Decision Analysis 1—Bayesian Updating

Dr. Dale M. Nesbitt
Management Science and Engineering
Stanford University
Huang Rm. 325, Stanford
221 State Street, Suite 1856
Los Altos, CA 94022
(650) 218-3069 mobile
dnesbitt@stanford.edu

Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis, Data, and Bayesian Updating

Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Problem As Communicated to Me

- Sandia National Laboratory during a job interview.
- Minuteman missiles armed with MIRV warheads.
- Unlike nuclear reactors, theses are specifically designed to go “boom”
- What is the probability that one will self detonate in our own silo?

2018, Dale M. Nesbitt, All Rights Reserved

Here's How We Have Approached the Problem

- Every year we select one at random
- We drag it up out of the hold and do comprehensive "destructive testing." We cut that sucker into tiny pieces and look at every piece and component looking for failure mechanisms. ("Sampling without replacement")
- We've been doing this about 20 times.
- We've never found even the slightest flaw or degradation anywhere.

2018, Dale M. Nesbitt, All Rights Reserved

Classical Statistician

- We have 0 failures out of 20 tests.
- The probability is therefore 0=0/20.
- “Congress and the President and the Joint Chiefs of Staff absolutely do not buy that.”
- How would you approach the problem?

Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

Here's My Prior on the Self Detonation Probability

- Mean: 1 in a billion.
- Standard deviation: ¼ in a billion.
- I will use a beta distribution to characterize my prior.

$$f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

- For integer values of α,

$$\Gamma(\alpha) = (\alpha-1)!$$

Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Parameters of the Beta Distribution

$$\mu = \frac{\alpha}{\alpha+\beta} = 1\times 10^{-9}$$

$$\text{std.dev.} = \sigma = \frac{1}{\alpha+\beta}\sqrt{\frac{\alpha\beta}{\alpha+\beta+1}} = 0.25\times 10^{-9}$$

$$\alpha = \frac{\mu}{\sigma^2}\left[(1-\mu)\mu - \sigma^2\right] = 16\left(1-10^{-9}\frac{17}{16}\right)$$

$$\beta = \frac{1-\mu}{\sigma^2}\left[(1-\mu)\mu - \sigma^2\right] = 16\left(1-10^{-9}\right)\left(10^9 - \frac{17}{16}\right)$$

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Your Success/Failure Probability Is Binomial

- It is your likelihood function, telling you the number of Failures and Successes you would have for a model with failure probability p

$$\{F,S \mid p\} = \binom{F}{S+F} p^{F} (1-p)^{S}$$

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Your Posterior Is the Product

$$\{p \mid F,S\} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \binom{F}{F+S} p^{F}(1-p)^{F+S}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \binom{F}{F+S} p^{\alpha+F-1}(1-p)^{\beta+F+S-1}$$

- All you have to do is add exponents to do Bayesian updating with the S/F data coming in

2018, Dale M. Nesbitt, All Rights Reserved

Is 3 Point Shooting Bernoulli?

- Does this look like a risk averse guy?
- What is his risk tolerance for points?

Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Likelihood Function (Past Shooting)

- Let Xi be 1 if we observe a "success" on the ith trial, otherwise 0, with probability p of success on each trial.
- Each X is 0 or 1; each X has a Bernoulli distribution. Suppose these Xs are conditionally independent given p.
- Bayes' theorem says that to find the conditional probability distribution of p given the data Xi, i = 1, ..., n, one multiplies the "prior" (i.e., marginal) probability measure assigned to p by the likelihood function

$$\{s \mid n, p\} = L(p) = \text{const} \times p^{s} (1-p)^{n-s}$$

- where s = x1 + ... + xn is the number of "successes" and n is of course the number of trials, and then normalizes, to get the "posterior" (i.e., conditional on the data) probability distribution of p.
- This is looking sort of “binomial,” isn't it. It is.
- We did this with multidetector trees

2018, Dale M. Nesbitt, All Rights Reserved

Is This Right?   Are We Done?

- Is this what we term an "unbiased" estimate of p? No; it is going to be "biased." (We are going to define "bias.")
- For s = 1, does this give the right answer to the law of succession?
- The answer is no, as Jaynes and Laplace showed!
- The reason is that the mode of the likelihood function is not enough to do the job.
- The mode of the likelihood is what maximum likelihood people are solely focused on

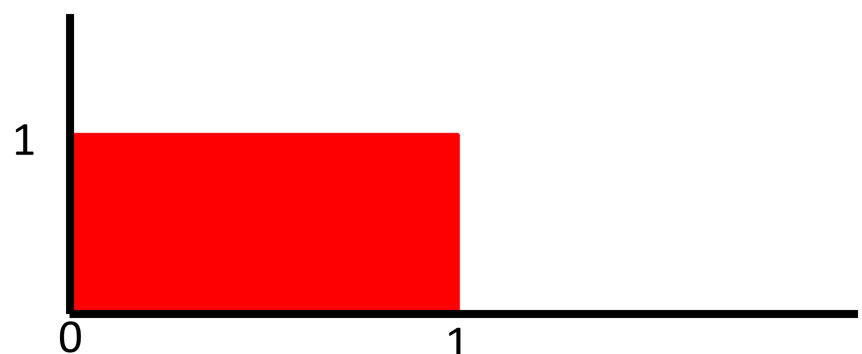
Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

A Laplacean Prior

- The prior probability density function that expresses total, abject ignorance of p except for the certain knowledge that it is neither 1 nor 0 (i.e., that we know that the experiment can in fact succeed or fail) is equal to 1 for $0 < p < 1$ and equal to 0 otherwise. To get the normalizing constant, we find
- Uniform density between 0 and 1



2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Define a Prior and Posterior in Light of the Form of the Likelihood

$$\text{Likelihood}(s \mid n) = \text{const} \times p^{s} (1-p)^{n-s}$$

$$\text{Prior}(p) = \text{const} \times p^{A} (1-p)^{B}$$

$$\text{Posterior}(p) = \text{const} \times p^{A+s} (1-p)^{B+n-s}$$

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Posterior Equals Prior Times Likelihood

- The probability distribution over p after we have seen s successes in n trials is binomial, derived from n Bernoulli trials

$$\{p \mid s, n\} = \text{const} \times p^{s} (1-p)^{n-s}$$

s successes

n trials

- This is looking binomial in structure, but it ISN'T. This pdf is over p, not s.
- The likelihood was binomial, but the prior and posterior are NOT

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Getting the Integrating Constant Is Tough

$$\{p \mid s,n\} = \frac{(n+1)!}{n!(n-s)!} p^{s}(1-p)^{n-s} = (n+1)\binom{n}{s} p^{s}(1-p)^{n-s}$$

- This can be manipulated to be the Beta distribution

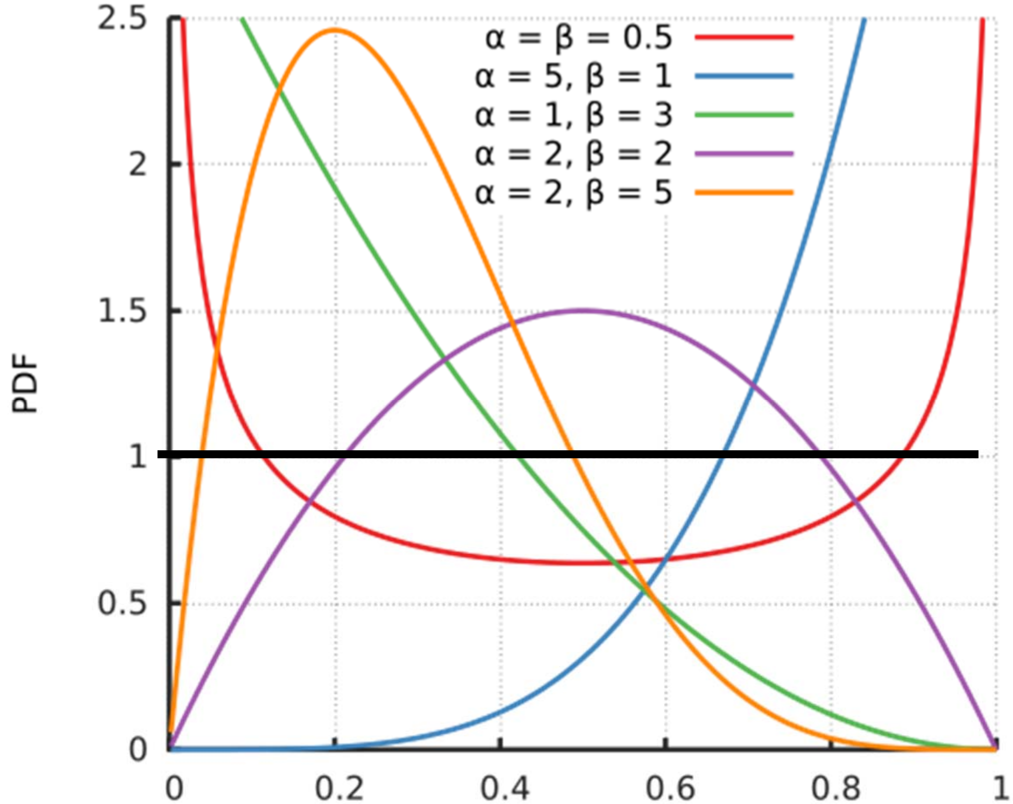
$$\{p \mid s,n\} = \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)!} p^{(s+1)-1}(1-p)^{(n-s+1)-1}$$

$$= \text{Beta}(s+1, n-s+1)$$

2018, Dale M. Nesbitt, All Rights Reserved

Beta Is Pretty Rich

- The uniform prior has the parameters of the Beta set to a=1,b=1. The previous calculation embedded that assumption.



Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

Where in the Heck Do You Think the Beta Distribution Came From???

- It came from repeated success and failure trials from the Bernoulli/Binomial.
- Bayes did this by 1761. How smart are we?
- You now know where it came from—conjugate prior for repeated Bernoulli trials with counting
  - Stephen Curry's scoring and Bryce Harper's hitting probability are governed by binomial probabilities, we think with beta prior and posterior
  - There was a study of Tim Hardaway that argued that his 3 point shots were indeed Bernoulli
  - Call up the Cavs and you can probability get a job!

