output of ChatGPT (Generative Pre-trained Transformer) be tilted to reduce bias? What do you think?
- What can you do with a random sample?
  - visualization that reflects reality: stem-and-leaf, box plot, histogram...
  - testing $\qquad H_0 : X \sim F$

Why does MLE work?

- MLE gives values of θ that are 'best'?

Why does MLE work?

- MLE gives values of θ that are ‘best’?
- Presumably $\hat{\theta} \approx \theta_0$ must be good.

Why does MLE work?

- MLE gives values of θ that are 'best'?
- Presumably $\hat{\theta} \approx \theta_0$ must be good.
- What if $f_\theta(x)$ has the wrong parametric form?

Why does MLE work?

- MLE gives values of θ that are 'best'?
- Presumably $\hat{\theta} \approx \theta_0$ must be good.
- What if $f_\theta(x)$ has the wrong parametric form?
- As $n \to \infty$? variance? bias?

Why does MLE work?

- MLE gives values of θ that are 'best'?
- Presumably $\hat{\theta} \approx \theta_0$ must be good.
- What if $f_\theta(x)$ has the wrong parametric form?
- As $n \to \infty$? variance? bias?
- What does that actually mean about the quality of

$$\hat{f}_\theta \qquad \text{or} \qquad f_{\hat{\theta}} \qquad \text{at various } x's?$$

Why does MLE work?

- MLE gives values of θ that are 'best'?
- Presumably $\hat{\theta} \approx \theta_0$ must be good.
- What if $f_\theta(x)$ has the wrong parametric form?
- As $n \to \infty$? variance? bias?
- What does that actually mean about the quality of

$$\hat{f}_\theta \qquad \text{or} \qquad f_{\hat{\theta}} \qquad \text{at various } x's?$$

- Consider Gaussian model.

Normal Model

Memorize this amazingly useful identity:

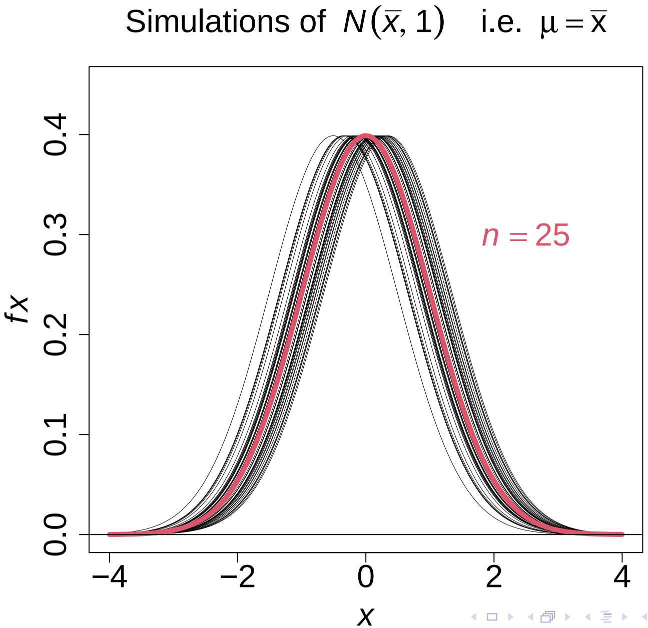
$$\int \phi(x|\mu_1, \sigma_1^2) \times \phi(x|\mu_2, \sigma_2^2)\, dx = \phi(0|\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)\,.$$

