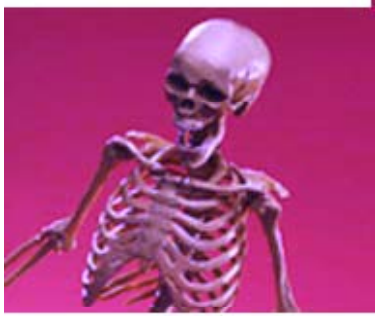
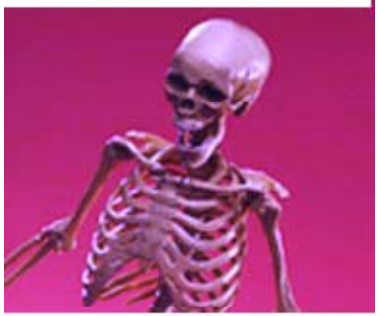


September 16, 2009

Do we really need statistics in science?

"FWF Graduate Seminar"

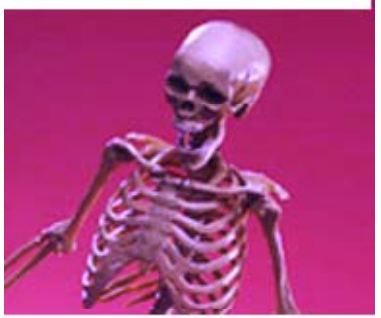
Timothy M. Young, Ph.D.
Associate Professor
Department of Forestry, Wildlife & Fisheries
Forest Products Center



"Don't worry, it will be OK……."

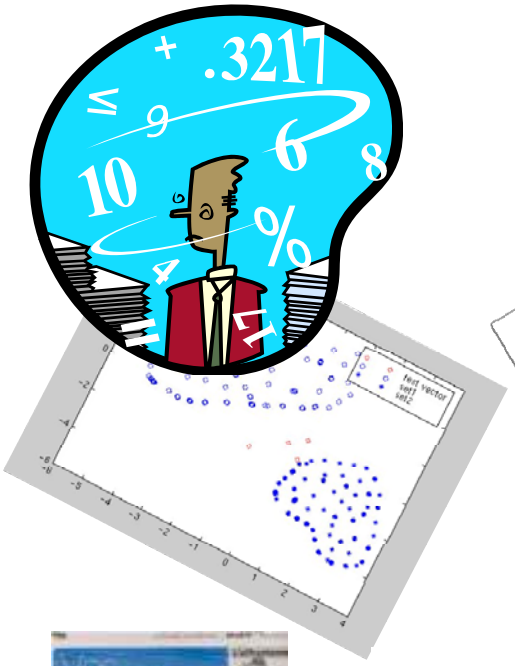
Overview

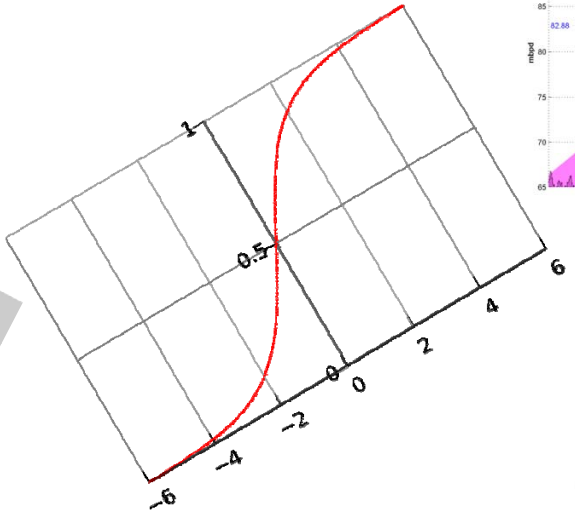
- Definition of Statistics
  - Variance (σ^2)
  - Random variable (*sample space*)
  - Probability
- Definition of Science
- 1st Law of Statistics
- Key Assumptions
- Research Program

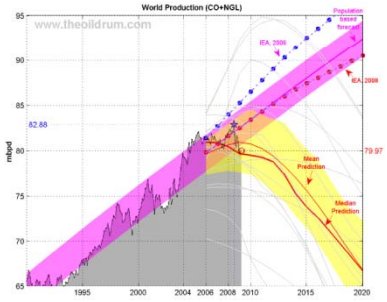


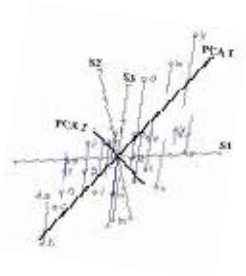
Statistics

“The Measurement of Uncertainty”





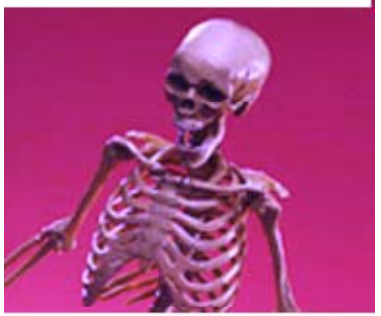




2009 Tennessee Volunteers
Tennessee Red-Zone Results (as of Sep 12, 2009)
All games

Tennessee Inside Opponent Red-Zone

Date	Opponent		Score	Times In RZ	Times Scored	Total Pts	TDs	Rush TDs	Pass TDs	FGs Made	---- Failed to score inside RZ ---- FGA	Downs	Int	Fumb	Half	Game
Sept 5	Western Kentucky.....	W	63-7	9	8	56	8	3	5	0	0	0	1	0	0	0
Sep 12	UCLA.....	L	15-19	4	3	13	1	1	0	2	0	1	0	0	0	0
	Totals			13	11	69	9	4	5	2	0	1	1	0	0	0
	11 of 13 (84.6%)															

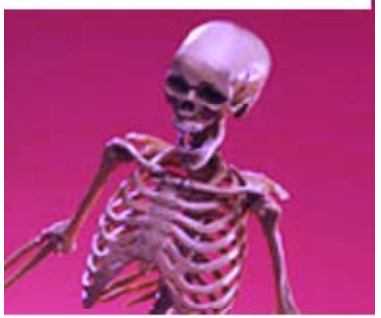


Definition of Statistics?

- Many, many definitions…….most people in this room would have different definitions
- Common theme of most definitions:

  "….study of variance (σ^2)……."

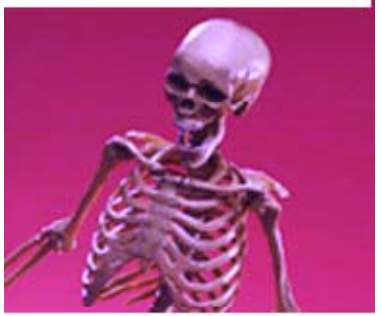
  "……..quantifying variance……"



Francis Galton
1822-1911

Variance (σ^2)

Components of a System



- If X and Y are two random variables,

(X,Y dependent):

$$Var(X + Y) = VarX + VarY + 2Cov(X,Y)$$

or,

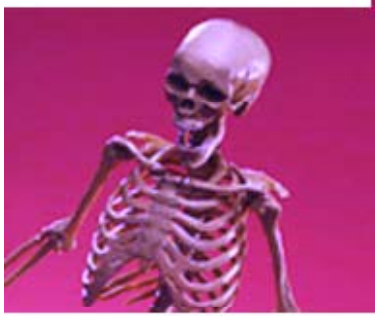
$$Var(aX + bY) = a^2VarX + b^2VarY + 2abCov(X,Y)$$

(X,Y independent):

$$Var(X + Y) = VarX + VarY$$

Variance (σ^2)

Components of a System



- Generalization

(dependent random variables):

$$\mathrm{var}[\sum_i^n X_i] = \sum_i^n \mathrm{var}[X_i] + 2\sum\sum_{i<j} \mathrm{cov}[X_i, X_j]$$

(independent random variables):

$$\mathrm{var}[\sum_i^n X_i] = \sum_i^n \mathrm{var}[X_i]$$

Variance (σ^2)

Components of a System

- Reduce (or increase) variance

  (dependent random variables):

  ⬇$Var(X + Y) = VarX +$ ⬇$VarY +$ ⬇⬆$2Cov(X,Y)$

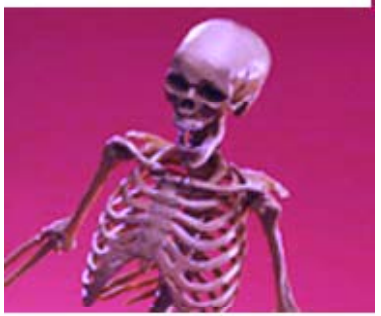
  (X,Y independent):

  ⬇$Var(X + Y) = VarX +$ ⬇$VarY$

  or

  ⬆$Var(X + Y) = VarX +$ ⬆$VarY$

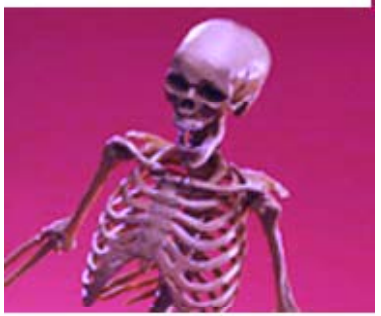
- What is the difficulty: "quantifying σ^2"



Random Variable

Random variable (*X)* is allowed to vary within a sample space for the set of real numbers

e.g., weight, height, moisture content, number of spots, distance traveled, survival rate, etc.

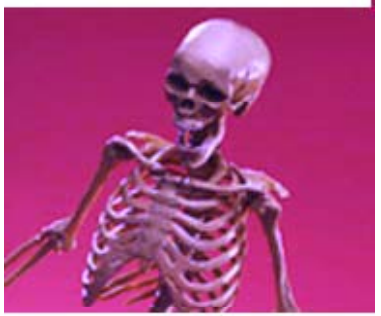


Random Variable

Two dice: 