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

We Have Just Derived the Beta Density

- The Beta probability density function (over Steph's shot probability p) is

$$\{p\} = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

2018, Dale M. Nesbitt, All Rights Reserved

Suppose

- Steph attempts 11.2 per game and makes 5.1 of them.
- In an 81 game season, Steph hits 5.1*81=413 three pointers out of 11.2*81=907 attempts. (He misses 494 three point shots)
- His likelihood function is

$$\{s \mid n, p\} = \binom{907}{413} p^{413} (1-p)^{494}$$

- The maximum likelihood estimate of p is p=413/907 = 0.455347, which is wrong

2018, Dale M. Nesbitt, All Rights Reserved

Under This Estimate, His Shot pdf Is Binomial

- So Steph walks onto the court and shoots 12 3 pointers. What is his pdf over points and shots made?
- Let's have a look at the classical prediction.

$$\{s \mid n\} = \frac{\Gamma(n+1)}{\Gamma(s+1)\Gamma(n-s+1)}(0.455347)^{s}(0.544653)^{n-s}$$

Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Spreadsheet

2018, Dale M. Nesbitt, All Rights Reserved

Parameters of the Beta Density over Shot Probability

- Mean $$\langle x \rangle = \frac{\alpha}{\alpha + \beta}$$
- Variance $$\mathrm{Var} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

2018, Dale M. Nesbitt, All Rights Reserved

Suppose the Prior is Beta

Decision Analysis

Dale M. Nesbitt

- Prior is Beta

$$\{p\} = \frac{\Gamma(\alpha_P + \beta_P)}{\Gamma(\alpha_P)\Gamma(\beta_P)} p^{\alpha_P - 1}(1-p)^{\beta_P - 1}$$

- Likelihood is Binomial

$$\{s \mid n, p\} = \frac{\Gamma(n+1)}{\Gamma(s+1)\Gamma(n-s+1)} p^{s}(1-p)^{n-s}$$

- Posterior is Beta

$$\{p\} = \frac{\Gamma(\alpha_P + \beta_P + n)}{\Gamma(\alpha_P + s)\Gamma(\beta_P + n - s)} p^{(\alpha_P + s) - 1}(1-p)^{(\beta_P + n - s) - 1}$$

2018, Dale M. Nesbitt, All Rights Reserved

Is the Beta Density Sufficiently Rich So That You Can Approximate a Wide Range of Priors?

- Generally yes.
- You can have the uniform prior we had before with the two beta parameters set to 1

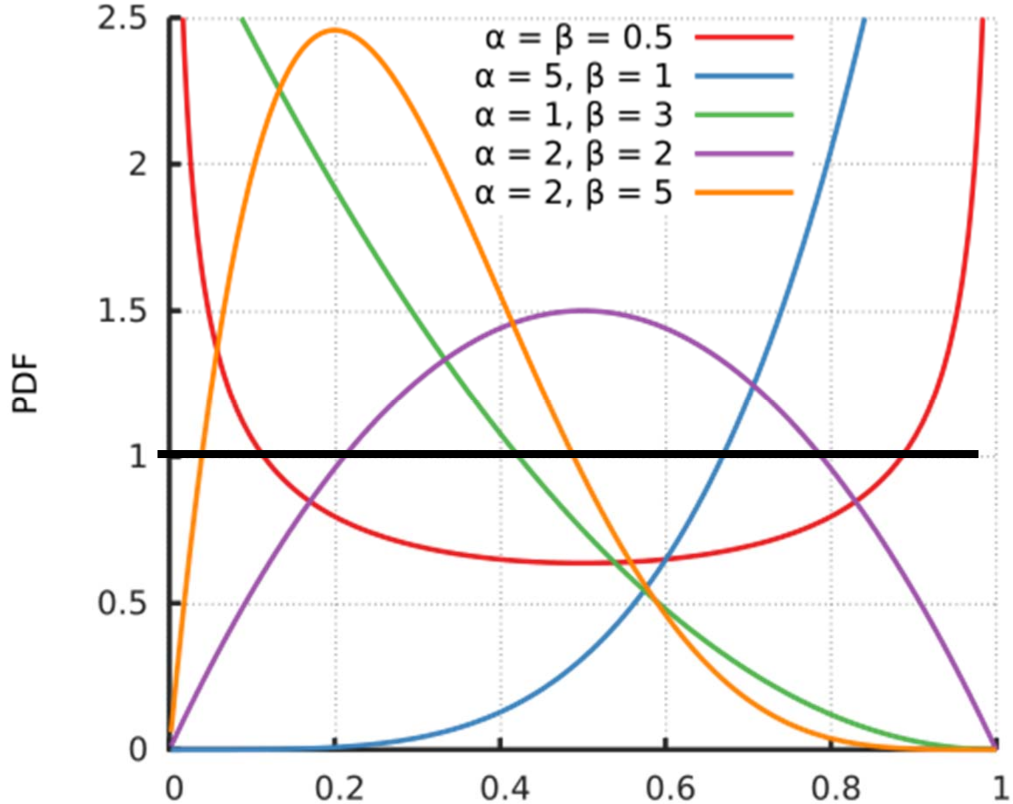
Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

Beta Is Pretty Rich

- The uniform prior has the parameters of the Beta set to a=1,b=1. The previous calculation embedded that assumption.



2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Classic Example of Conjugate Prior

Exact same functional forms; only numerical parameters are different

$$Prior * Likelihood = Posterior$$

Requires a very specific likelihood function

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Classic Example of Conjugate Prior

Exact same functional forms; only numerical parameters are different

Binomial

$$\underbrace{\text{Prior}}_{\text{Beta}} * \text{Likelihood} = \underbrace{\text{Posterior}}_{\text{Beta}}$$

Requires a very specific likelihood function

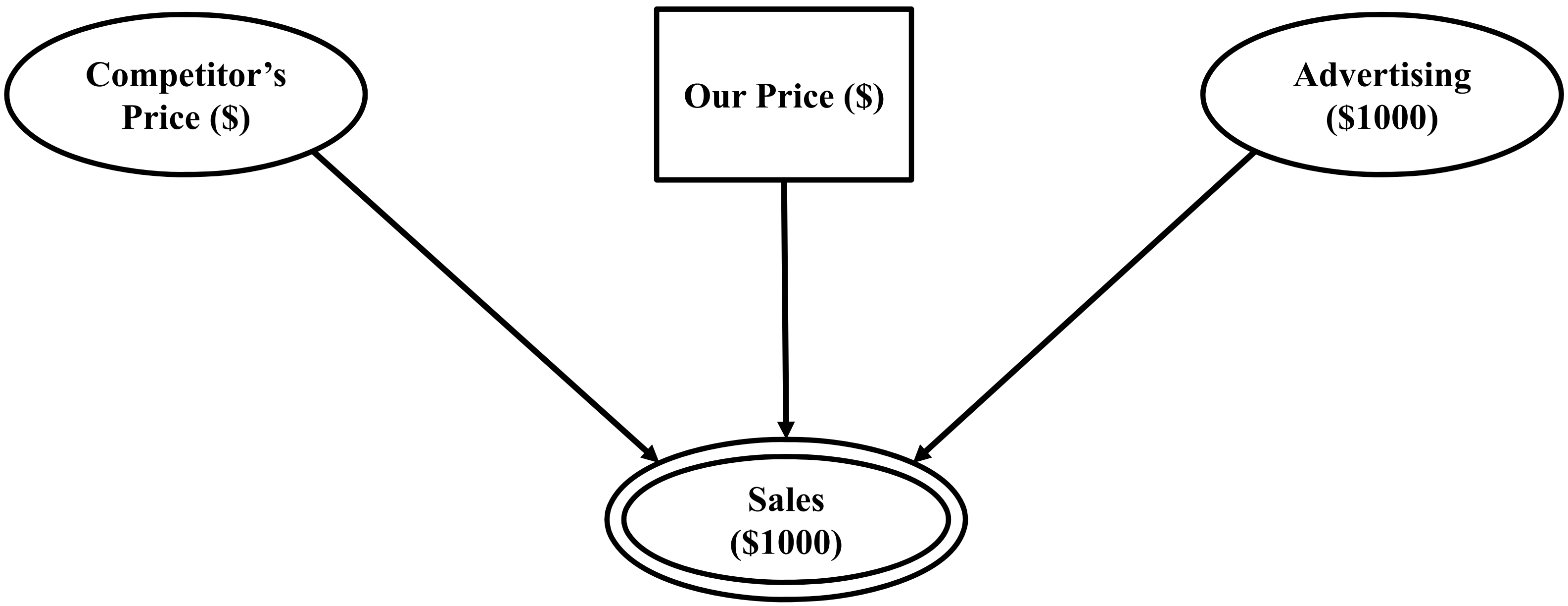
2018, Dale M. Nesbitt, All Rights Reserved

Can We Use Regression? Statistics? Excel to get a pdf?

Yes, sort of.

2018, Dale M. Nesbitt, All Rights Reserved

Clemen’s Advertising Problem



- We want a model in the Sales node. It is a linear function of coefficients.

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Clemen Has Historical Data Claimed to Be Relevant

- Here is the data base that he has collected regarding advertising, our price, competition price, and sales

Observation	Constant Int	Advertising ($1000s) Ad	Price ($) P	Competition Price ($) CP	Sales ($1000s) S
1	1	366	90.99	96.95	10541
2	1	377	90.99	93.99	8891
3	1	387	94.99	90.99	5905
4	1	418	96.99	97.95	8251
5	1	434	92.99	97.95	11461
6	1	450	95.95	93.95	6924
7	1	457	93.95	90.99	7347
8	1	466	91.95	96.95	10972
9	1	467	96.95	94.99	7811
10	1	468	92.95	96.95	10559
11	1	468	97.99	98.95	9825
12	1	475	91.95	90.99	9130
13	1	479	99.95	91.95	5116
14	1	479	96.99	95.95	7830
15	1	481	91.95	90.95	8388
16	1	490	96.99	96.99	8588
17	1	494	96.95	91.95	6945
18	1	502	98.95	95.95	7697
19	1	505	94.99	96.99	9655
20	1	529	93.99	97.95	11516
21	1	532	91.99	95.99	11952
22	1	533	92.99	97.99	13547
23	1	542	93.99	92.95	9168
24	1	544	90.95	95.95	11942
25	1	547	94.99	93.95	9917
26	1	554	89.95	90.95	10666
27	1	556	96.95	95.95	9717
28	1	560	91.99	97.95	13457
29	1	561	98.99	97.95	10319
30	1	566	93.95	91.99	9731
31	1	566	94.99	94.99	10279
32	1	582	98.99	91.99	7202
33	1	609	89.95	92.99	12103
34	1	612	92.95	92.99	11482
35	1	617	92.95	94.95	11944
36	1	623	94.99	91.99	9188

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

He Wonders Whether a Statistical Model Is Suitable for Decision Analysis

- Can you just fit the coefficients and run a bunch of sensitivities and get the answer.
- Everybody does it!
- Can they all be wrong?
- Yep.
- Hint: Is there anything of a probabilistic nature that can be inferred from regression?

2018, Dale M. Nesbitt, All Rights Reserved

Here's What Excel Gives You

Decision Analysis

Dale M. Nesbitt

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.999027
R Square	0.998056
Adjusted R	0.966623
Standard E	459.0979
Observatio	36

ANOVA

	df	SS	MS	F	ignificance F
Regression	4	3.46E+09	8.66E+08	4106.45	9.54E-42
Residual	32	6744669	210770.9		
Total	36	3.47E+09			

	Coefficients	andard Erro	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	lpper 95.0%
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Int	2199.342	3839.736	0.572785	0.570794	-5621.94	10020.63	-5621.94	10020.63
Ad	15.0466	1.172569	12.83216	3.67E-14	12.65816	17.43505	12.65816	17.43505
P	-503.764	28.34356	-17.7735	3.84E-18	-561.498	-446.03	-561.498	-446.03
CP	499.6713	30.55929	16.35088	4.29E-17	437.424	561.9185	437.424	561.9185

2018, Dale M. Nesbitt, All Rights Reserved

Classical Statistics/Regression

- This is a fundamental review of classical linear regression
- It is very, very hard to find this in the literature in a form that is accessible to decision analysts and Bayesians
- I have worked hard to get this together and definitive
- https://onlinecourses.science.psu.edu/stat501/node/250

Dale M. Nesbitt

Decision Analysis

2018, Dale M. Nesbitt, All Rights Reserved

The Classic Linear Regression Model

- Suppose we take a series of n observations of some performance measure (designated y) together with a vector of p attendant independent variables that prospectively have a contributing effect to that performance measure.
- The linear regression model that attempts to characterize these observations conjectures that the dependent variable y can be "predicted" by the following linear equation in which the unknowns are the coefficients β_1, β_2,…, β_p

$$y = \beta_1 + \sum_{k=2}^{p} \beta_k x_k$$

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Eliminating the Intercept

- We think of x_1 as being unity for every observation.
- Setting x_1 to unity allows us to write the foregoing linear equation in the general form

$$y = \sum_{k=1}^{p} \beta_k x_k$$

- Secure in the knowledge we can consider an intercept or not at our discretion without loss of generality.
- Under this assumption, it must be kept in mind that p counts the constant as well as the nonconstant coefficients.