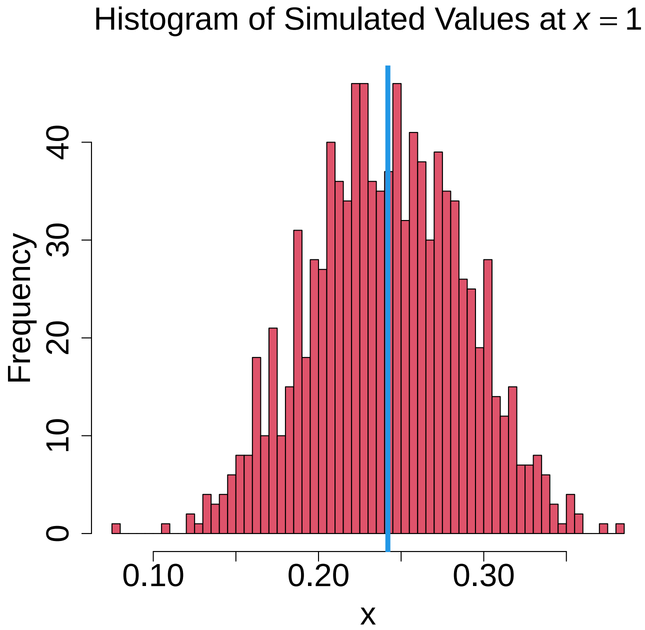
Multivariate version true: replace μ_k with $\boldsymbol{\mu}_k$ and σ_k^2 with Σ_k.

Normal model with known $\sigma = 1$.

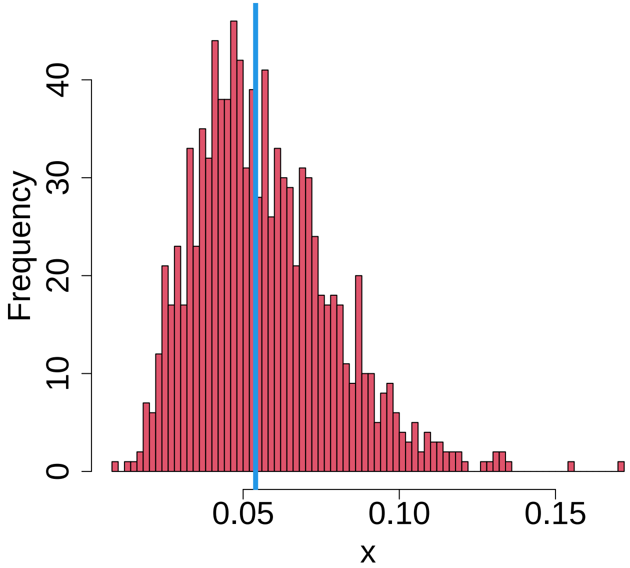
- For the normal model at a fixed point x:
  - The unknown to be estimated is $\phi(x|\mu, 1)$, call it θ_x
  - The MLE of θ_x is
$$\hat{\theta}_x = \phi(x|\bar{x}, 1)\,.$$
- Wow!!! Need to estimate $\theta_x = \phi(x, \mu, 1)$ at an infinite number of x values!!! Does infinite imply nonparametric?

How to Evaluate $\hat{\theta}$? Simulations!

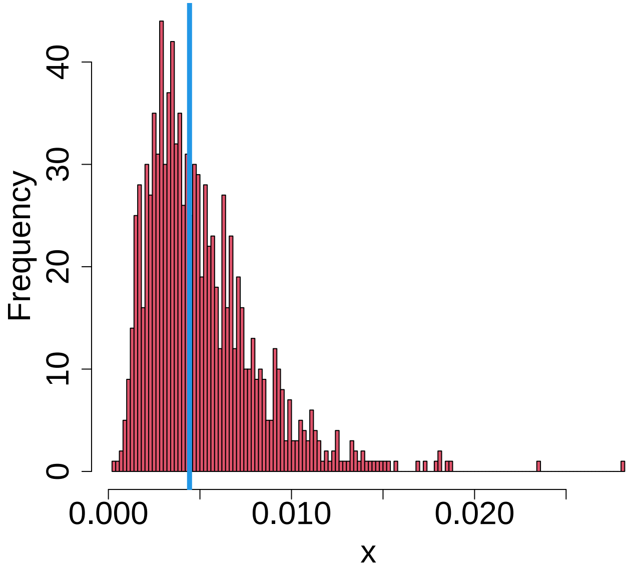



Histogram of Simulated Values at $x = 1$
Frequency
0
10
20
30
40
0.10
0.20
0.30
x