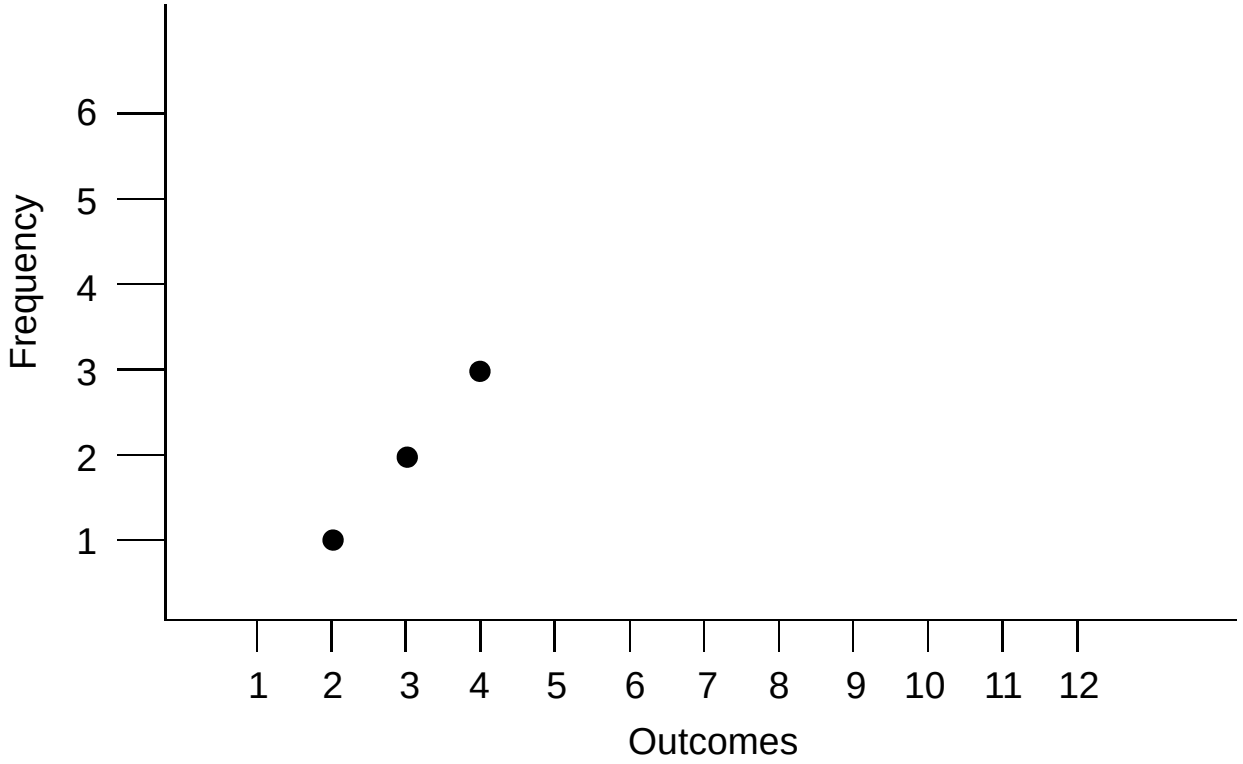
Outcomes: (2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)

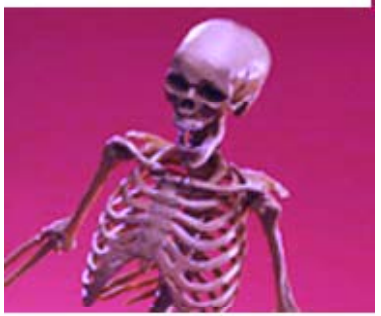
Probability of an outcome?



Random Variable

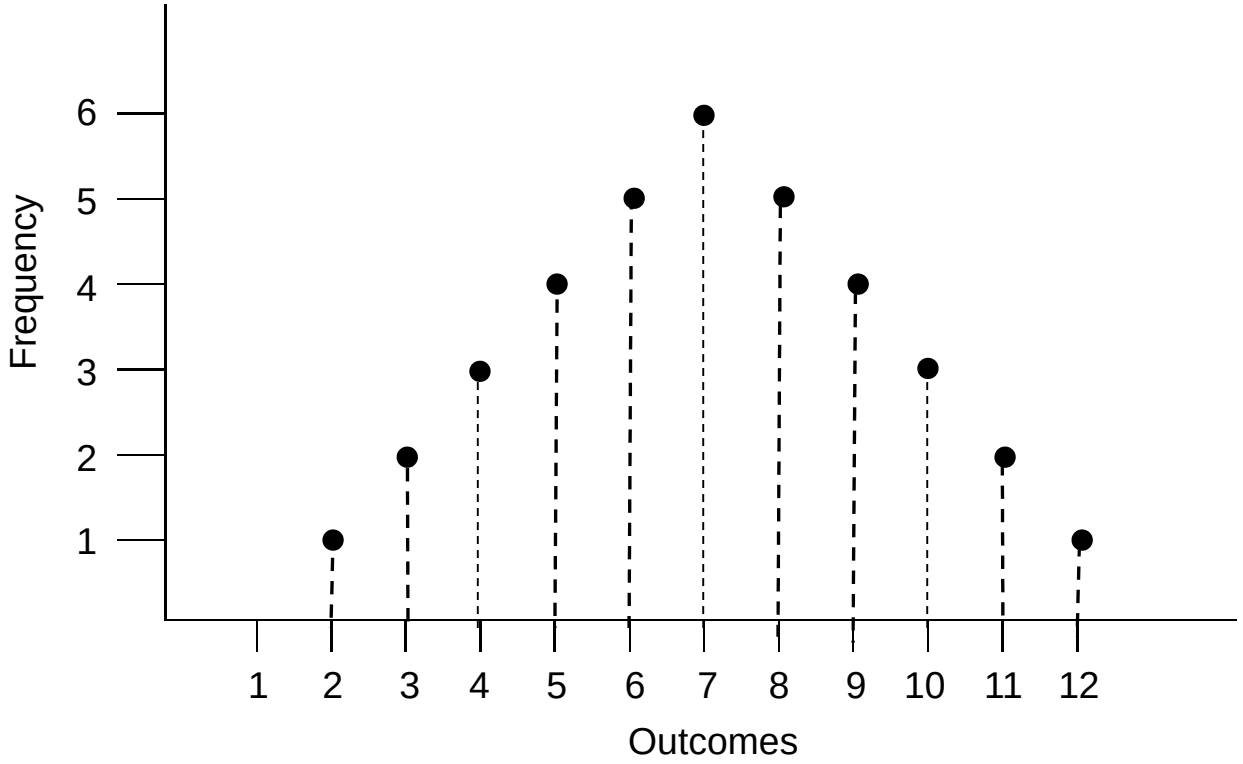
Two dice:

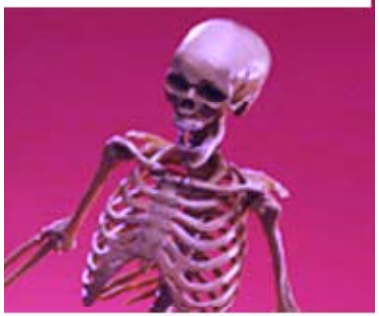




Random Variable

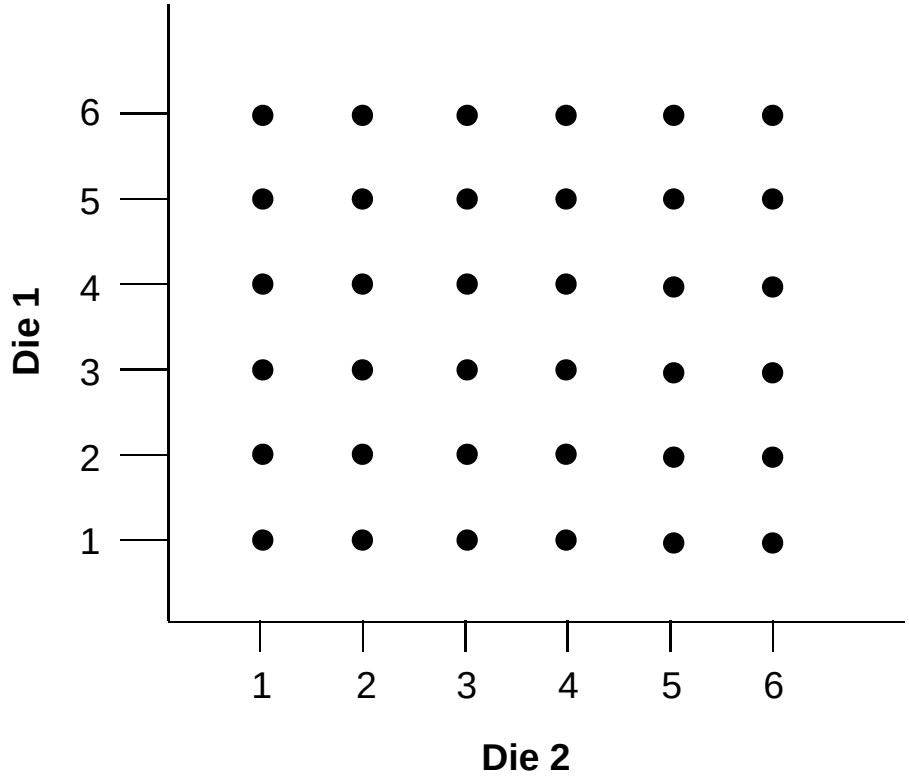
Two dice: 

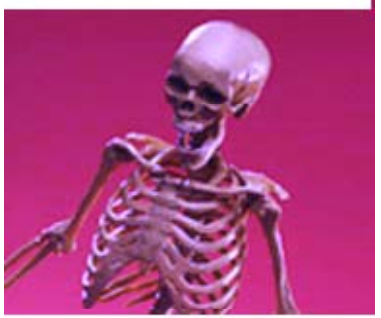




Sample Space

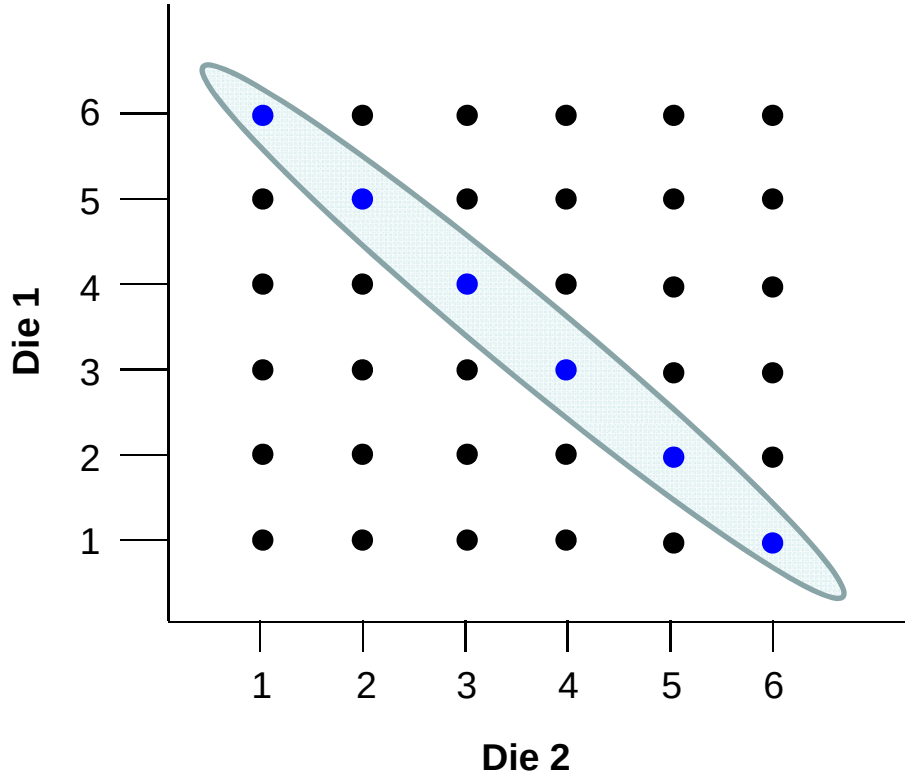
Two dice:

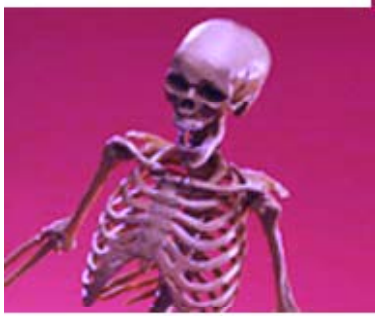




Sample Space

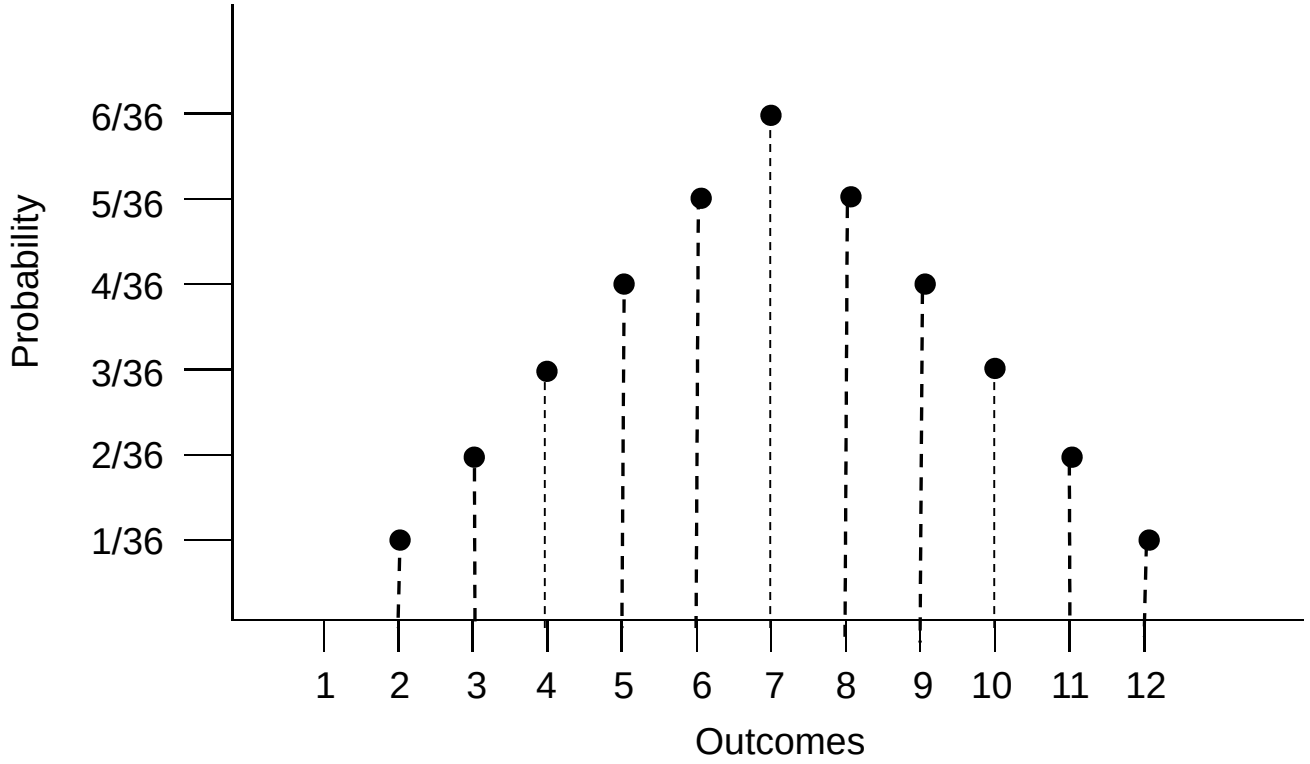
Two dice:  *Lucky 7?*

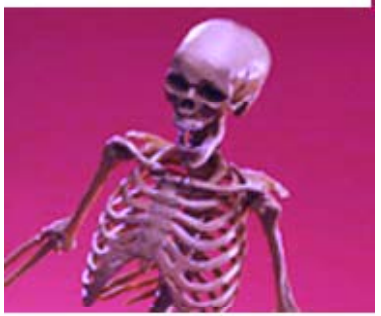




Probability Density Function (pdf)

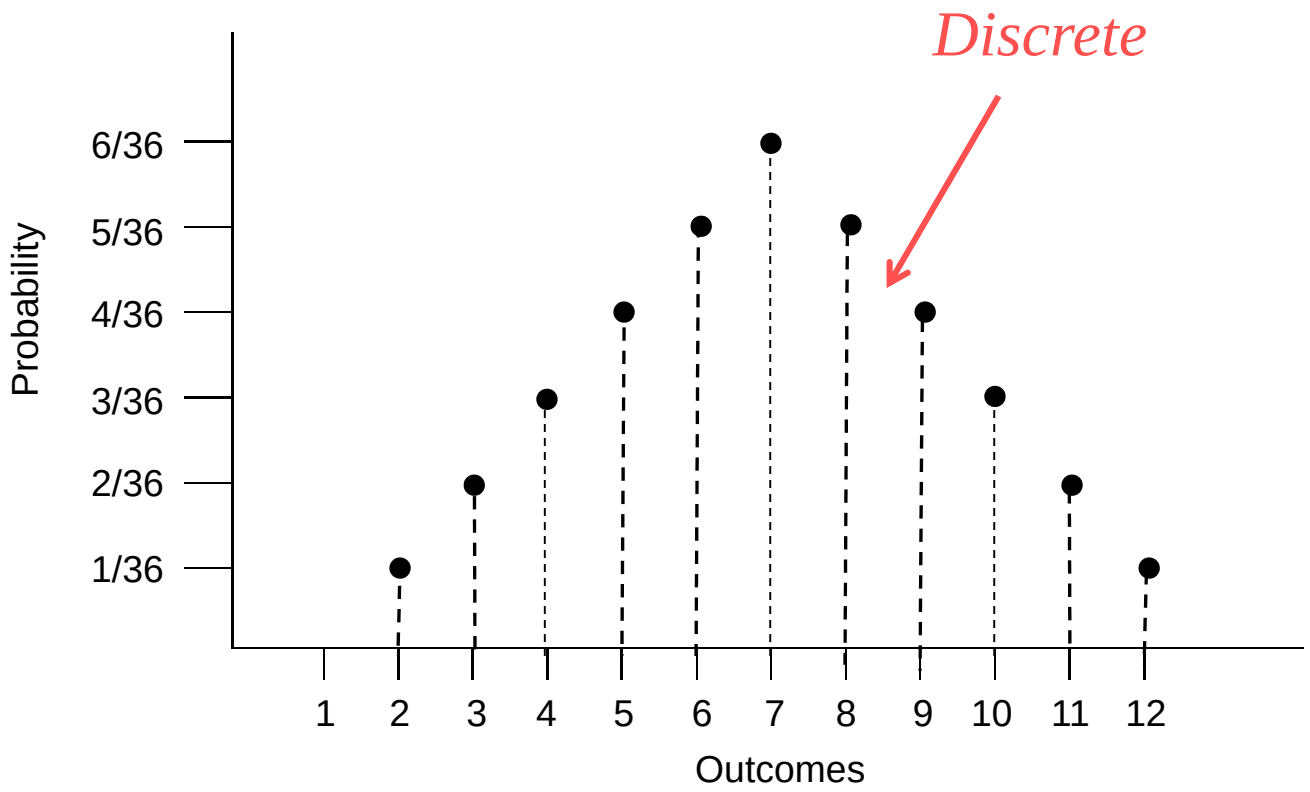
Two dice:

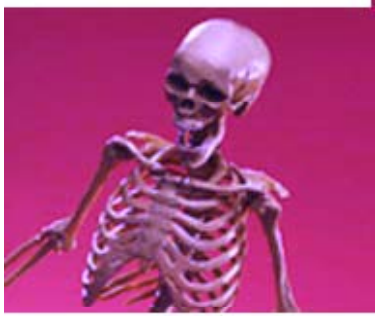




Probability Density Function (pdf)

Two dice:





Probability Density Function

(Discrete pdfs)

Bernoulli $f(x) = p^x (1-p)^{1-x}$

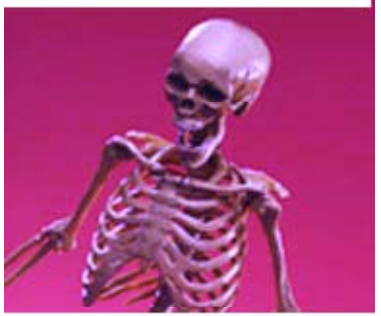
Binomial $f(x) = \left(\frac{n}{x}\right) p^x (1-p)^{n-x}$

Poisson $f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$

Geometric (or Pascal) $f(x) = p(1-p)^x$

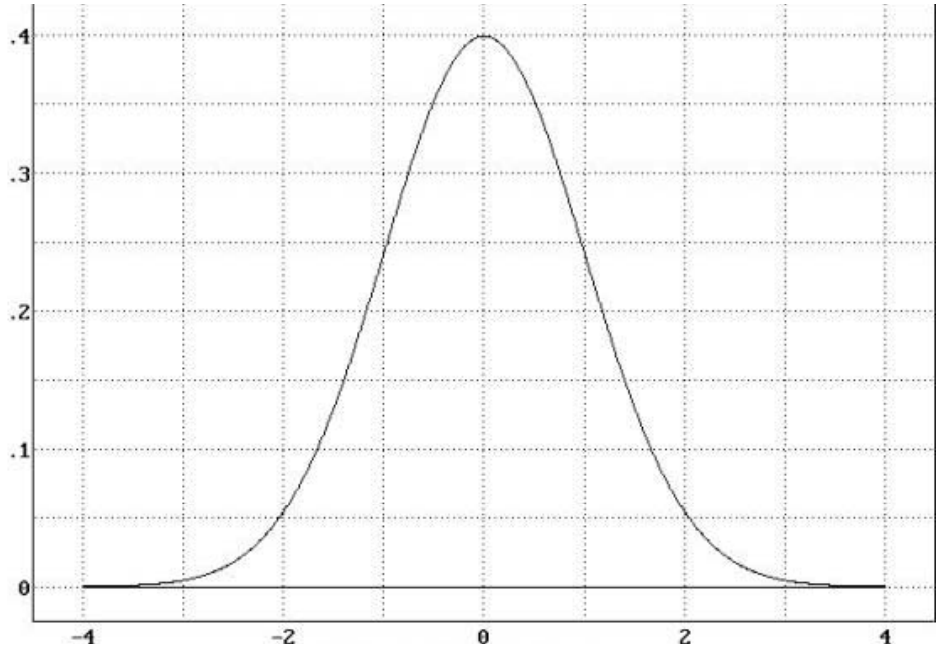
Hypergeometric $f(x) = \frac{\left(\frac{K}{x}\right)\left(\frac{M-K}{n-x}\right)}{\left(\frac{M}{n}\right)}$

etc., etc. etc……..