2018, Dale M. Nesbitt, All Rights Reserved

The Observations

Observation	Independent Variables	Dependent Variable
1	$x_{11}, \ldots, x_{1p}$	y_1
2	$x_{21}, \ldots, x_{2p}$	y_2
.	.	.
.	.	.
.	.	.
n	$x_{n1}, \ldots, x_{np}$	y_n

2018, Dale M. Nesbitt, All Rights Reserved

Dale M. Nesbitt

Table of n Observations

	y	x_1	x_2	…	x_p
1		1			
2		1			
.		.			
.		.			
.		.			
n		1			

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

This Implies the Overdetermined Set of Equations

Observed **Predicted** **"Error"**

$$y_1 - \left(x_{11},\ldots, x_{1p}\right)\beta = \varepsilon_1$$

$$y_2 - \left(x_{21},\ldots, x_{2p}\right)\beta = \varepsilon_2$$

$$\ldots$$

$$y_n - \left(x_{n1},\ldots, x_{np}\right)\beta = \varepsilon_n$$

2018, Dale M. Nesbitt, All Rights Reserved

Assume the Error is Governed by a Normal with Mean 0 and Nonzero Variance

Decision Analysis

Dale M. Nesbitt

$$N(0,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\varepsilon_i^2}{2\sigma^2}}$$

The joint forecasting error, with NO RELEVANCE BETWEEN ERRORS, is thereby assumed to be.

$$f(\varepsilon_1,\ldots,\varepsilon_n) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\varepsilon_1^2}{2\sigma^2}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\varepsilon_2^2}{2\sigma^2}} \ldots \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\varepsilon_n^2}{2\sigma^2}}$$

$$= \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\left(\varepsilon_1^2+\varepsilon_2^2+\ldots+\varepsilon_n^2\right)}$$

There are problems here (like observation relevance or technique) but let us proceed as they do.

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Joint Forecasting Error

$$\varepsilon_1^2 + \varepsilon_2^2 + \ldots + \varepsilon_n^2 = \varepsilon^T \varepsilon$$

- Thus the joint error (with no relevance between terms) is

$$f\left(\varepsilon_1, \ldots, \varepsilon_n\right) = \frac{1}{\sigma^n \left(2\pi\right)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\varepsilon^T \varepsilon}$$

2018, Dale M. Nesbitt, All Rights Reserved

The Previous Overdetermined Set of Equations Is Written

$$\varepsilon = y - X\beta$$

2018, Dale M. Nesbitt, All Rights Reserved

The Observations

Observation	Independent Variables	Dependent Variable
1	$x_{11},\ldots,x_{1p}$	y_1
2	$x_{21},\ldots,x_{2p}$	y_2
.	.	.
.	.	.
.	.	.
n	$x_{n1},\ldots,x_{np}$	y_n

Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Table of n Observations

	Int	Ad	P	CP	Sales
Int		X			y
Ad					
P					
CP					

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Define

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}$$

$$X = \begin{bmatrix} x_{11} & x_{12} & . & . & x_{1p} \\ x_{21} & x_{22} & . & . & x_{2p} \\ . & . & & & . \\ . & . & & & . \\ . & . & & & . \\ x_{n1} & x_{n2} & . & . & x_{np} \end{bmatrix}. \qquad X = \begin{bmatrix} 1 & x_{12} & . & . & x_{1p} \\ 1 & x_{22} & . & . & x_{2p} \\ . & . & & & . \\ . & . & & & . \\ . & . & & & . \\ 1 & x_{n2} & . & . & x_{np} \end{bmatrix}.$$

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

The (Overdetermined) Equation

$$y \; - \; X \times \beta \; = \; \varepsilon$$

2018, Dale M. Nesbitt, All Rights Reserved

The Calculations That Excel Does Under the Covers (Stat. 101)

	Int	Ad	P	CP
Int	36	18296	3400.96	3411.8
Ad	18296	9454524	1728295.16	1733148.44
P	3400.96	1728295.16	321558.0044	322343.6064
CP	3411.8	1733148.44	322343.6064	323576.314

$$X^T X$$

	Int	Ad	P	CP
Int	69.95068263	-0.005567355	-0.319147816	-0.389810452
Ad	-0.005567355	6.52329E-06	1.35852E-06	2.24088E-05
P	-0.319147816	1.35852E-06	0.00381152	-0.000439171
CP	-0.389810452	2.24088E-05	-0.000439171	0.004430736

$$\left(X^T X\right)^{-1}$$

	XTy
Int	345966
Ad	177849135
P	32561320.38
CP	32878377.14

$$X^T y$$

n	36	n = number of observatior
p	4	p = number of coefficients
$\nu_C = n - p$	32	ν = deg rees of freedom
$\nu_C s_C^2$	6744669.357	$R = (y - X\beta)^T (y - X\beta)$

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis
Dale M. Nesbitt

Decision Analysis

Dale M. Nesbitt

The Results—A Students t Distribution with the Following Mean and Variance

	Mean
Int	2199.342251
Ad	15.04660288
P	-503.7640378
CP	499.6712512

$$\bar{\beta} = \left(X^T X\right)^{-1} X^T y$$

210770.9174 $s_C^{\;2} = \dfrac{R}{\nu_C}$

224822.3119 $\dfrac{\nu_C}{\nu_C - 2} s_C^2$

3610362280 $\dfrac{2}{\nu_C - 4}\left(s_C^2 \dfrac{\nu_C}{\nu_C - 2}\right)^2$

Variance Covariance

	Int	Ad	P	CP
Int	15726474.19	-1251.66552	-71751.54976	-87638.08706
Ad	-1251.66552	1.466580347	0.305426674	5.038003627
P	-71751.54976	0.305426674	856.9148069	-98.7353528
CP	-87638.08706	5.038003627	-98.7353528	996.1283795

$$\frac{\nu_C}{\nu_C - 2} s_C^2 \left(X^T X\right)^{-1}$$

Correlation

	Int	Ad	P	CP
Int	1.0000	-0.2606	-0.6181	-0.7002
Ad	-0.2606	1.0000	0.0086	0.1318
P	-0.6181	0.0086	1.0000	-0.1069
CP	-0.7002	0.1318	-0.1069	1.0000

2018, Dale M. Nesbitt, All Rights Reserved

What Does This Mean?

- It means that we have derived a probability distribution over the coefficients. (Nobody ever told you that, but they certainly should have.)
- That means that with settings of the independent variables, we have a probability distribution over the dependent variable (sales).
- Nobody in statistics really tells you what to do with that.
  - Decision analysis will tell you what to do with that.
  - Make a probabilistic projection!

2018, Dale M. Nesbitt, All Rights Reserved

The Joint Density Over all n Observations Is Assumed to Be a Product of Independent Normal Distributions

Decision Analysis

Dale M. Nesbitt

- The likelihood function was the joint density over all n observations

$$f(\varepsilon_1,\ldots,\varepsilon_n)=\prod_{i=1}^{n}\frac{1}{\sigma(2\pi)^{\frac{1}{2}}}e^{-\frac{\varepsilon_i^2}{2\sigma^2}}=\frac{1}{\sigma^n(2\pi)^{\frac{n}{2}}}\prod_{i=1}^{n}e^{-\frac{\varepsilon_i^2}{2\sigma^2}}=\frac{1}{\sigma^n(2\pi)^{\frac{n}{2}}}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2}$$

$$=\frac{1}{\sigma^n(2\pi)^{\frac{n}{2}}}e^{-\frac{1}{2\sigma^2}\varepsilon^T\varepsilon}=\frac{1}{\sigma^n(2\pi)^{\frac{n}{2}}}e^{-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)}$$

Sum of squared errors—Where do you suppose OLS came from?

- From a pdf perspective, the likelihood function is

$$\{\text{Observations}\mid\text{Coefficients}\}=\{y,X\mid\beta,\sigma\}$$

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Joint Forecasting Error Rewritten

- This joint forecasting error is written

$$\{y, X \mid \beta, \sigma\} = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}(y - X\beta)^T (y - X\beta)}$$

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Check Out That Exponent in the pdf

- It sure as heck looks like a multivariate quadratic in β, doesn't it?

$$\{y, X \mid \beta, \sigma\} = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}(y - X\beta)^T (y - X\beta)}$$

2018, Dale M. Nesbitt, All Rights Reserved

Remember When You Completed the Square?

- You added and subtracted an unknown number a from x in a quadratic equation and set a so as to eliminate linear terms

$$y=x^2+4x-7=\left[\left(x-a\right)+a\right]^2+4\left[\left(x-a\right)+a\right]-7$$

$$=\left(x-a\right)^2+2a\left(x-a\right)+a^2+4\left(x-a\right)+4a-7$$

$$=\left(x-a\right)^2+\left(2a+4\right)\left(x-a\right)+a^2+4a-7$$

Let $a=-2$ to eliminate linear term

$$\Rightarrow y=\left(x+2\right)^2+0\left(x+2\right)+\left(-2\right)^2+4\left(-2\right)-7=\left(x+2\right)^2-11$$

Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

Complete the Square in the Matrix Sense

$$y - X\beta = y - X\left(\beta - \bar{\beta} + \bar{\beta}\right) = \left(y - X\bar{\beta}\right) - X\left(\beta - \bar{\beta}\right)$$

Add and subtract

- So if $z = \beta - \bar{\beta}$ then

$$\left(y - X\beta\right) = \left(y - X\bar{\beta}\right) - Xz \Rightarrow \left(y - X\beta\right)^T = \left(y - X\bar{\beta}\right)^T - z^T X^T$$

$$\Rightarrow \left(y - X\beta\right)^T \left(y - X\beta\right) = \left[\left(y - X\bar{\beta}\right)^T - z^T X^T\right]\left[\left(y - X\bar{\beta}\right) - Xz\right]$$

$$= \underbrace{\left(y - X\bar{\beta}\right)^T \left(y - X\bar{\beta}\right)}_{\text{Constant}} - \underbrace{2z^T X^T \left(y - X\bar{\beta}\right)}_{\text{Linear}} + \underbrace{z^T X^T X z}_{\text{Quadratic}}$$

- Set $\bar{\beta}$ so that the middle (linear) term is zero

$$X^T\left(y - X\bar{\beta}\right) = 0 \Rightarrow X^T y - \left(X^T X\right)\bar{\beta} = 0 \Rightarrow \bar{\beta} = \left(X^T X\right)^{-1} X^T y$$

- This is the classical regression solution. This is what comes out of Excel
  - It is the highest point on the likelihood function. The mode.
  - We get it simply by completing the square. No statistics.
  - We didn't have to maximize anything or invent any "estimators" or anything like that

2018, Dale M. Nesbitt, All Rights Reserved

Completing the Square (in a Matrix Sense) Shows Us the Classical Regression Mean Coefficient Values

Decision Analysis

Dale M. Nesbitt

- Substituting for $\bar{\beta}$

$$(y - X\beta)^T (y - X\beta) = \left(\beta - \bar{\beta}\right)^T X^T X \left(\beta - \bar{\beta}\right) + R$$

where

$$\bar{\beta} = \left(X^T X\right)^{-1} X^T y$$

$$R = \left(y - X\bar{\beta}\right)^T \left(y - X\bar{\beta}\right) = \text{"residual" sum of squared error}$$

- We have not altered the exponent at all; we have merely restructured it. No statistics or regression have been done! We have merely completed the square in the likelihood function.
- Substitute the exponent back into the likelihood function.