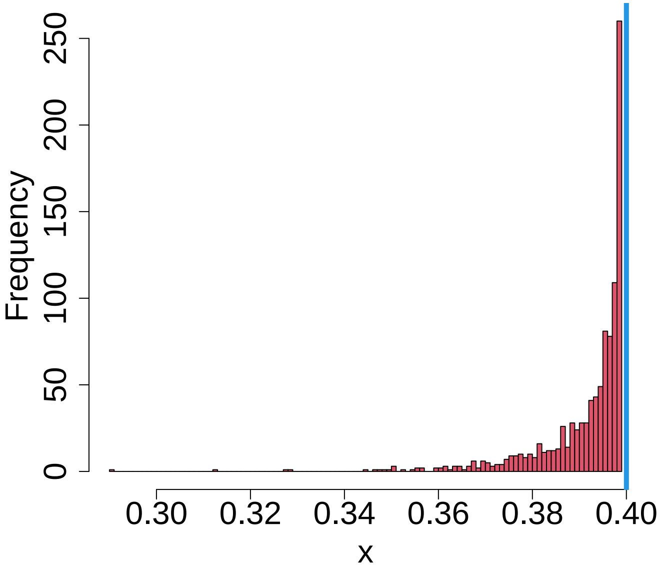
Histogram of Simulated Values at $x = 2$



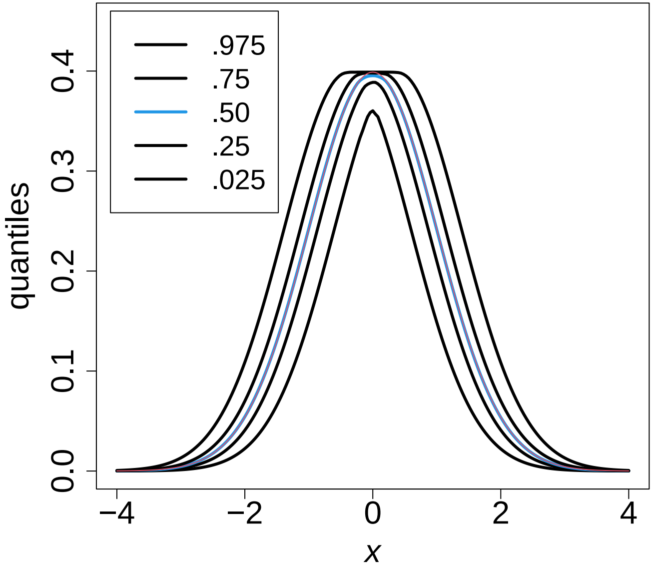
Histogram of Simulated Values at $x = 3$