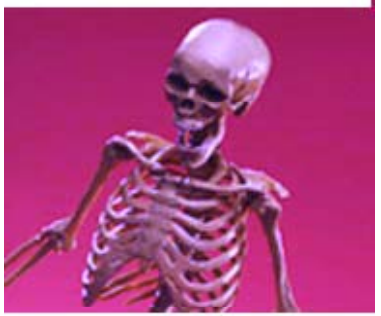
Probability Density Function (pdf)

Continuous



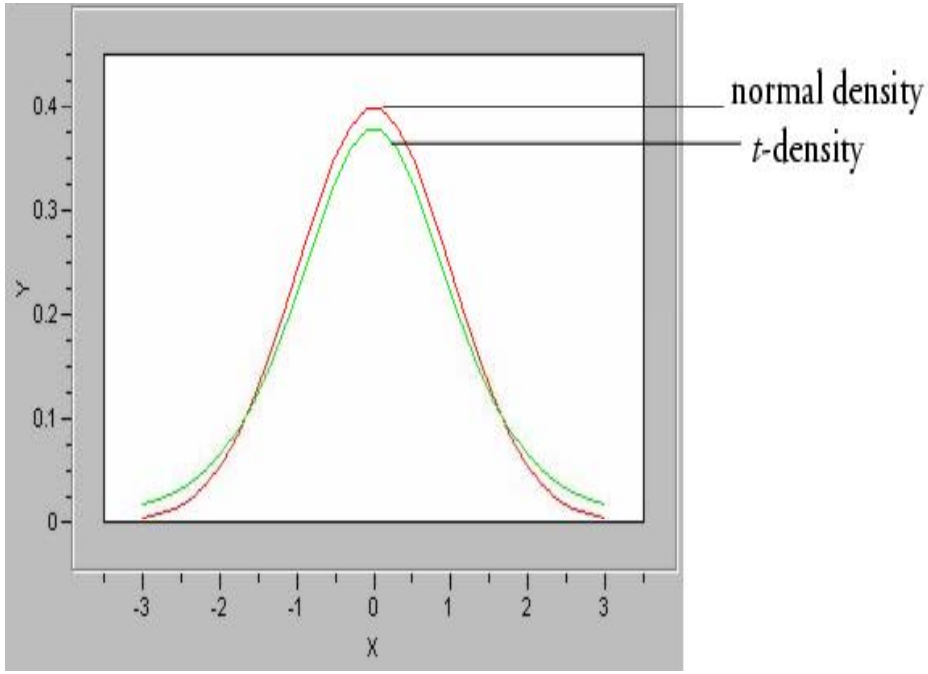
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2)[(x-\mu)/\sigma]^2}$$

"Normal "or "Gaussian" (with $\mu = 0$; $\sigma = 1$)

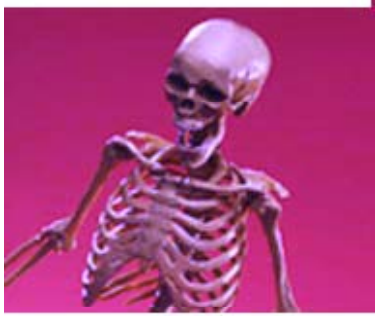


Probability Density Function (pdf)

Continuous (e.g., t distribution)



The *t* statistic was invented by William Sealy Gosset (1876 – 1937) for cheaply monitoring the quality of beer brews. “Student” was his pen name.



Probability Density Function

(Continuous pdfs)

Normal

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2)[(x-\mu)/\sigma]^2}$$

Standard Normal

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)[(x]^2} \quad (\mu = 0;\ \sigma = 1)$$

Student's t

$$f(x) = \frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2})} \frac{1}{p\pi^{1/2}} \frac{1}{(1+x^2/p)^{(p+1)/2}}$$

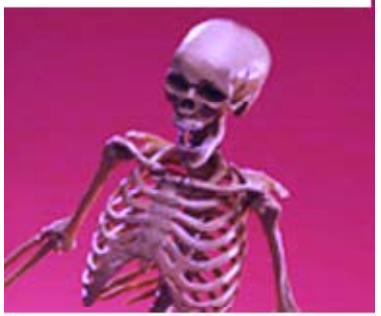
F distribution

$$f(x) = \frac{\Gamma(\frac{p+q}{2})}{\Gamma(\frac{p}{2})\Gamma(\frac{q}{2})} \left(\frac{p}{q}\right)^{p/2} \frac{x^{(p/2)-1}}{[1+(p/q)x]^{(p+q)/2}}$$

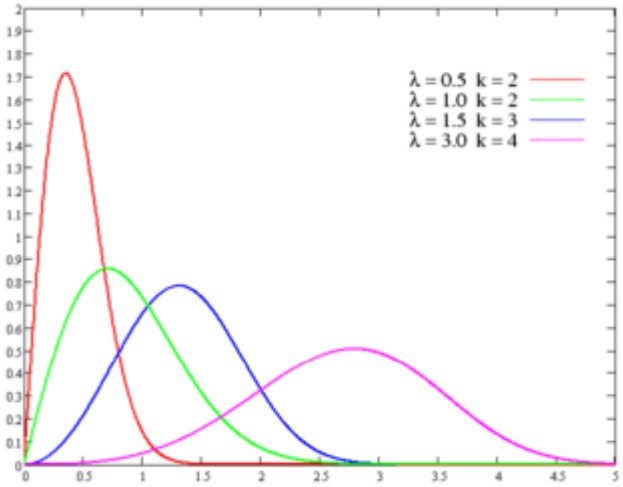
Weibull

$$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$$

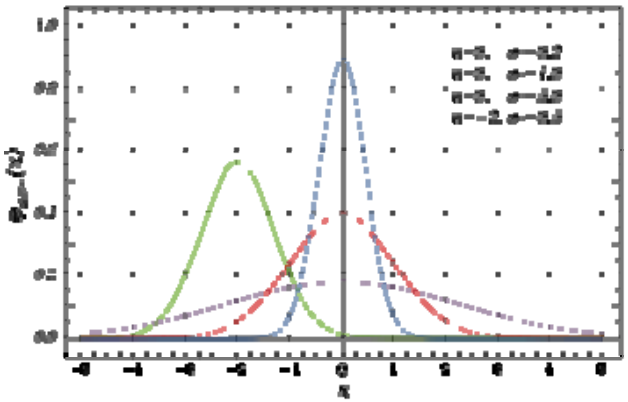
etc., etc. etc……..



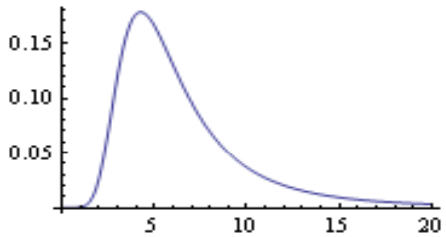
Why are the probability density functions important?



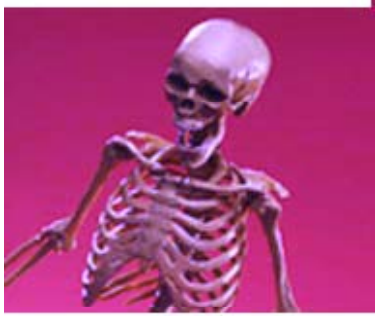
Weibull



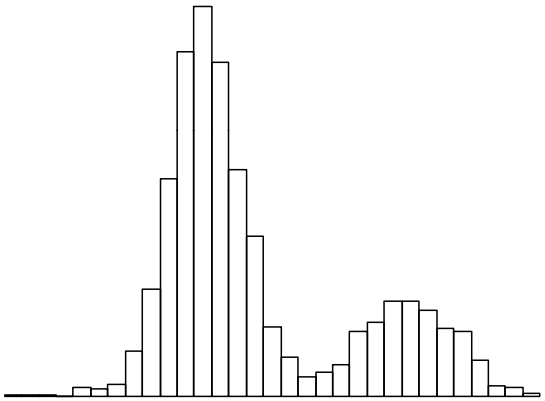
Normal

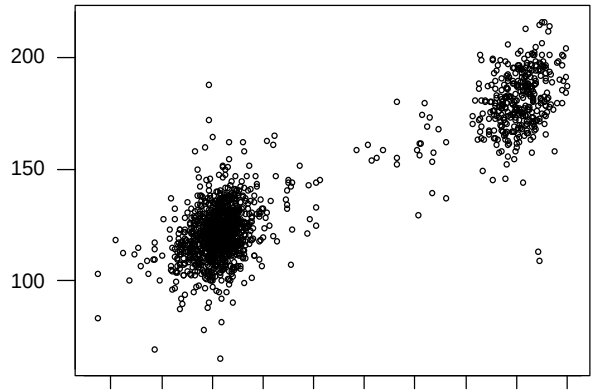


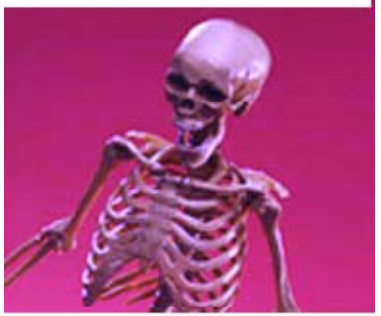
Non-central t



How are your data collected from some sample space?







Statistical Methods

<u>Parametric Methods</u>

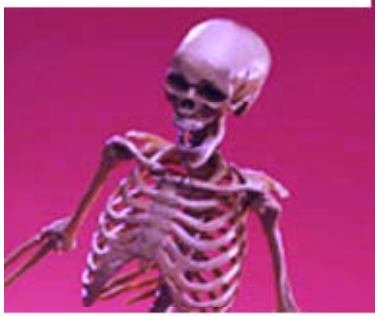
Analysis of variance (ANOVA)
Chi-square test
Correlation
Factor Analysis
Mann-Whitney U
Mean Square Weighted Deviation MSWD
Pearson product-moment correlation coefficient
Regression analysis
Logistic regression
Spearman's rank correlation coefficient
Student's t-test
Time Series Analysis
etc., etc., etc…….

<u>Non-Parametric Methods</u>

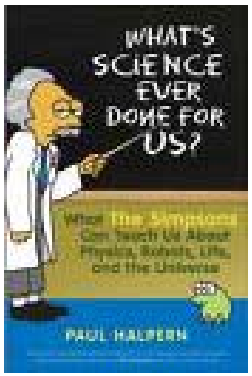
Anderson-Darling test
Cochran's Q
Cohen's kappa
Efron-Petrosian test
Friedman two-way analysis
Kendall's tau
Kendall's W
Kolmogorov-Smirnov test
Kruskal-Wallis one-way analysis
Kuiper's test
Wilcoxon rank sum test
Pitman's permutation test
Rank products
Siegel-Tukey test
Wilcoxon signed-rank test.
etc., etc., etc…….

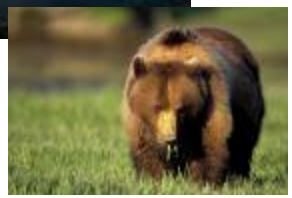
Statistical Inference

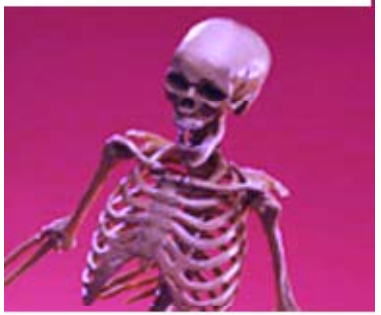
- Statistical inference (or statistical induction) is the use of statistics and random sampling to make inferences concerning some unknown aspect of the population.