2018, Dale M. Nesbitt, All Rights Reserved

The Likelihood Function

- It is the product of a gamma distribution times a normal distribution

$$\ell\{y, X \mid \beta, \sigma\} = \frac{1}{(2\pi)^{\frac{n}{2}}} \sigma^{-n} e^{-\frac{1}{2\sigma^2}R} e^{-\frac{1}{2\sigma^2}(\beta-\bar{\beta})^T X^T X(\beta-\bar{\beta})}$$

Univariate gamma distribution over σ^2

Multivariate normal distribution over the β coefficients

This separation is going to be profound

Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

The Likelihood Function

$$\{y,X\mid\beta,\sigma\}=\frac{1}{\sigma^{n}\left(2\pi\right)^{\frac{n}{2}}}e^{-\frac{1}{2\sigma^{2}}R}e^{-\frac{1}{2}\left(\beta-\bar{\beta}\right)^{T}\left(\frac{X^{T}X}{\sigma^{2}}\right)\left(\beta-\bar{\beta}\right)}\equiv L\left(\beta,\sigma\right)$$

- Aggregate the constant and write

$$\{y,X\mid\beta,\sigma\}=c_{2}\sigma^{-n}e^{-\frac{1}{2\sigma^{2}}R}e^{-\frac{1}{2}\left(\beta-\bar{\beta}\right)^{T}\left(\frac{X^{T}X}{\sigma^{2}}\right)\left(\beta-\bar{\beta}\right)}$$

2018, Dale M. Nesbitt, All Rights Reserved

Clemen Wants to Do Some Linear Regression to Fit His Model

Decision Analysis

Dale M. Nesbitt

- He calculates the standard statistical results (which he could get automatically with regression in Excel—almost, but not quite!)

$$\left(X^T X\right)^{-1} \qquad X^T X \qquad X^T y \qquad \bar{\beta} = \left(X^T X\right)^{-1} X^T y$$

$$n = \text{number of observations}$$

$$p = \text{number of coefficients}$$

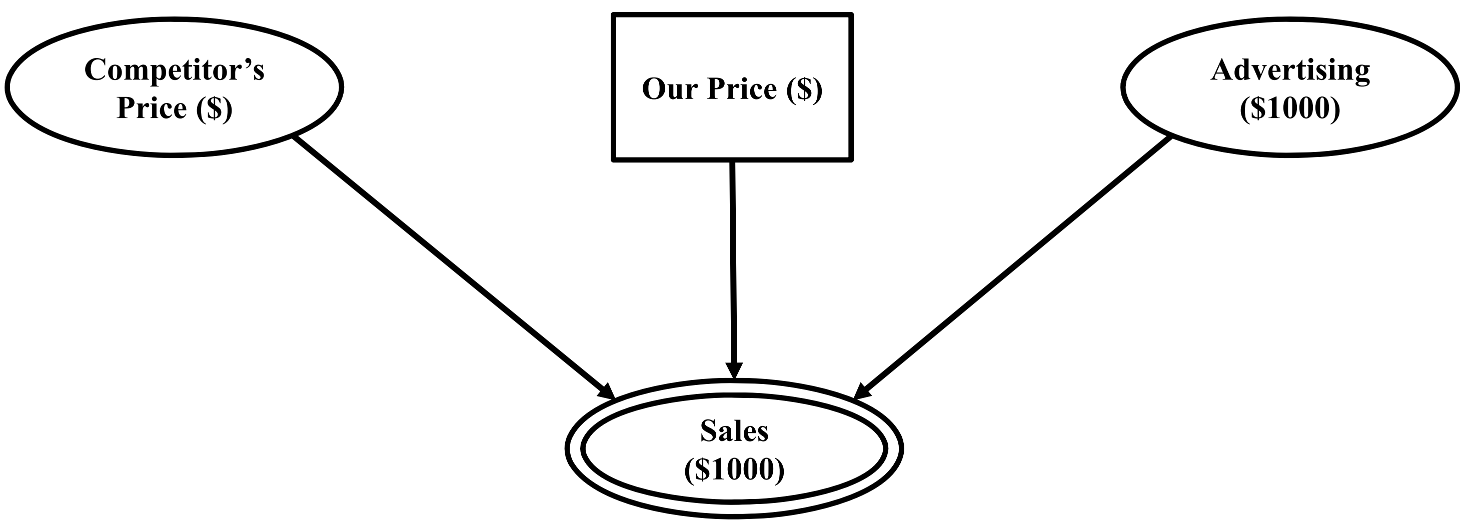
$$\nu = \text{degrees of freedom}$$

2018, Dale M. Nesbitt, All Rights Reserved

Step 1: Postulate an Elemental Possibility

- For each elemental possibility (with the intercept frozen at 1)

D	Int	Ad	P	CP
1	1	505	95	97



Dale M. Nesbitt

Decision Analysis

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

You Have the Joint pdf Over Coefficients

- You could sample using Monte Carlo if you wanted.
- That would track out a derived density over sales given the conditioning variables D.
- However, there is a closed form that precludes this.
- We will not derive it, but you can use it.

2018, Dale M. Nesbitt, All Rights Reserved

Step 2: Implement the Predictive Distribution (Which Is Univariate)

$$\{d|X,y,D\} = \frac{\Gamma\left(\frac{\nu_C+q}{2}\right)}{\left\|I+D(X^TX)^{-1}D^T\right\|^{\frac{1}{2}}\Gamma\left(\frac{\nu_C}{2}\right)\left(\frac{\nu_C s_C^2}{2}\right)^{\frac{q}{2}}(2\pi)^{\frac{q}{2}}}\left[1+(d-D\bar{\beta})^T\frac{\left[I+D(X^TX)^{-1}D^T\right]^{-1}}{\nu_C s_C^2}(d-D\bar{\beta})\right]^{-\frac{\nu_C+q}{2}}$$

$$= c_0\left[1+(d-D\bar{\beta})^T\frac{\left[I+D(X^TX)^{-1}D^T\right]^{-1}}{\nu_C s_C^2}(d-D\bar{\beta})\right]^{-\frac{\nu_C+q}{2}}$$

- D is a row vector, so this equation is a univariate distribution
- It gives you the PDF over d for the elemental possibility D.
- It is univariate Students' t in form.
- You can find the mean and the variance and use the approximate formula for certain equivalent
- An exact formula for certain equivalent does not exist.

Decision Analysis

Dale M. Nesbitt

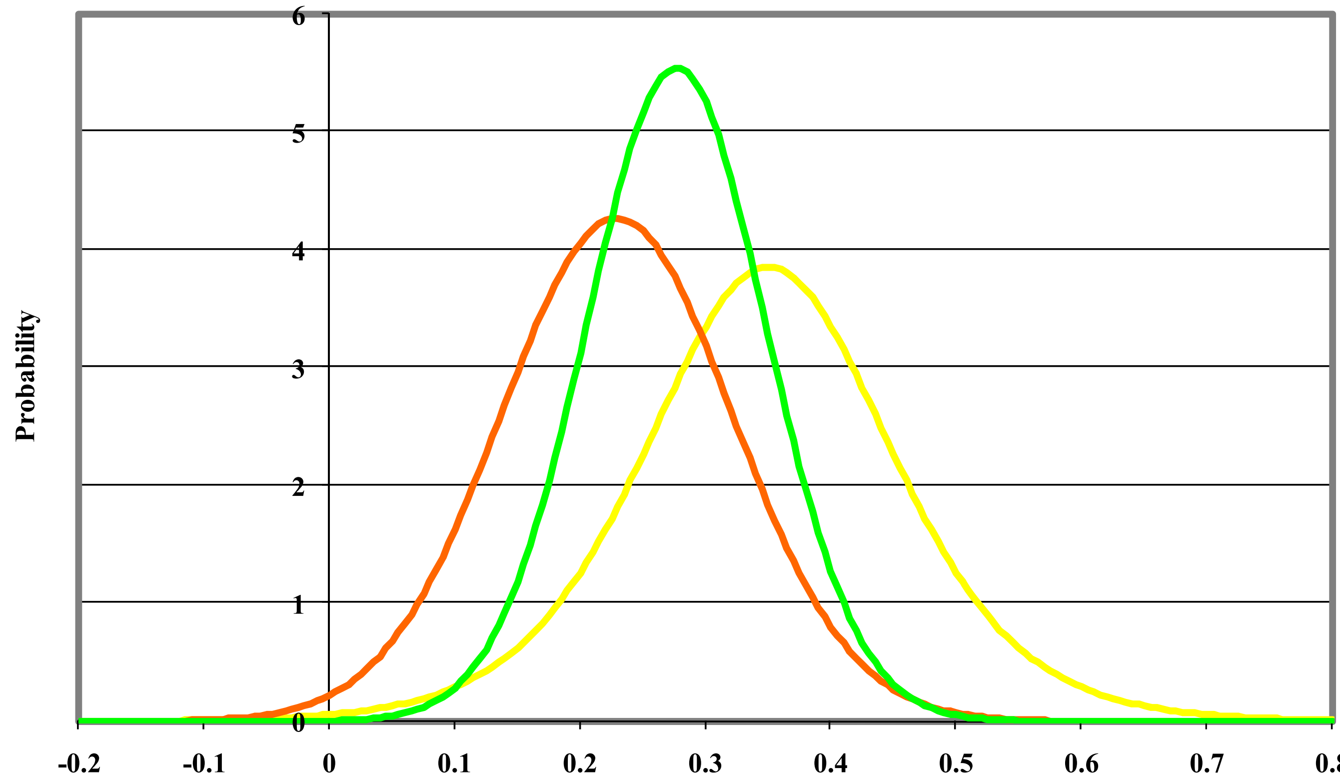
2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