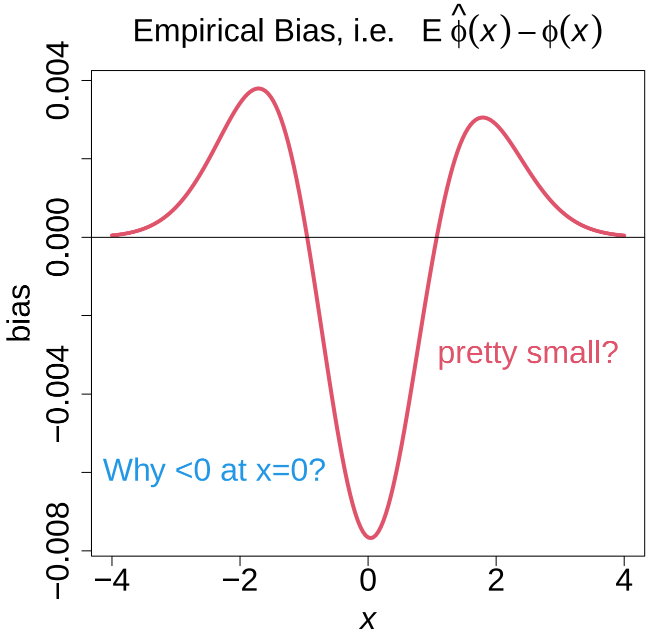


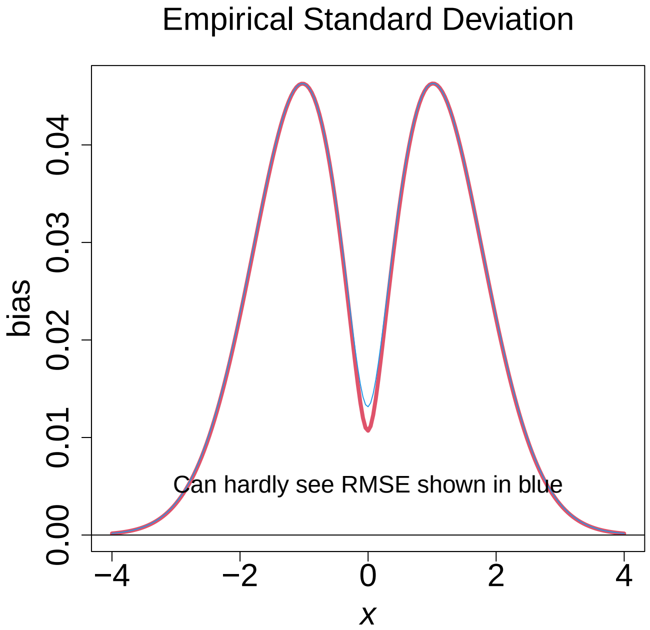
Histogram of Simulated Values at $x = 0$



Quantiles of 1000 Simulated Curves




Empirical Bias, i.e. $E\,\hat{\phi}(x) - \phi(x)$
pretty small?
Why <0 at x=0?
bias
0.004
0.000
−0.004
−0.008
−4
−2
0
2
4
x


Empirical Standard Deviation
bias
0.00
0.01
0.02
0.03
0.04
−4
−2
0
2
4
x
Can hardly see RMSE shown in blue

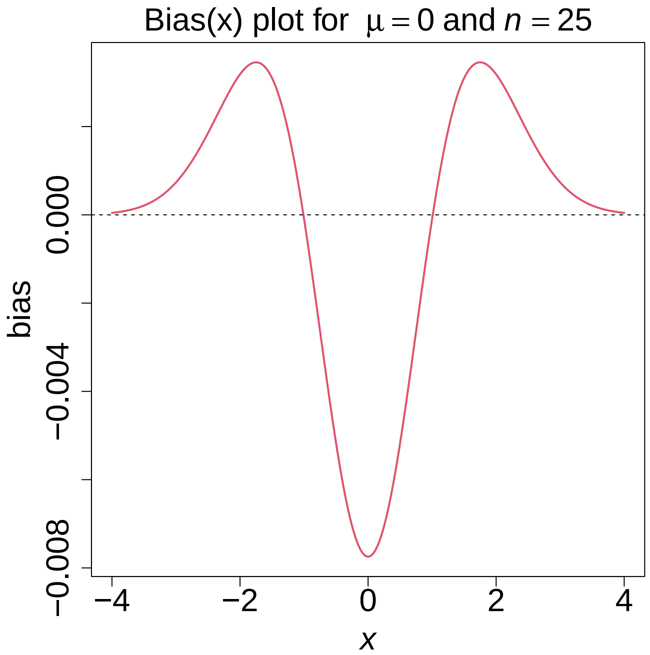
How to Evaluate $\hat{\theta}$? Theoretically. (Watch for $\bar{x}$ vs $\bar{X}$?)