Definition of Science?







Definition of Science?

The word science comes from Latin "scientia," meaning knowledge.

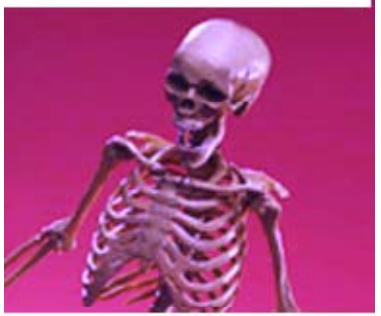
"Science is an *intellectual activity carried on by humans* that is designed to discover information about the natural world in which humans live and to discover the ways in which this *information* can be *organized* into *meaningful patterns*"

"It is done through *observation of natural phenomena*, and/or through *experimentation* that tries to simulate natural processes under controlled conditions"

"A primary aim of science is to collect *facts (data)*"

"An ultimate purpose of science is to *discern the order* that exists *between and amongst* the various *facts*"

"……..systematic knowledge-base or prescriptive practice that is capable of resulting in a *prediction* or predictable type of outcome……"

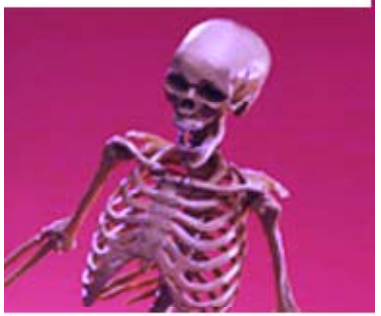


Definition of Science?

Scientific Reasoning or Inference

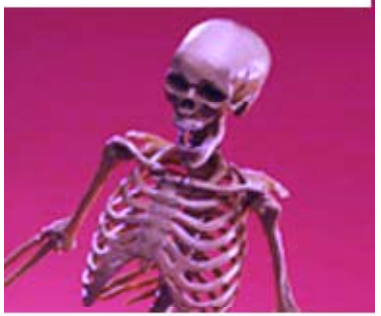
<u>Deductive Reasoning</u>:

Knowledge → Theory → Hypothesis → Data



Definition of Science?

Scientific Reasoning or Inference



Deductive Reasoning:

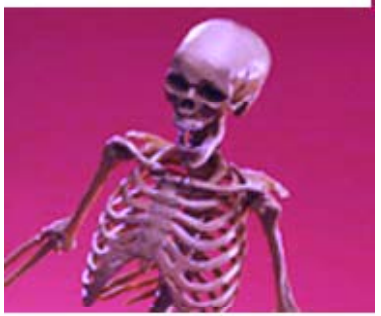
Knowledge → Theory → Hypothesis → Data

Inductive Reasoning:

Data → Hypothesis → Theory → Knowledge

Do we really need Statistics in Science?

- Observations (data)
- Meaningful patterns
- Experimentation
- Discern order between and amongst facts
- Hypothesis testing (or generation)
- Prediction
- etc., etc., etc…….

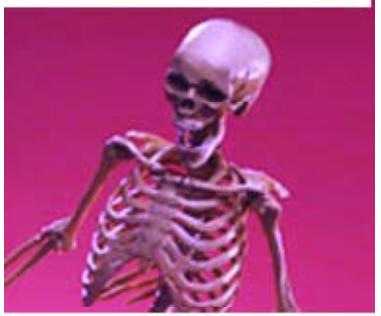


Do we really need Statistics in Science?

How will you quantify variance (σ^2) of observational data without the use of statistical methods?

"An approximate answer to the right question is worth a good deal more than the exact answer to an approximate problem"

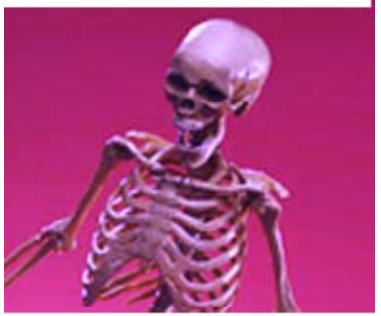
John W. Tukey (1915-2000)



First Law of Applied Statistics

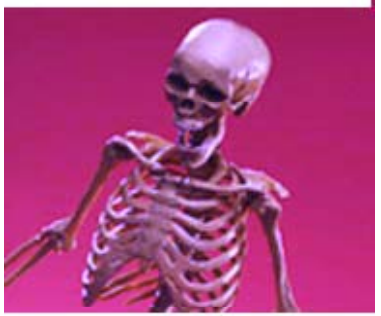
(Gleser 1996)

"….two individuals using the same statistical method on the same data should arrive at the same conclusion."



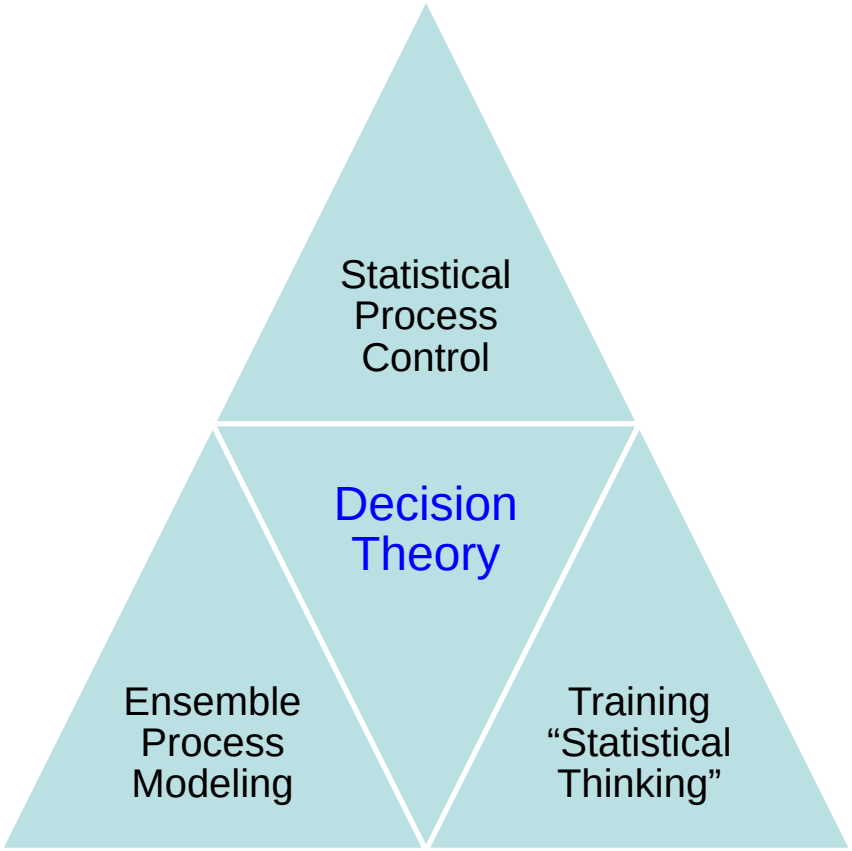
Key Assumptions

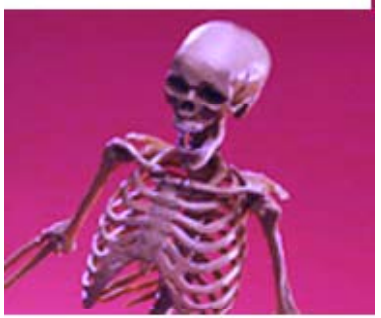
- What is the question (problem definition)?
- Sample space (bias?)
- Data quality
- Parametric (pdf) assumption?
- Most appropriate method to provide the approximate answer to a well-defined question



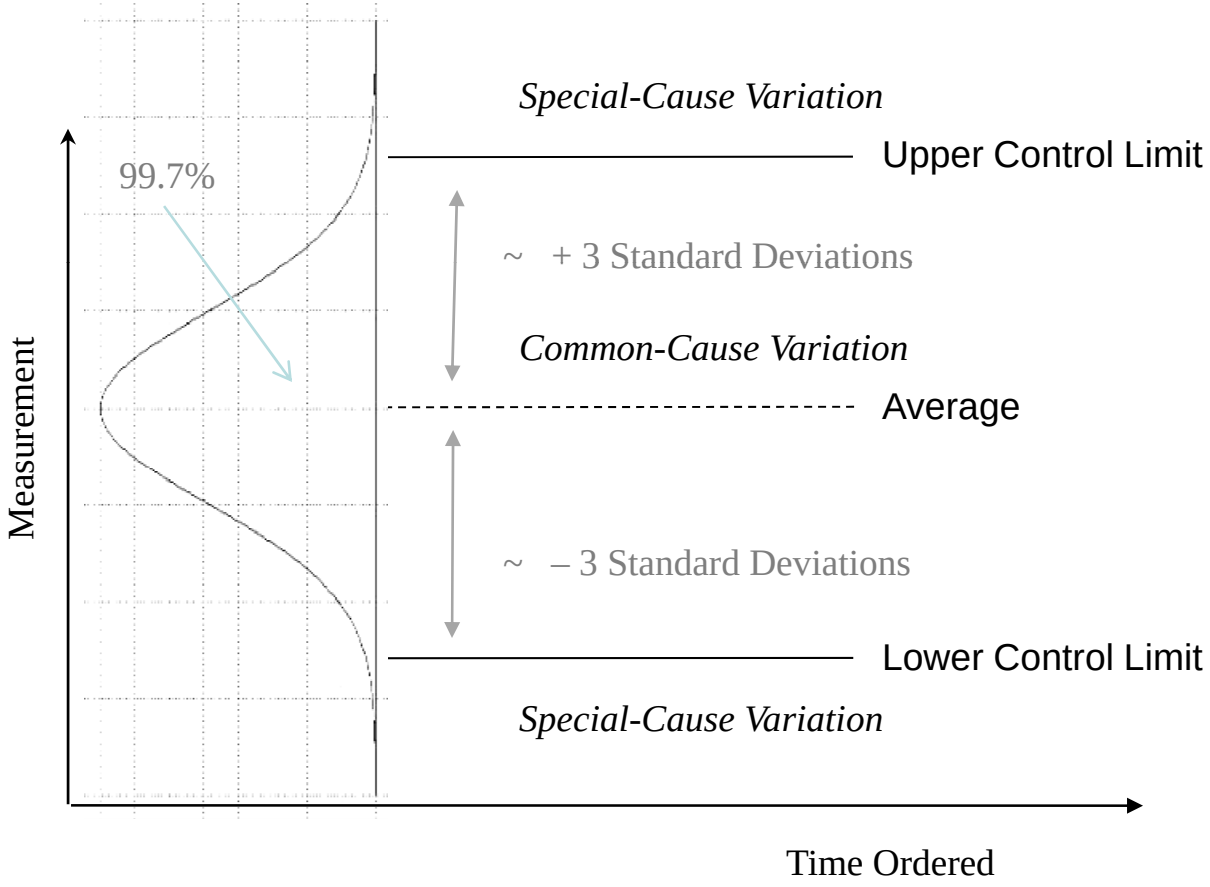
Research Program

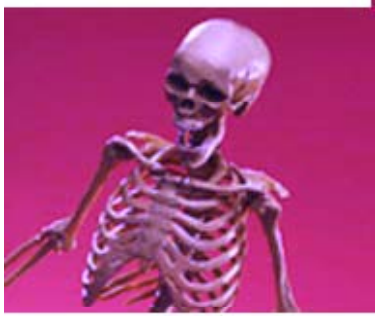
(Bio-based Products Industries)