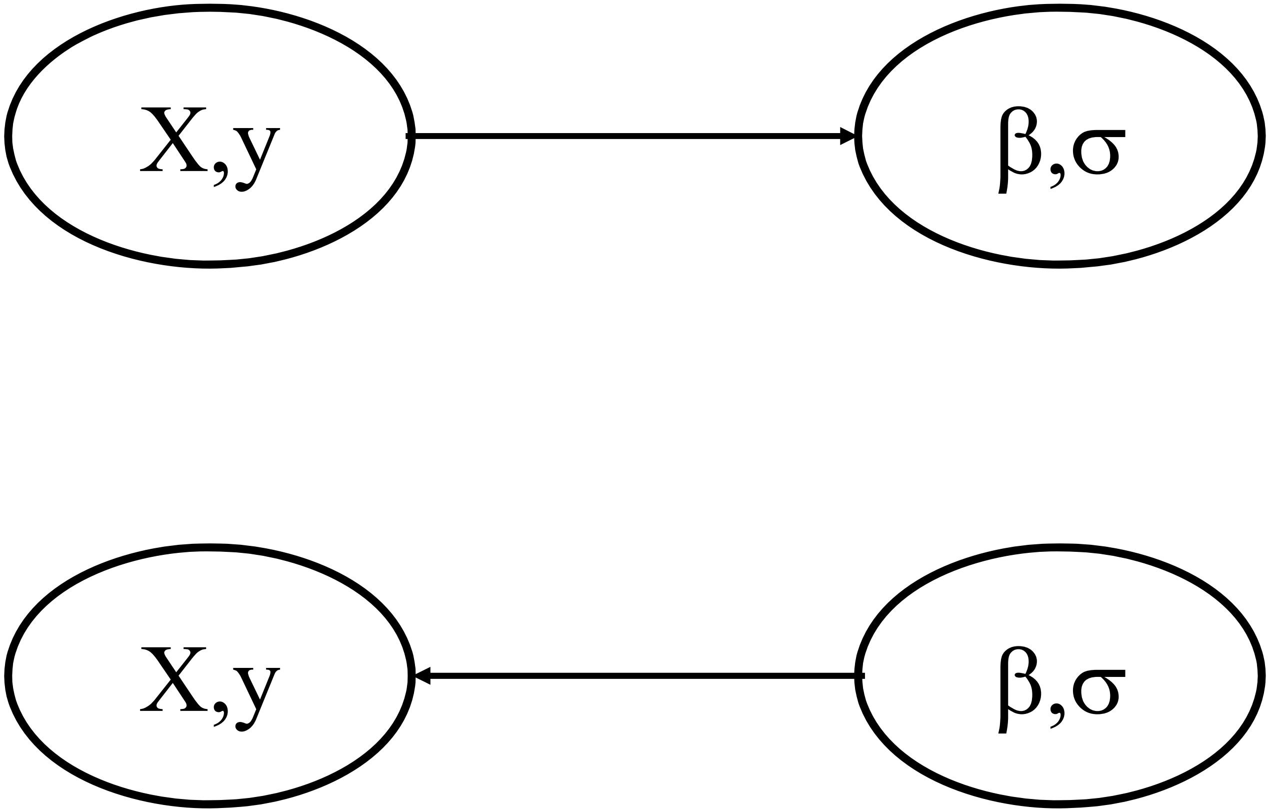
It Gives Distributions That Look Like This

- These distributions can be discretized and used in tree and relevance diagram calculations (e.g., simulation, moment matching)



2018, Dale M. Nesbitt, All Rights Reserved

We Know All About Relevance, Don't We?



2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Bayes Theorem—The Most Fundamental View

Observations

Model coefficients

Bayes Theorem

$$\{X, y, \beta, \sigma\} = \{\beta, \sigma | X, y\}\{X, y\} = \{X, y | \beta, \sigma\}\{\beta, \sigma\}$$

so

$$\{\beta, \sigma | X, y\} = \frac{\{X, y | \beta, \sigma\}\{\beta, \sigma\}}{\{X, y\}}$$

$$= \text{const} * \{X, y | \beta, \sigma\}\{\beta, \sigma\}$$

"Likelihood function"

"Prior"

- Bayes approaches the problem at the outset from a probabilistic perspective; no approximations other than the linear model (which can be extended)

2018, Dale M. Nesbitt, All Rights Reserved

Decision
Analysis

Dale M. Nesbitt

Let's Coalesce the Constants to See What This Functional Form Looks Like

$$\{X, y \mid \beta, \sigma\} = \text{const} * \sigma^{-n} e^{-\frac{1}{2\sigma^2}\left[\nu_C s_C^2 + (\beta-\bar{\beta})^T X^T X (\beta-\bar{\beta})\right]}$$

Negative power
Scalar
Mean
Matrix quadratic

- It occurred to Zellner and others: "Why don't we think about a prior with an entirely parallel type of form?"

$$\{\beta, \sigma\} = \text{const} * \sigma^{-m} e^{-\frac{1}{2\sigma^2}\left[M + (\beta-\beta_0)^T Q (\beta-\beta_0)\right]}$$

Negative power
Scalar
Mean
Matrix quadratic

- The prior needs to characterize what you think the coefficients are with your OLD data (or just a guess).

2018, Dale M. Nesbitt, All Rights Reserved

Prior Density Over Coefficients in Exactly Parallel (Conjugate) Form

Constant plus a quadratic

$$\{\beta,\sigma\} = const * \sigma^{-m} e^{-\frac{1}{2\sigma^2}\left[M+(\beta-\beta_0)^T Q(\beta-\beta_0)\right]}$$

- There are four parameters we must subjectively specify to comprise our prior
  - The constant scalar power on the s term: m
  - The additive scalar constant in the exponent: M
  - The vector of means (length p) in the quadratic portion of the exponent: β_0
  - The (p x p) matrix in the quadratic portion of the exponent: Q
  - The knowledge of the experts should be embedded in the values of m, M, β_0, and Q that are assumed.
  - They comprise judgment regarding what the model parameters should be based on experience, knowledge, etc.

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Multiply Prior Times Likelihood to Get Posterior—Bayes Theorem

$$\{X, y | \beta, \sigma\}\{\beta, \sigma\} = \{\beta, \sigma | X, y\}$$

$$= \left\{ \text{const} * \sigma^{-n} e^{-\frac{1}{2\sigma^2}\left[\nu_C s_C^2 + (\beta-\bar{\beta})^T X^T X(\beta-\bar{\beta})\right]} \right\} \left\{ \text{const} * \sigma^{-m} e^{-\frac{1}{2\sigma^2}\left[M + (\beta-\beta_0)^T Q(\beta-\beta_0)\right]} \right\}$$

$$= \text{const} * \sigma^{-(n+m)} e^{-\frac{1}{2\sigma^2}\left[\nu_C s_C^2 + M + (\beta-\bar{\beta})^T X^T X(\beta-\bar{\beta}) + (\beta-\beta_0)^T Q(\beta-\beta_0)\right]}$$

- This posterior is a probability distribution over model coefficients given model observations (after model observations).
  - It has a mean, which we are going to denote b* even though we don’t know what it is yet.
  - It has a variance/covariance matrix, and we don’t know what that is yet either.

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Complete the Square

Let $z=\beta-\beta^*$

$$\text{Exponent}=\nu_C s_C^2+M+\left[z+\left(-\bar\beta+\beta^*\right)\right]^T\left(X^TX\right)\left[z+\left(-\bar\beta+\beta^*\right)\right]$$

$$+\left[z+\left(-\beta_0+\beta^*\right)\right]^T Q\left[z+\left(-\beta_0+\beta^*\right)\right]$$

$$=\nu_C s_C^2+M+\left[z+\left(-\bar\beta+\beta^*\right)\right]^T\left[\left(X^TX\right)z+\left(X^TX\right)\left(-\bar\beta+\beta^*\right)\right]$$

$$+\left[z+\left(-\beta_0+\beta^*\right)\right]^T\left[Qz+Q\left(-\beta_0+\beta^*\right)\right]$$

$$=\nu_C s_C^2+M+z^T\left(X^TX\right)z+\left(-\bar\beta+\beta^*\right)^T\left(X^TX\right)z+z^T\left(X^TX\right)\left(-\bar\beta+\beta^*\right)+\left(-\bar\beta+\beta^*\right)^T\left(X^TX\right)\left(-\bar\beta+\beta^*\right)$$

$$+z^TQz+\left(-\beta_0+\beta^*\right)^TQz+z^TQ\left(-\beta_0+\beta^*\right)+\left(-\beta_0+\beta^*\right)^TQ\left(-\beta_0+\beta^*\right)$$

$$=\nu_C s_C^2+M+\left(-\bar\beta+\beta^*\right)^T\left(X^TX\right)\left(-\bar\beta+\beta^*\right)+\left(-\beta_0+\beta^*\right)^TQ\left(-\beta_0+\beta^*\right)$$

$$+\left(-\bar\beta+\beta^*\right)^T\left(X^TX\right)z+z^T\left(X^TX\right)\left(-\bar\beta+\beta^*\right)+\left(-\beta_0+\beta^*\right)^TQz+z^TQ\left(-\beta_0+\beta^*\right)$$

$$+z^T\left(X^TX\right)z+z^TQz$$

$$=\nu_C s_C^2+M+\left(\beta^*-\bar\beta\right)^T\left(X^TX\right)\left(\beta^*-\bar\beta\right)+\left(\beta^*-\beta_0\right)^TQ\left(\beta^*-\beta_0\right)\quad \textbf{Constant term}$$

$$+2z^T\left[\left(X^TX\right)\left(\beta^*-\bar\beta\right)+Q\left(\beta^*-\beta_0\right)\right]\quad \textbf{Linear term}$$

$$+z^T\left(X^TX+Q\right)z\quad \textbf{Quadratic term}$$

2018, Dale M. Nesbitt, All Rights Reserved

Zero Out the Linear Term (Complete the Square)

Decision Analysis

Dale M. Nesbitt

$$+2z^T\left[\left(X^TX\right)\left(\beta^*-\bar{\beta}\right)+Q\left(\beta^*-\beta_0\right)\right]=0$$

$$\left(X^TX\right)\left(\beta^*-\bar{\beta}\right)+Q\left(\beta^*-\beta_0\right)=0$$

$$\left(X^TX+Q\right)\beta^*=\left(X^TX\right)\bar{\beta}+Q\beta_0=\left(X^TX\right)\left(X^TX\right)^{-1}X^Ty+Q\beta_0$$

$$\left(X^TX+Q\right)\beta^*=\left(X^Ty+Q\beta_0\right)$$

$$\beta^*=\left(X^TX+Q\right)^{-1}\left(X^Ty+Q\beta_0\right)$$

2018, Dale M. Nesbitt, All Rights Reserved

Completing the Square

- Here is that linear term rewritten

$$\beta^* = \left(X^T X + Q\right)^{-1} \left(X^T y + Q\beta_0\right)$$

- This is the mean value of the Bayesian posterior, the Bayesian posterior mean value of the linear coefficients.
- **This is FANTASTIC!!!!!!!!!!!**
- It is sort of a “weighted average of the prior and the classical, but it is a very precise and special weighted average.