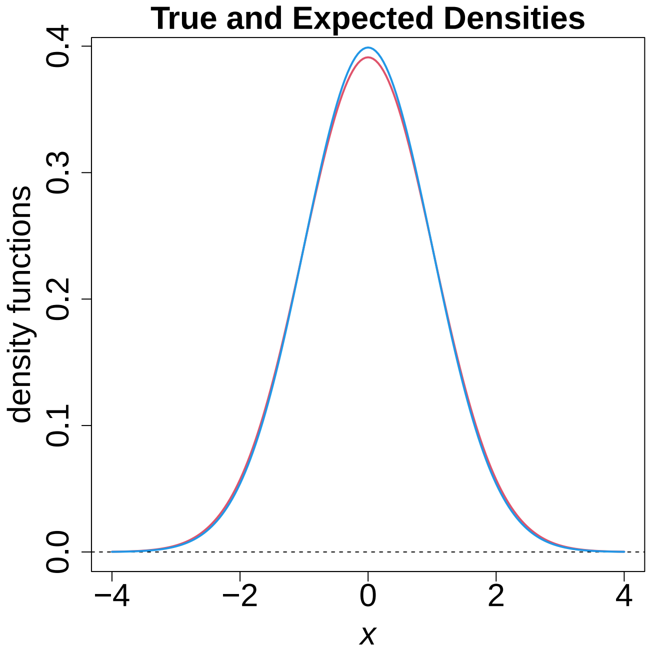
$$
\begin{aligned}
\hat{\theta} &= \phi(x|\bar{X},1) \\
Bias(x) &= E\left[\hat{\theta}-\theta\right] \qquad \text{what is random?} \\
&= \int_{-\infty}^{\infty}\left[\phi(x|\bar{x},1)-\phi(x,|\mu,1)\right]\times\phi\left(\bar{x}\middle|\mu,\frac{1}{n}\right)d\bar{x} \\
&= \phi\left(0\middle|x-\mu,1+\frac{1}{n}\right)-\phi(x|\mu,1) \\
&= \phi\left(x\middle|\mu,1+\frac{1}{n}\right)-\phi(x|\mu,1)\,.
\end{aligned}
$$

Looks a little strange, but we used the symmetric of $(x-\mu)^2$ in the normal exponent to obtain

$$
\phi(x|\bar{x},1)=\phi(\bar{x}|x,1) \quad \text{and} \quad \phi(0|x-\mu,1)=\phi(x|\mu,1)\,;
$$

so can use the cool identity cleverly twice.

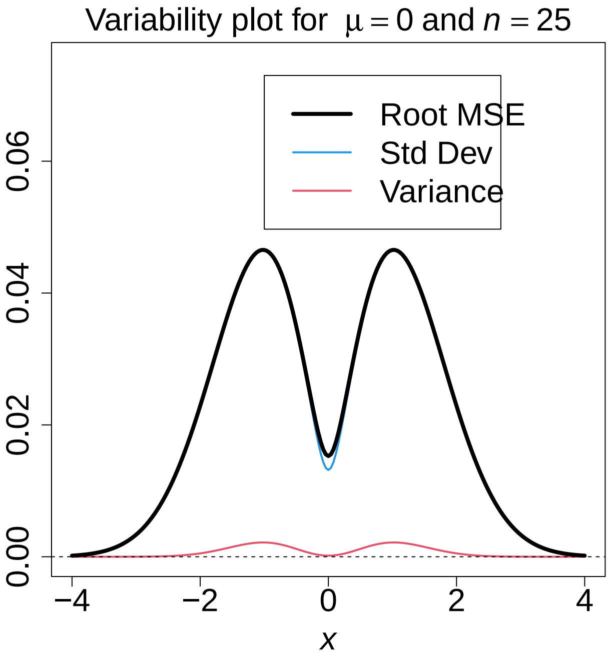

Bias(x) plot for μ = 0 and n = 25
bias
0.000
−0.004
−0.008
−4
−2
0
2
4
x


True and Expected Densities
density functions
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x