Statistical Process Control



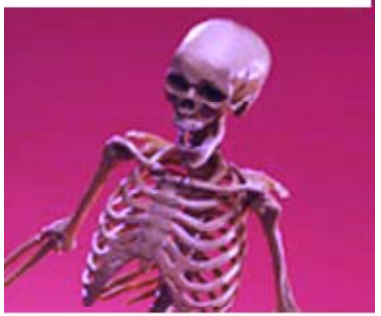


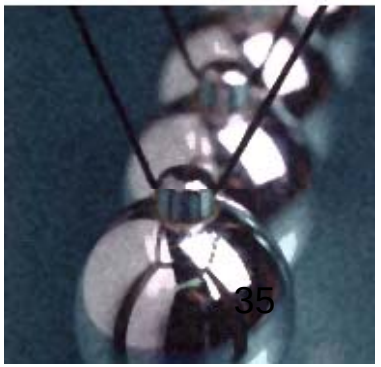
Statistical Process Control

Hardwood/Softwood Sawmills

Hypothesis, a priori,

"…the use of real-time statistical process control to monitor and reduce lumber thickness variation does not improve lumber recovery, lumber quality or financial performance."





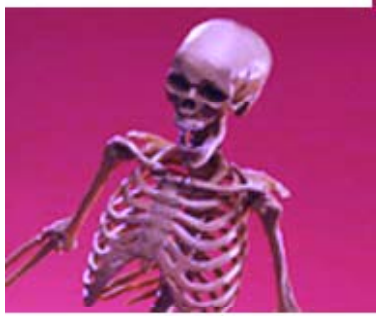
Statistical Process Control

Hardwood/Softwood Sawmills

Summary - Sawmill A

Quercus rubra

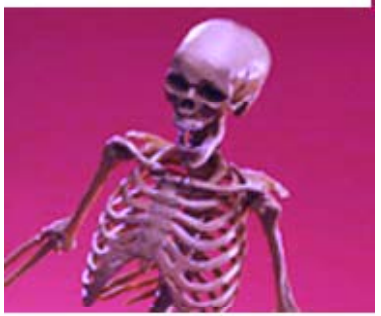
Month-Year	n	Average $\bar{x}$	Changes in $\bar{x}$ **	Median (M)	Changes in M ***	Standard Deviation $\hat{\sigma}_x$	Changes in $\hat{\sigma}_x$ ****
August – 1999	171	1.134”	a	1.135”	a	0.044	a
September – 1999	230	1.124”		1.117”		0.044	ab
October – 1999	61	1.126”	a	1.123”	ab	0.027	c
November - 1999	188	1.108”	ab	1.109”	c	0.040	ab d
December - 1999	641	1.121”	a	1.121”	ab	0.045	ab d
January - 2000	--*	--	--	--	--	--	
February – 2000	--	--	--	--	--	--	
March - 2000	--	--	--	--	--	--	
April – 2000	328	1.103”	ab	1.115”	abc	0.035	cd
May – 2000	414	1.114”	a	1.112”	abc	0.033	cd
June – 2000	431	1.180”	a	1.106”	c	0.035	cd
July – 2000	44	1.096”	b	1.097”		0.024	c



Statistical Process Control

Hardwood/Softwood Sawmills

Company	Investment	Return	ROI
A	$15,000	$180,000	12:1
B	$27,000	$752,000	28:1
C (softwood)	$13,000	$210,000	16:1
D	$21,000	$147,000	7:1
Total:	**$76,000**	**$1,289,000**	**17:1**

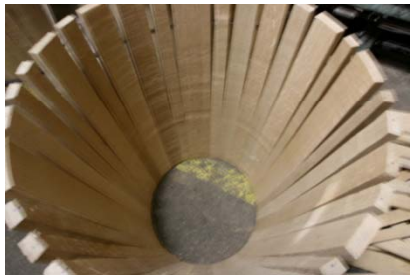


Statistical Process Control

Staves for Bourbon Barrels

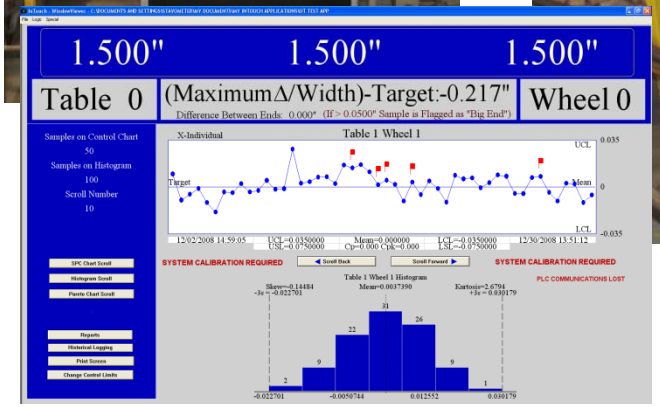
BROWN-FORMAN

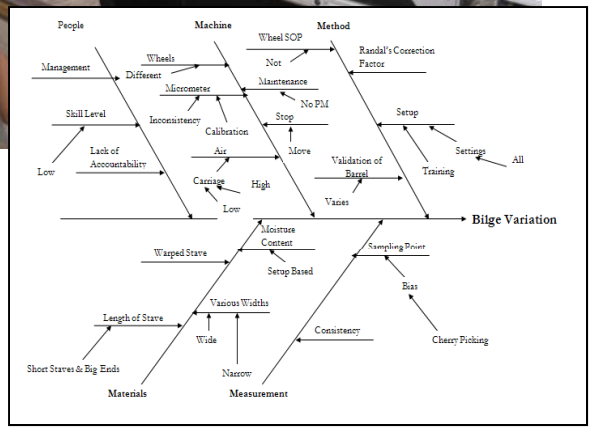
Hypothesis: A reduction in stave width variability and bilge variability in any of the 15 jointer wheels will reduce both within and between barrel circumference variability

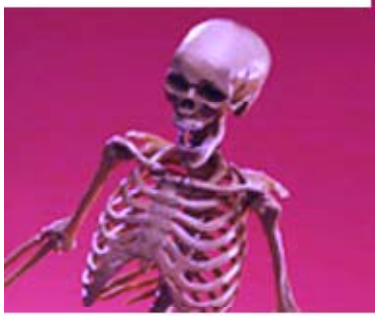


Statistical Process Control

Staves for Bourbon Barrels

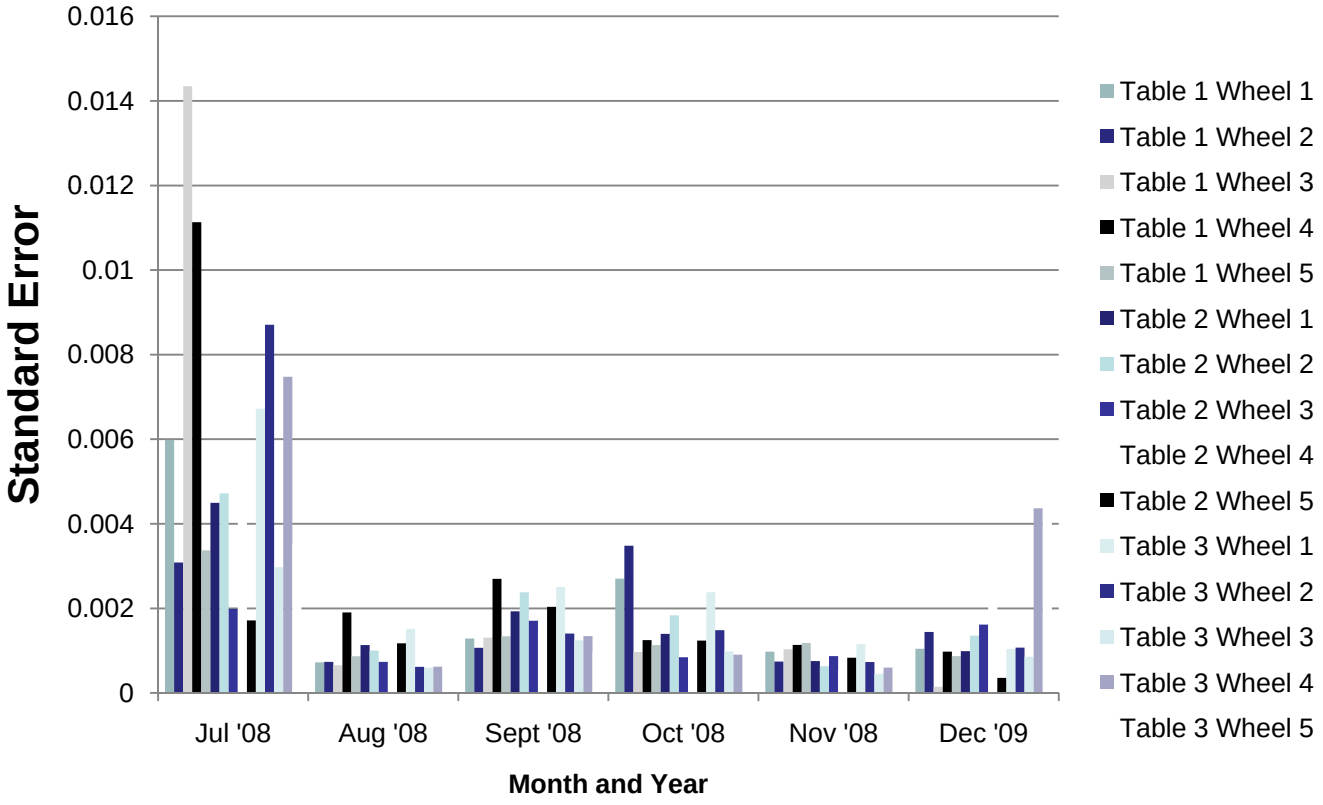


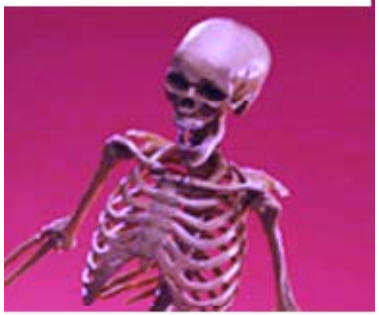




Statistical Process Control

Staves for Bourbon Barrels

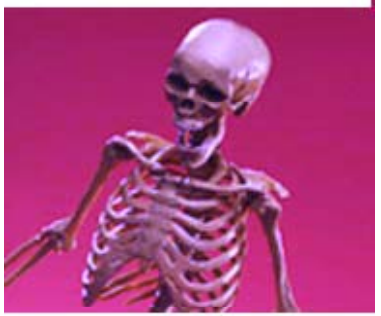




Statistical Process Control

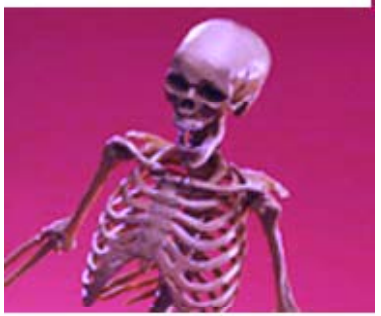
Staves for Bourbon Barrels

- Barrel circumference variation was reduced
- Allowed for increase in barrel target circumference size
- Yield per barrell (Official Proof Gallons) at Jack Daniels improved by 0.3 OPG after SPC
- Additional 938 barrels of Jack Daniels and approximately $300,000 of cost savings over the six-month study period (estimated by Brown Forman)