Dale M. Nesbitt

Decision Analysis

2018, Dale M. Nesbitt, All Rights Reserved

Substitute This Expression (Which Eliminates the First Order, Linear Term) into the Posterior

Decision Analysis

Dale M. Nesbitt

- The final expression is

$$\text{Exponent} = A + \left(\beta - \beta^*\right)^T \left(X^T X + Q\right)\left(\beta - \beta^*\right)$$

- in which

$$A = \nu_C s_C^2 + M + \left(\beta^* - \bar{\beta}\right)^T \left(X^T X\right)\left(\beta^* - \bar{\beta}\right) + \left(\beta^* - \beta_0\right)^T Q \left(\beta^* - \beta_0\right)$$

2018, Dale M. Nesbitt, All Rights Reserved

When We Complete the Square, Here Is the Posterior Density Quadratic

- We haven't done ANY statistics yet. We have just multiplied prior times likelihood to get posterior and all we have done is completed the square. This is so elegant!
- Prior, likelihood, and posterior all have the same mathematical form—conjugate.

$$\{\beta,\sigma|X,y\} = \text{const} * \sigma^{-(n+m)} e^{-\frac{1}{2\sigma^2}A} e^{-\frac{1}{2\sigma^2}(\beta-\beta^*)^T\left(X^TX+Q\right)(\beta-\beta^*)}$$

in which

$$\beta^* = \left(X^TX+Q\right)^{-1}\left(X^Ty+Q\beta_0\right)$$

$$A = \nu_C s_C^2 + M + \left(\beta^*-\bar{\beta}\right)^T\left(X^TX\right)\left(\beta^*-\bar{\beta}\right)+\left(\beta^*-\beta_0\right)^T Q\left(\beta^*-\beta_0\right)$$

Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

This Is Profound—Posterior is "Mix" of Prior and Likelihood

Decision Analysis

Dale M. Nesbitt

- This is the mean (and mode) of the posterior
- It is a very special "matrix weighted average" of the prior and likelihood.
- This is so, so, so intuitive when you think of prior times likelihood and think of these terms in the exponent.
- It allows an arbitrary number of variables in your linear model.

$$Q^{-1}Q\beta_0 = \beta_0$$

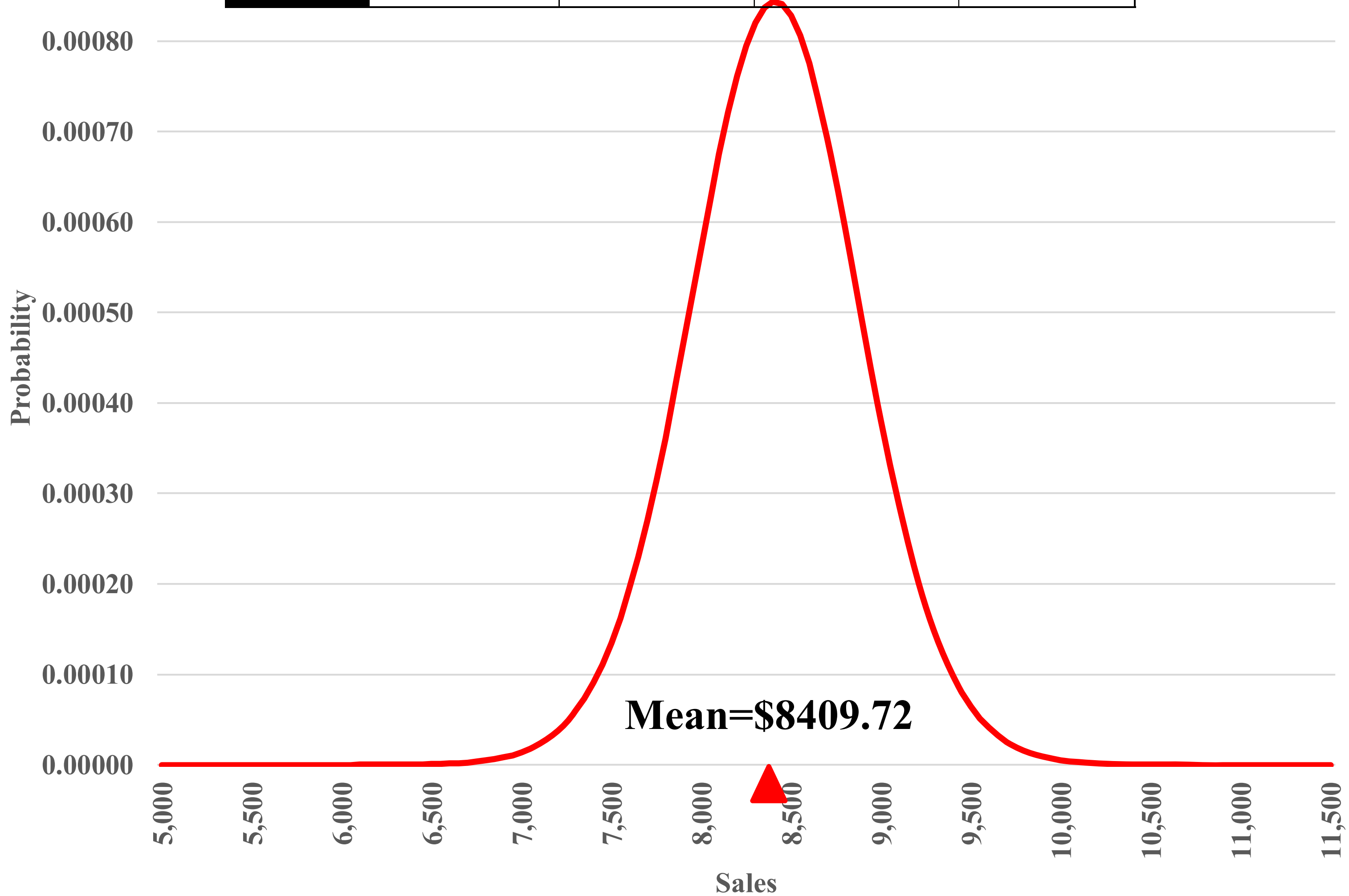
$$\beta^* = \left(X^T X + Q\right)^{-1}\left(X^T y + Q\beta_0\right)$$

$$\left(X^T X\right)^{-1} X^T y = \bar{\beta}$$

2018, Dale M. Nesbitt, All Rights Reserved

Conditional PDF Over Sales ($1000) Using Classical Regression (We'll See Later)

	Int	Ad	P	CP
1	1	505	95	97



Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Clemen

- He does not recognize the reality that we get an entire pdf over sales conditional on the three inputs to the sales node.
- He only considers that we get an expected value conditional on the three inputs.
- Knowing that we get the whole distribution really buys us the farm.
- We have a perfect model of conditional density, which is what we want.
- We might have to discretize.

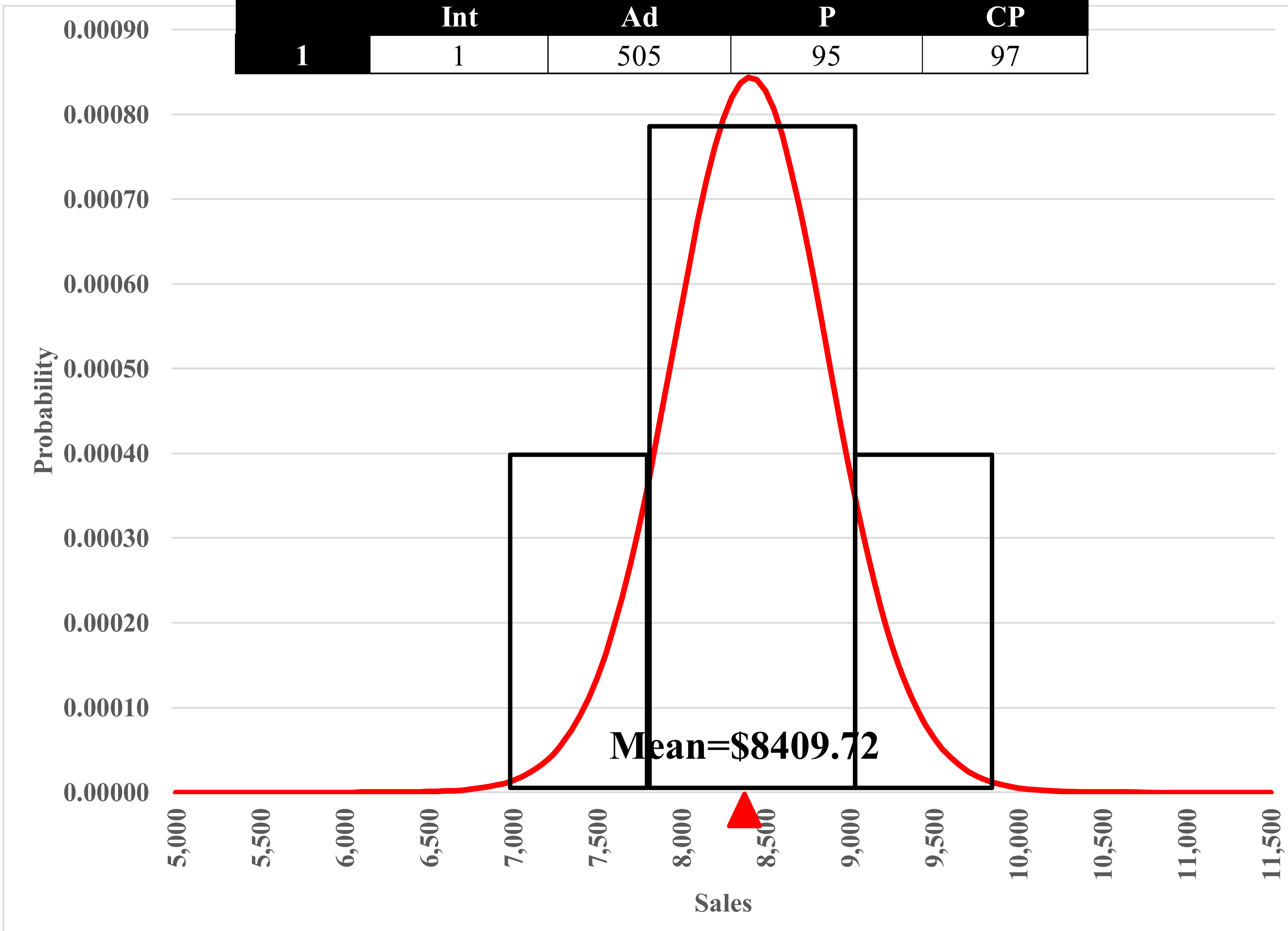
2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Discretize the Conditional PDF Over Sales ($1000)

	Int	Ad	P	CP
1	1	505	95	97



2018, Dale M. Nesbitt, All Rights Reserved

So “Big Data” Can Work?

- Yes in theory, usually not in practice.
- In theory, the model that is linear in coefficients is pretty good, and the probabilistic predictions it makes are pretty good.
- However, in the real world, data is often troubled and incomplete
  - Multicollinearity
  - Omitted variables
  - Uneven time sequences
  - Adverse section bias
  - Too early in the life cycle
- “Big Data” is harder than Decision Analysis!