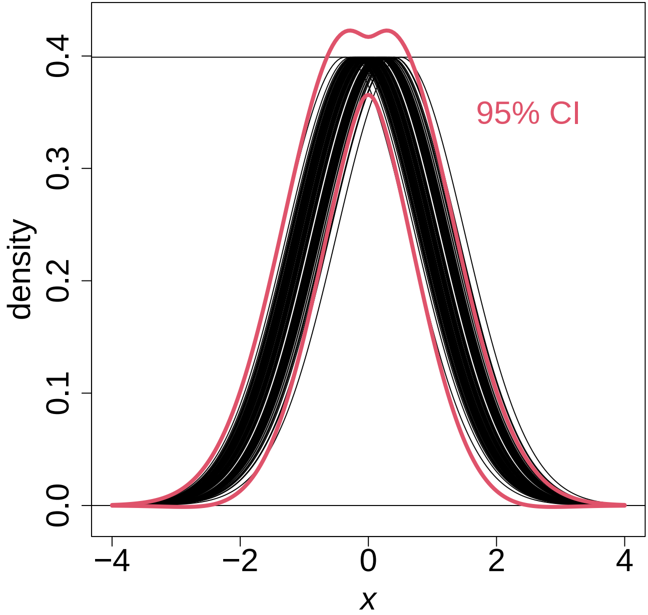
Variance of $\hat{\theta}$

$$
\begin{aligned}
\hat{\theta} - \theta &= \phi(x|\bar{X}, 1) - \phi(x, |\mu, 1) \\
Var(\hat{\theta}) &= E\left[(\hat{\theta} - \theta)^2\right] \\
&= E\left[\phi(x|\bar{X}, 1)^2 - 2\phi(x|\bar{X}, 1)\phi(x, |\mu, 1) + \phi(x, |\mu, 1)^2\right] \\
&= \int_{-\infty}^{\infty} \left[\frac{1}{2\sqrt{\pi}}\phi(\bar{x}|x, 1/2)\right] \times \phi\left(\bar{x}\middle|\mu, \frac{1}{n}\right) d\bar{x} \\
&\quad - 2\int_{-\infty}^{\infty} \left[\phi(\bar{x}|x, 1)\phi(x, |\mu, 1)\right] \times \phi\left(\bar{x}\middle|\mu, \frac{1}{n}\right) d\bar{x} \\
&\quad + \phi(x, |\mu, 1)^2 \qquad \text{nothing random here, so integral 1} \\
&= \frac{1}{2\sqrt{\pi}}\phi\left(x|\mu, \frac{1}{2} + \frac{1}{n}\right) - 2\phi(x|\mu, 1)\phi\left(x|\mu, 1 + \frac{1}{n}\right) \\
&\quad + \frac{1}{2\sqrt{\pi}}\phi(x, |\mu, 1/2)\,,
\end{aligned}
$$

using the identity $\quad \phi(x|\bar{x}, 1)^2 = \dfrac{1}{2\sqrt{\pi}}\phi(\bar{x}|x, 1/2)\,.$


Variability plot for μ = 0 and n = 25
Root MSE
Std Dev
Variance
0.00
0.02
0.04
0.06
−4
−2
0
2
4
x

100 Simulations with $\mu = 0$ and $n = 25$



Final Thoughts

- While $\bar{X}$ is unbiased for μ, $\hat{\theta}_x$ is not unbiased for $\phi(x)$!!!
- We will use R extensively in this course. You are welcome to play with MatLab, etc.
- We will also find Mathematica very helpful in many situations. Again, MAPLE?
- JMP is great for quick analyses and graphics. All of these are available for free at `kb.rice.edu`
- Do not be shy about asking questions in real time. If you have a question or didn't catch something, then 99% confidence others are in the same boat. ASK! It will slow me down, too.
- This course is great at putting lots of other course material in a useful perspective, IMO.
- This lecture illustrates the book's subtitle: Theory, Practice, and Visualiztion. We used the book's material heavily already!