“Ensemble” Process Modeling

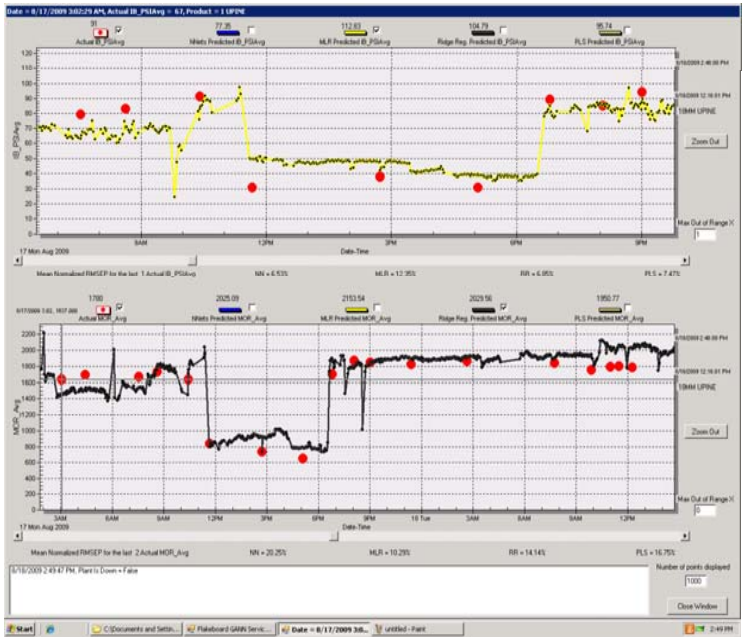
Question: Can improved methodologies for real-time process modeling improve the scientific understanding of undiscovered correlations in bio-based products manufacturing (facilitate improved causation investigation) ?

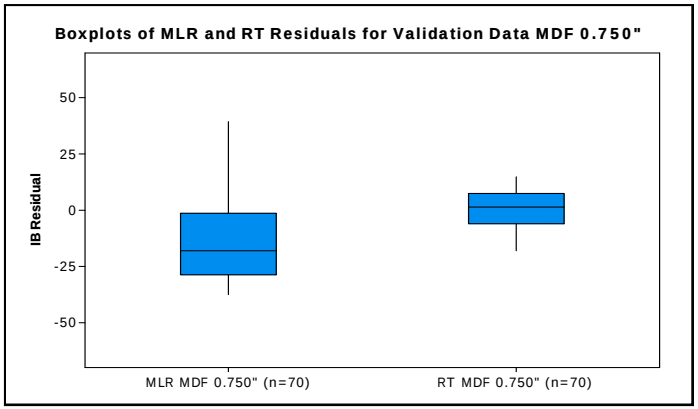


“Ensemble” Process Modeling

<u>Idea</u>: Reduce generalized error of prediction by combining predictions from several models and various types of algorithms into an “ensemble”

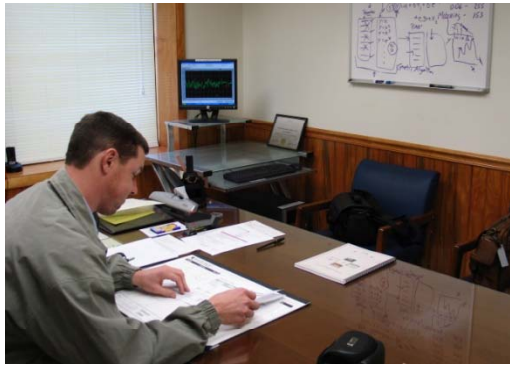
- MLR
- Regression Trees
- Partial Least Squares
- Ridge Regression
- Neural Networks

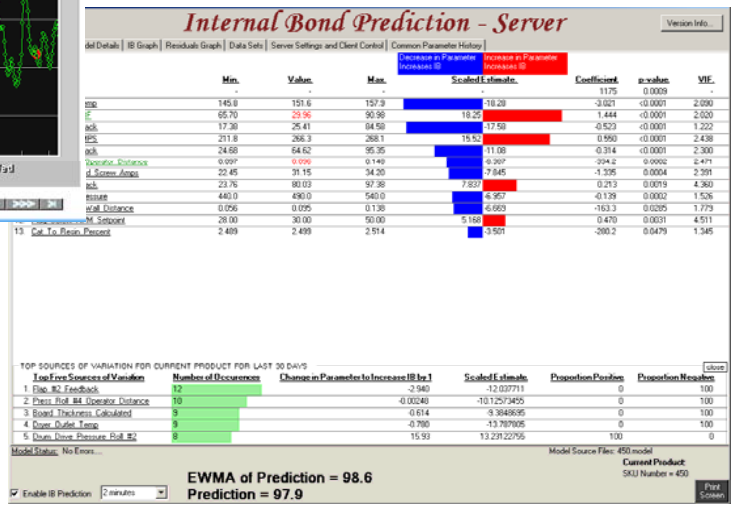


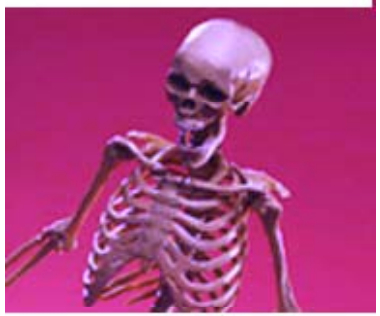


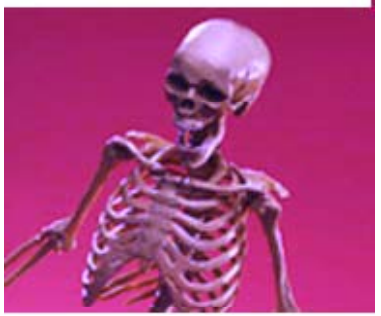
“Ensemble” Process Modeling

Several large projects with USDA SBIR competitive grants and private industry (T: $1.6M)



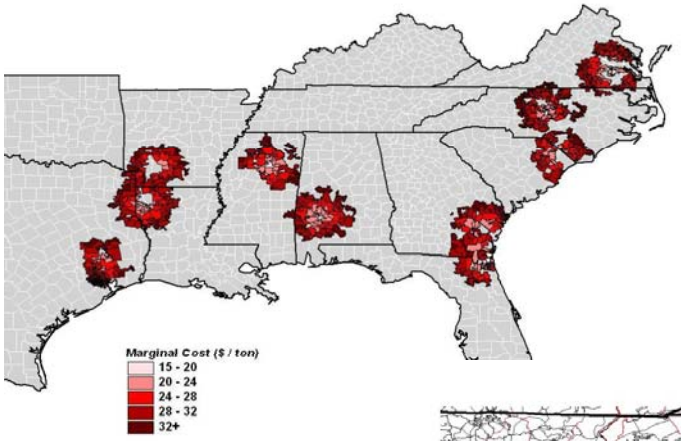


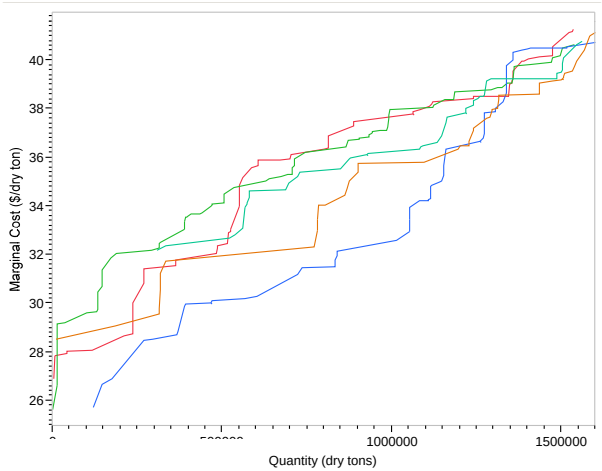


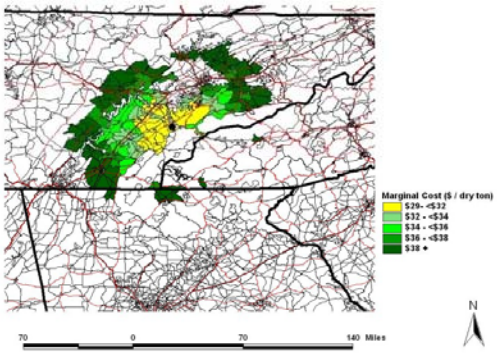


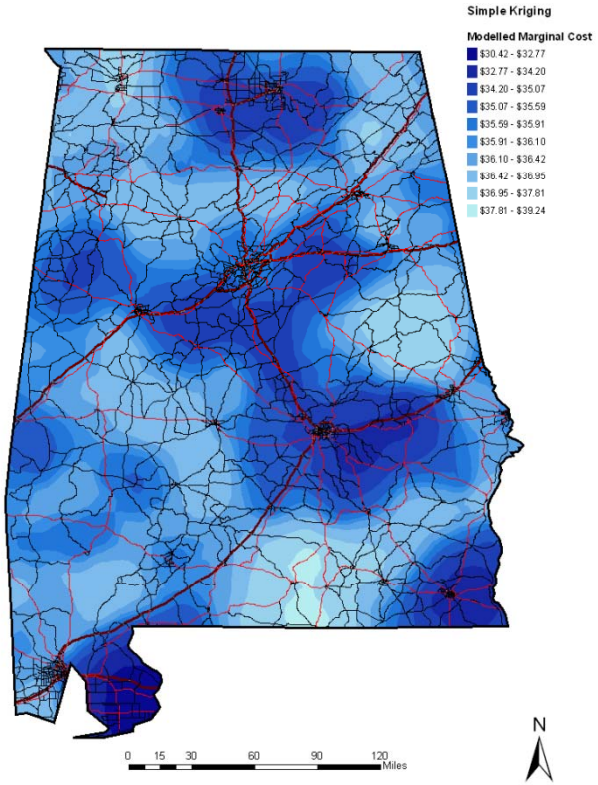
“Ensemble” Process Modeling

“BioSAT” Model: Modeling system for determining optimal locations for biomass using facilities in the eastern U.S. (www.biosat.net)