2018, Dale M. Nesbitt, All Rights Reserved

Back to Clemen's Problem

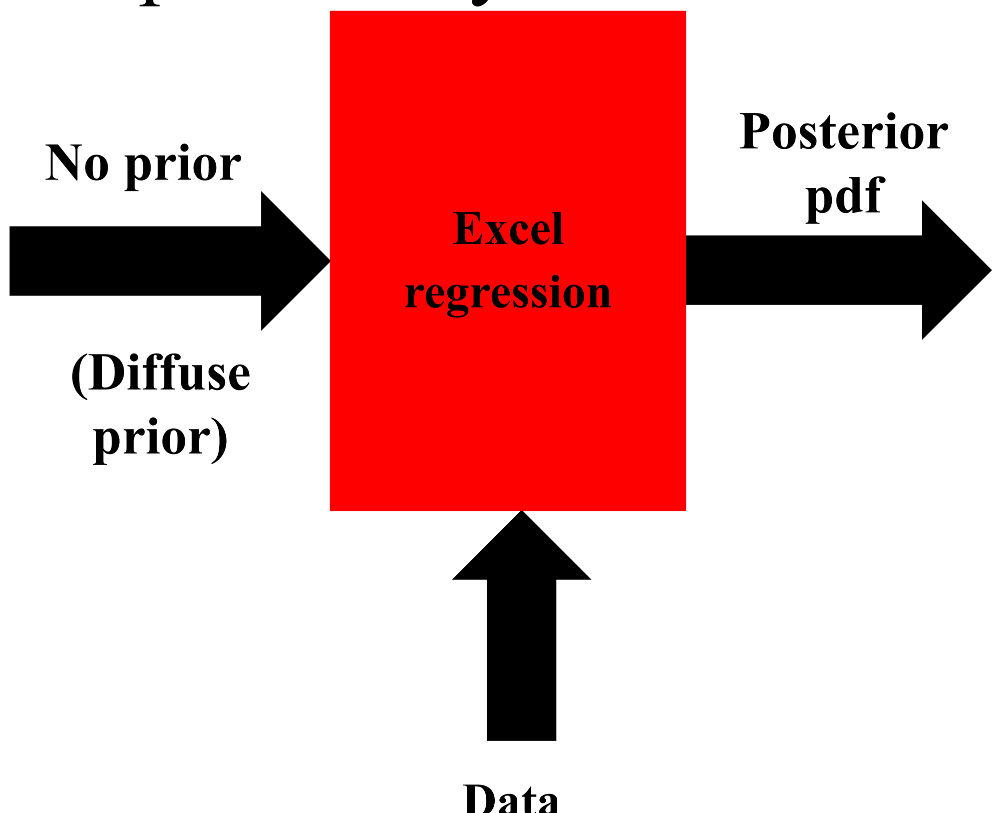
2018, Dale M. Nesbitt, All Rights Reserved

Decision
Analysis

Dale M. Nesbitt

Here Is How Classical Statistics Looks

- Gathering data is like an “experiment.” The more experimental results you have, the better predictor you have.



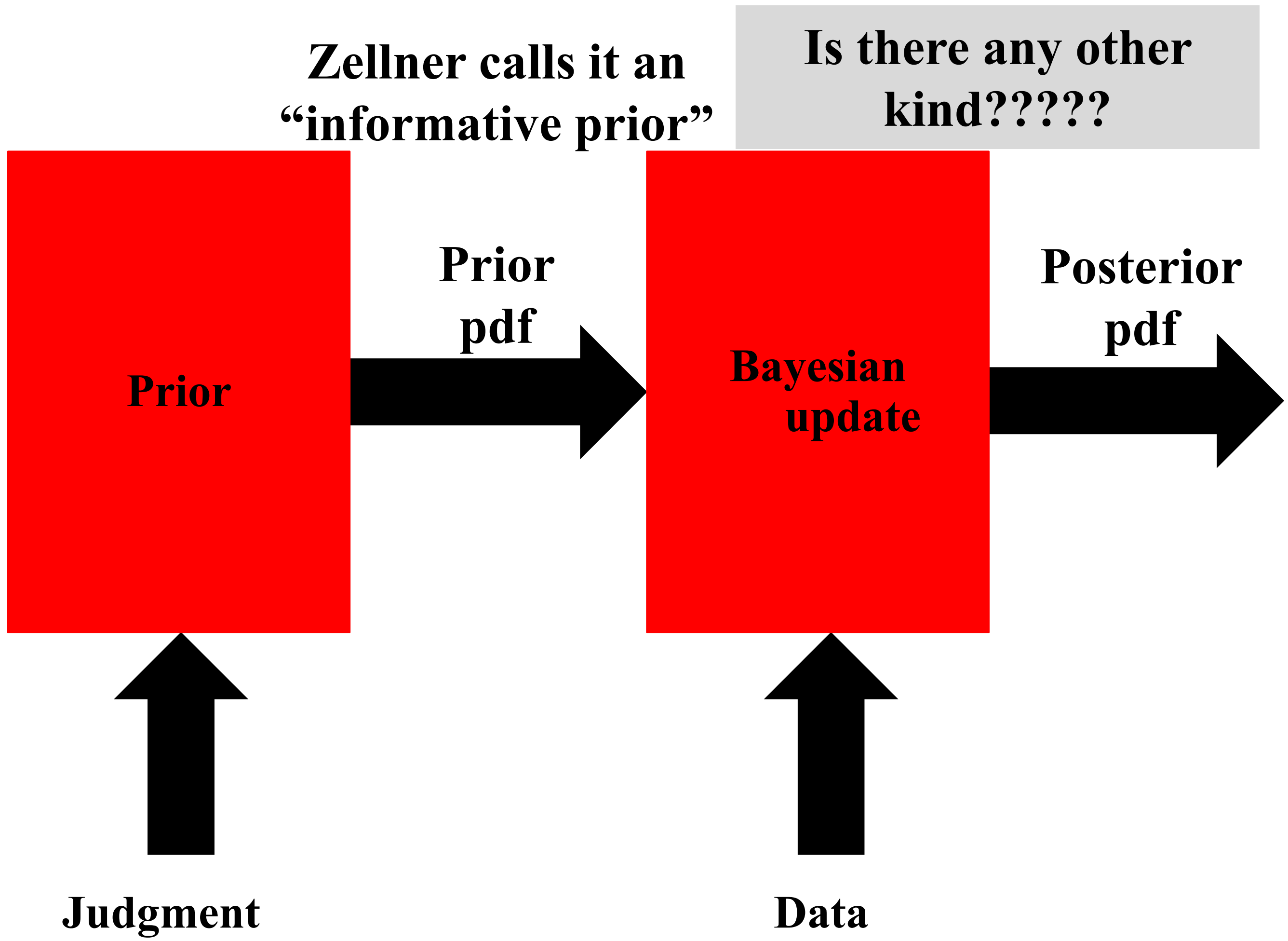
2018, Dale M. Nesbitt, All Rights Reserved

The Data Is an "Experiment"

- What if the data is problematic?
- What if the experiment gives you nothing?
- What if you need some probabilistic judgment?

2018, Dale M. Nesbitt, All Rights Reserved

Here Is How Bayesian Statistics Looks



2018, Dale M. Nesbitt, All Rights Reserved

You Start with a Prior Over the Model Coefficients

- You need a prior because your data may be problematic or incomplete.
- You may have knowledge of contributory relevances.
- You usually have some knowledge, perhaps with a very wide variance

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Assemble Your Prior Knowledge

	Int	Ad	P	CP
1	1	505	95	97

Means of coefficients

	β_0
Intercept	2100
Ad	20
P	-400
CP	400

Variances of coefficients

	VCV	(variance covariance)		
Intercept	784.00	0	0	0
Ad	0	0.07111	0	0
P	0	0	28.4444	0
CP	0	0	0	28.4444

Mean and Std. Dev. of Error Term

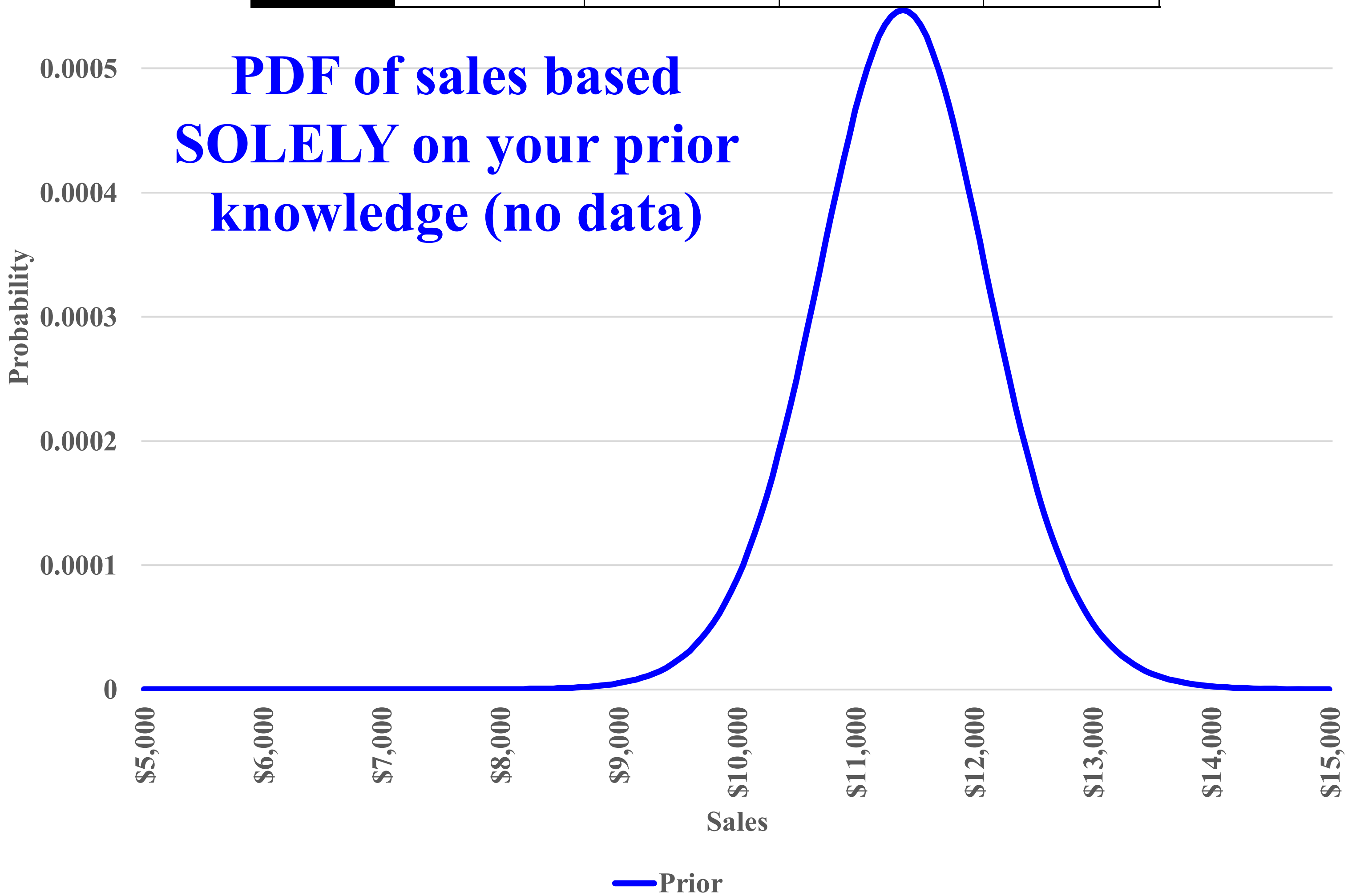
$\langle \sigma^2 \rangle$	50000
$\sqrt{Var_{s^2}}$	16667