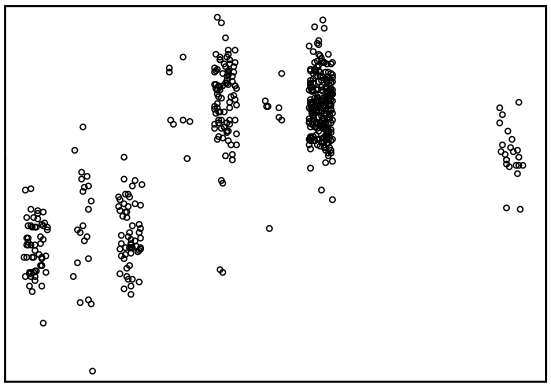


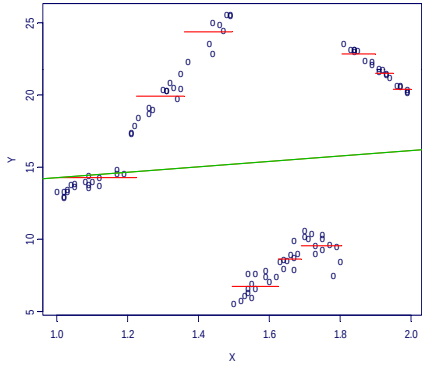




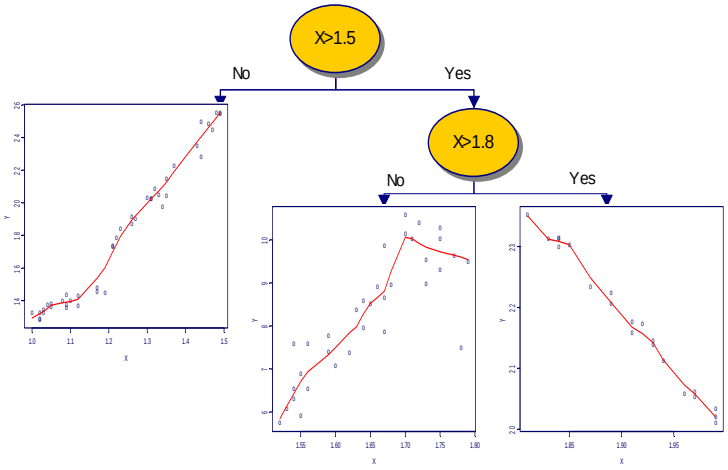
“Ensemble” Process Modeling

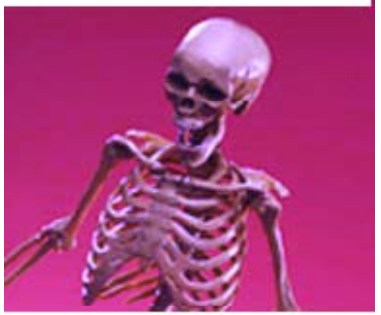
Regression Trees

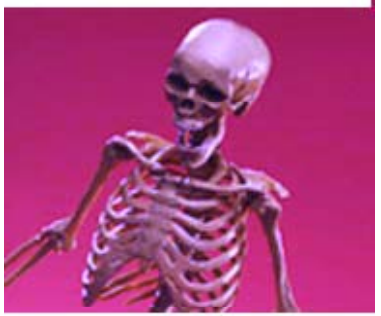




- Regression Tree:
  - Piecewise estimate of a regression function
  - Constructed by recursively partitioning the data and sample space

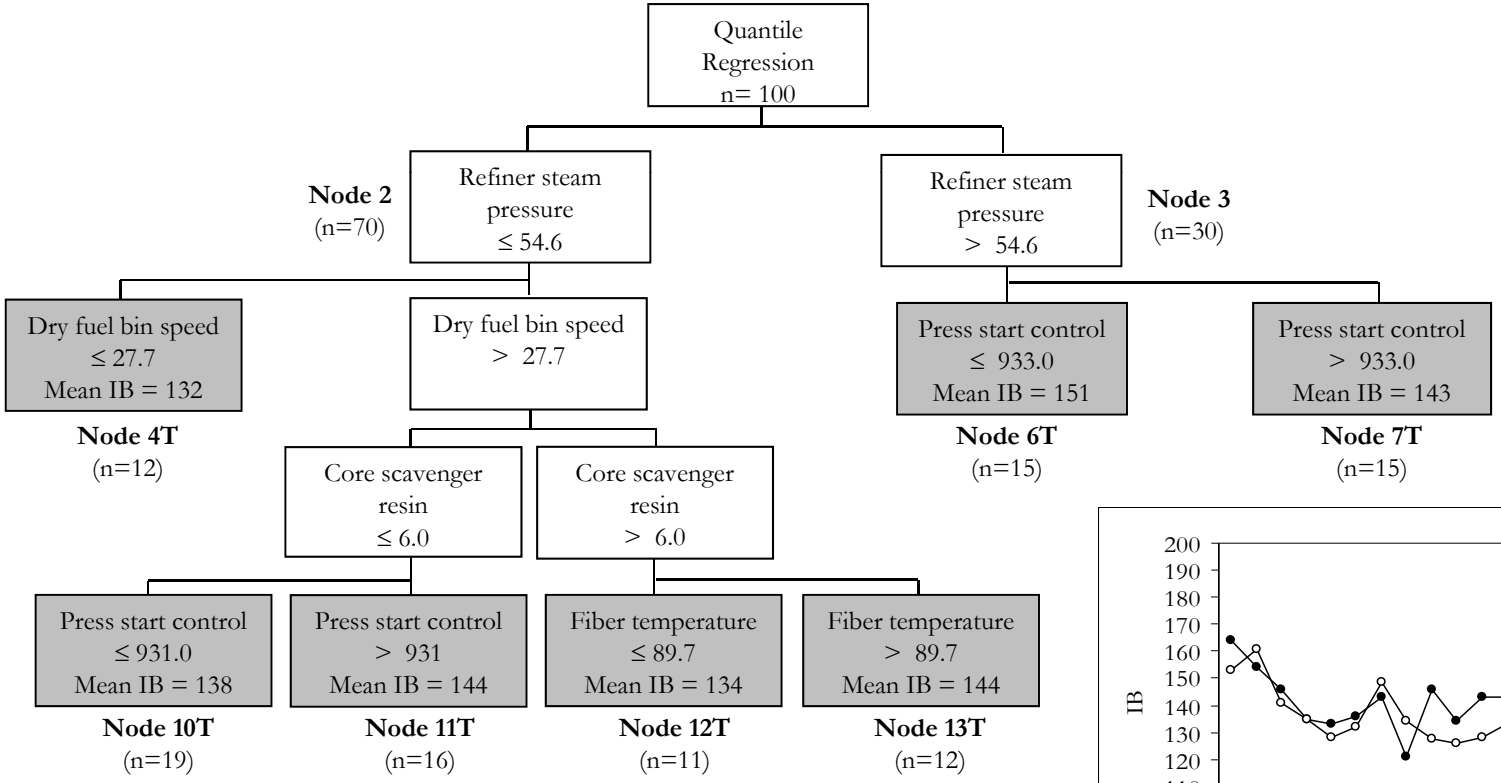


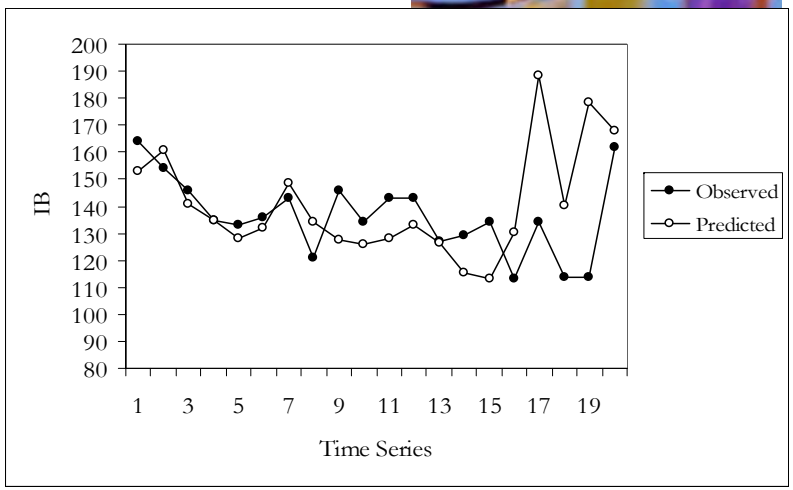




“Ensemble” Process Modeling

Regression Tree Models – “High Explanatory Value”





Box’s (1979): “All models are wrong but some are useful”

Training

"Statistical Thinking"

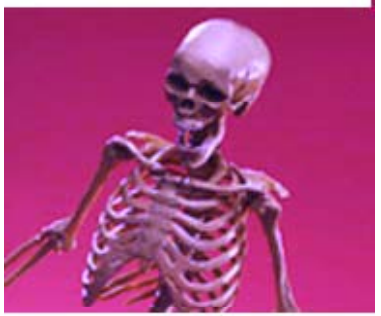
- Continuous improvement using statistical process control for the bio-based products industries
- Advanced statistical seminars for the bio-based products industries
- Applied design of experiments for the bio-based products industries

Conclusion

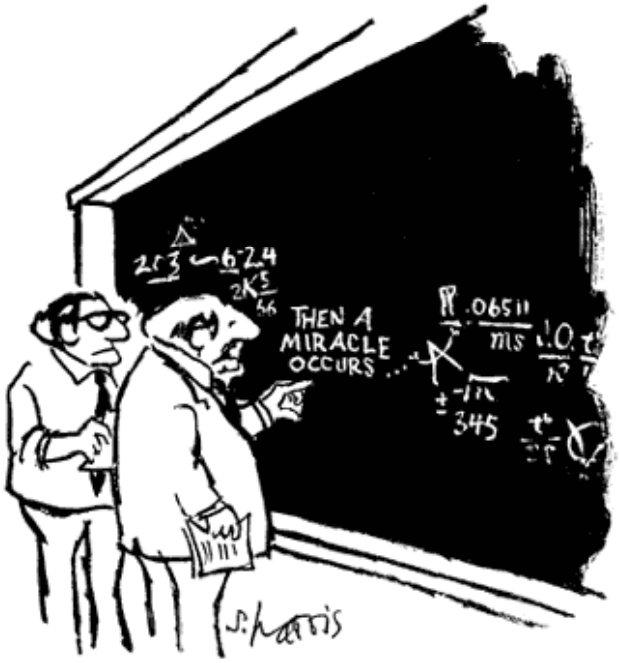
- Statistics is a key foundation of science
- <u>Bottom-line</u>: Statistics helps minimize the risk of being wrong

"What makes a scientist great is the care that he/she takes in telling you what is wrong with his/her results, so that you will not misuse them"

W. Edwards Deming (1900-1993)



Questions & Discussion



"I think you should be more explicit here in step two."

http://www.nytimes.com/2009/08/06/technology/06stats.html

"For Today's Graduate, Just One Word: Statistics"