I can make a Student's t density out of this

2018, Dale M. Nesbitt, All Rights Reserved

Here Is What Your Conditional Predictive Distribution Looks Like

	Int	Ad	P	CP
1	1	505	95	97



Decision Analysis

Dale M. Nesbitt

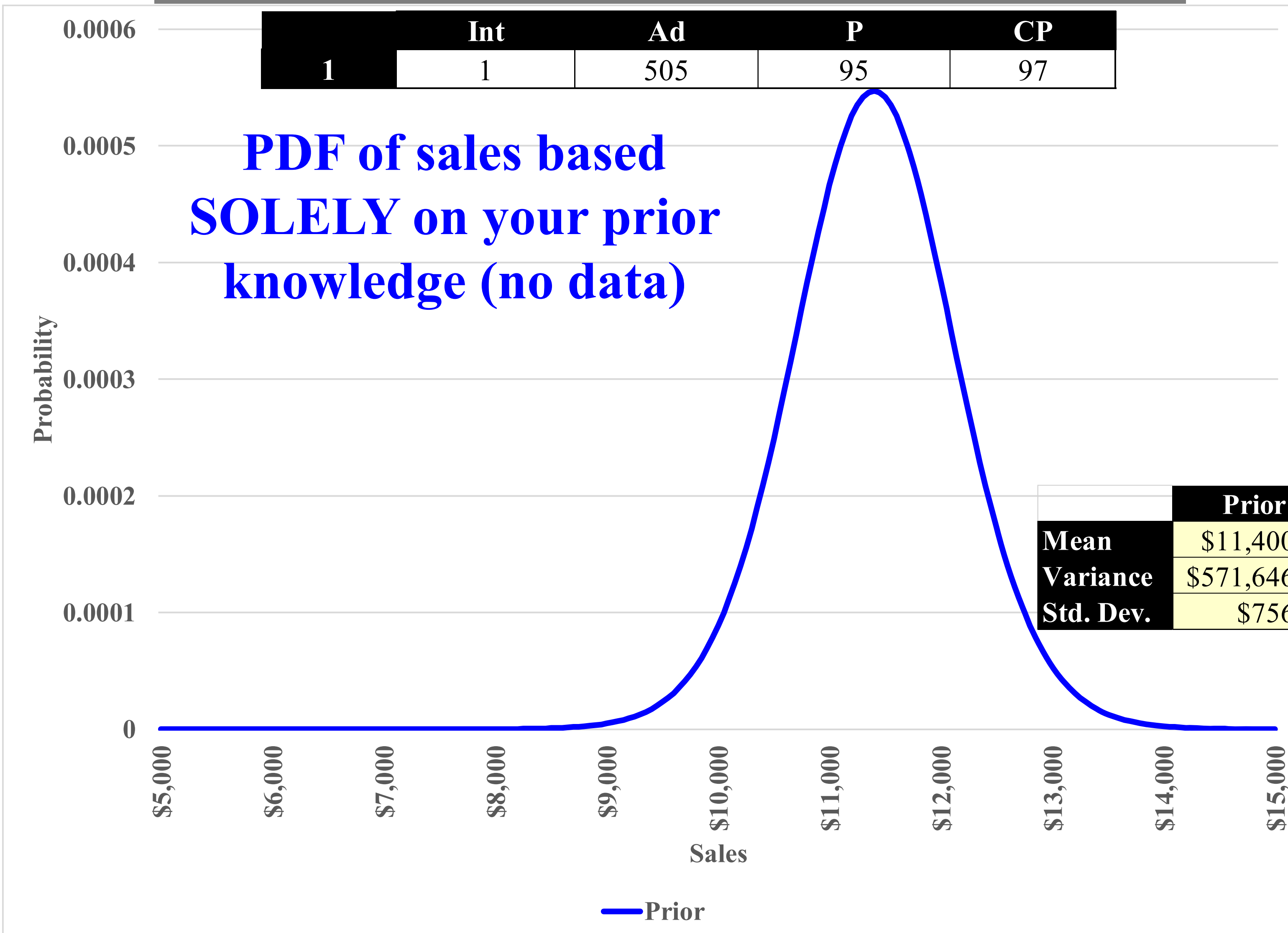
2018, Dale M. Nesbitt, All Rights Reserved

Here Is What Your Conditional Predictive Distributions Over Sales Looks Like

Decision Analysis

Dale M. Nesbitt

	Int	Ad	P	CP
1	1	505	95	97



	Prior
Mean	$11,400.00
Variance	$571,646.22
Std. Dev.	$756.07

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

The Prediction of Sales Based Solely on the Prior

	Int	Ad	P	CP
1	1	505	95	97

	Prior
Mean	$11,400.00
Variance	$571,646.22
Std. Dev.	$756.07

2018, Dale M. Nesbitt, All Rights Reserved

But, but, but, … There Is a Bunch of Data Out There

- Either it has appeared as a result of someone else's efforts.
- You have paid a fortune to gather it.
- You have bought it from a data vendor.
- "We want our decisions to be data driven."

2018, Dale M. Nesbitt, All Rights Reserved

Clemen Has Historial Data Claimed to Be Relevant

- Here is the data base that he has collected regarding advertising, our price, competitor price, and sales
- He is going to build a model linear in coefficients and fit them to this data!

Observation	Constant Int	Advertising ($1000s) Ad	Price ($) P	Competition Price ($) CP	Sales ($1000s) S
1	1	366	90.99	96.95	10541
2	1	377	90.99	93.99	8891
3	1	387	94.99	90.99	5905
4	1	418	96.99	97.95	8251
5	1	434	92.99	97.95	11461
6	1	450	95.95	93.95	6924
7	1	457	93.95	90.99	7347
8	1	466	91.95	96.95	10972
9	1	467	96.95	94.99	7811
10	1	468	92.95	96.95	10559
11	1	468	97.99	98.95	9825
12	1	475	91.95	90.99	9130
13	1	479	99.95	91.95	5116
14	1	479	96.99	95.95	7830
15	1	481	91.95	90.95	8388
16	1	490	96.99	96.99	8588
17	1	494	96.95	91.95	6945
18	1	502	98.95	95.95	7697
19	1	505	94.99	96.99	9655
20	1	529	93.99	97.95	11516
21	1	532	91.99	95.99	11952
22	1	533	92.99	97.99	13547
23	1	542	93.99	92.95	9168
24	1	544	90.95	95.95	11942
25	1	547	94.99	93.95	9917
26	1	554	89.95	90.95	10666
27	1	556	96.95	95.95	9717
28	1	560	91.99	97.95	13457
29	1	561	98.99	97.95	10319
30	1	566	93.95	91.99	9731
31	1	566	94.99	94.99	10279
32	1	582	98.99	91.99	7202
33	1	609	89.95	92.99	12103
34	1	612	92.95	92.99	11482
35	1	617	92.95	94.95	11944
36	1	623	94.99	91.99	9188

2018, Dale M. Nesbitt, All Rights Reserved

What If You Didn't Have The Data

- Or the data was "troubled."
- Wouldn't you want to start with direct assessments of the model coefficients.

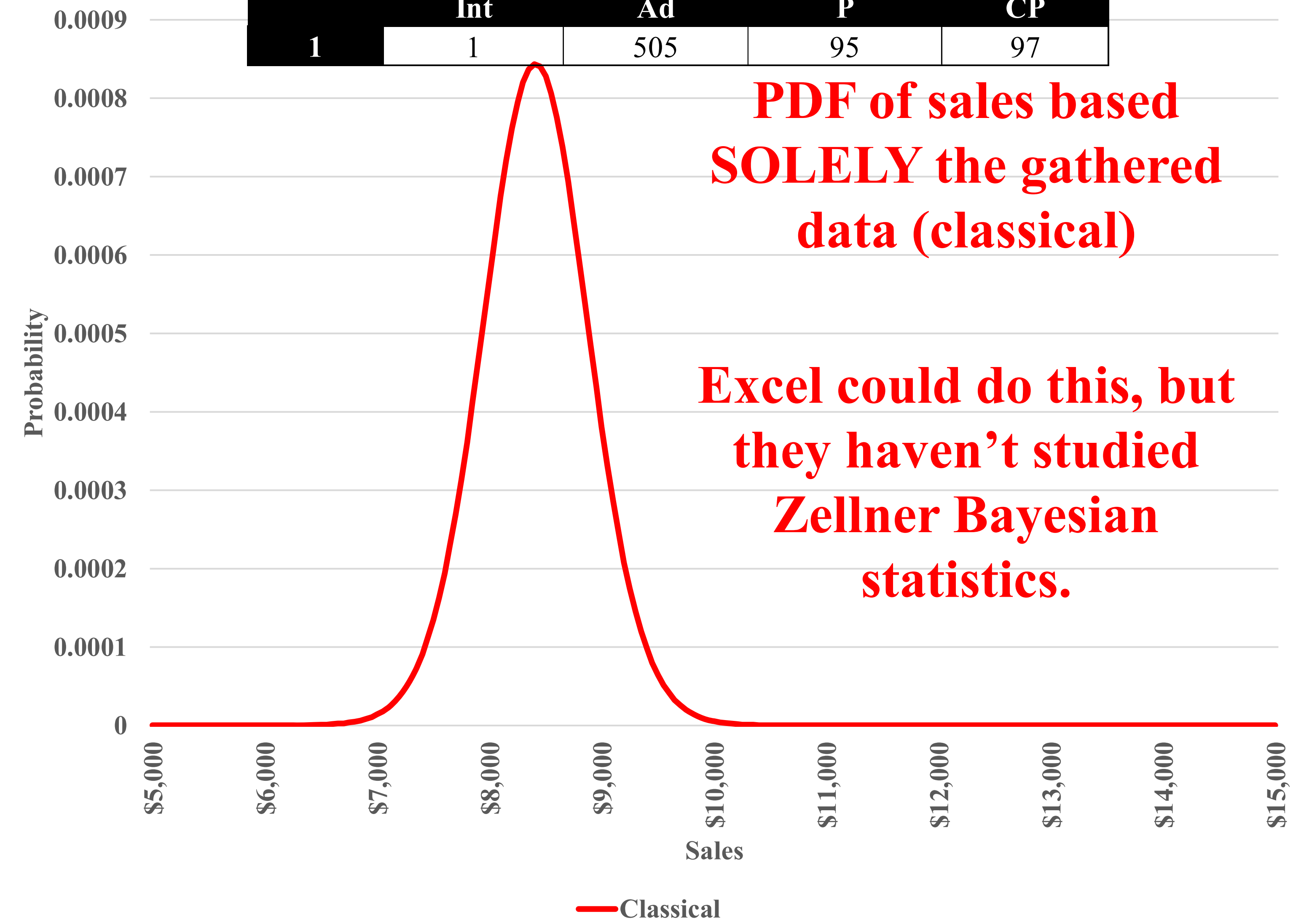
2018, Dale M. Nesbitt, All Rights Reserved

Here Is What Your Predictive Distribution Looks Like Based Solely on the Data

Decision Analysis

Dale M. Nesbitt

	Int	Ad	P	CP
1	1	505	95	97



PDF of sales based SOLELY the gathered data (classical)

Excel could do this, but they haven’t studied Zellner Bayesian statistics.

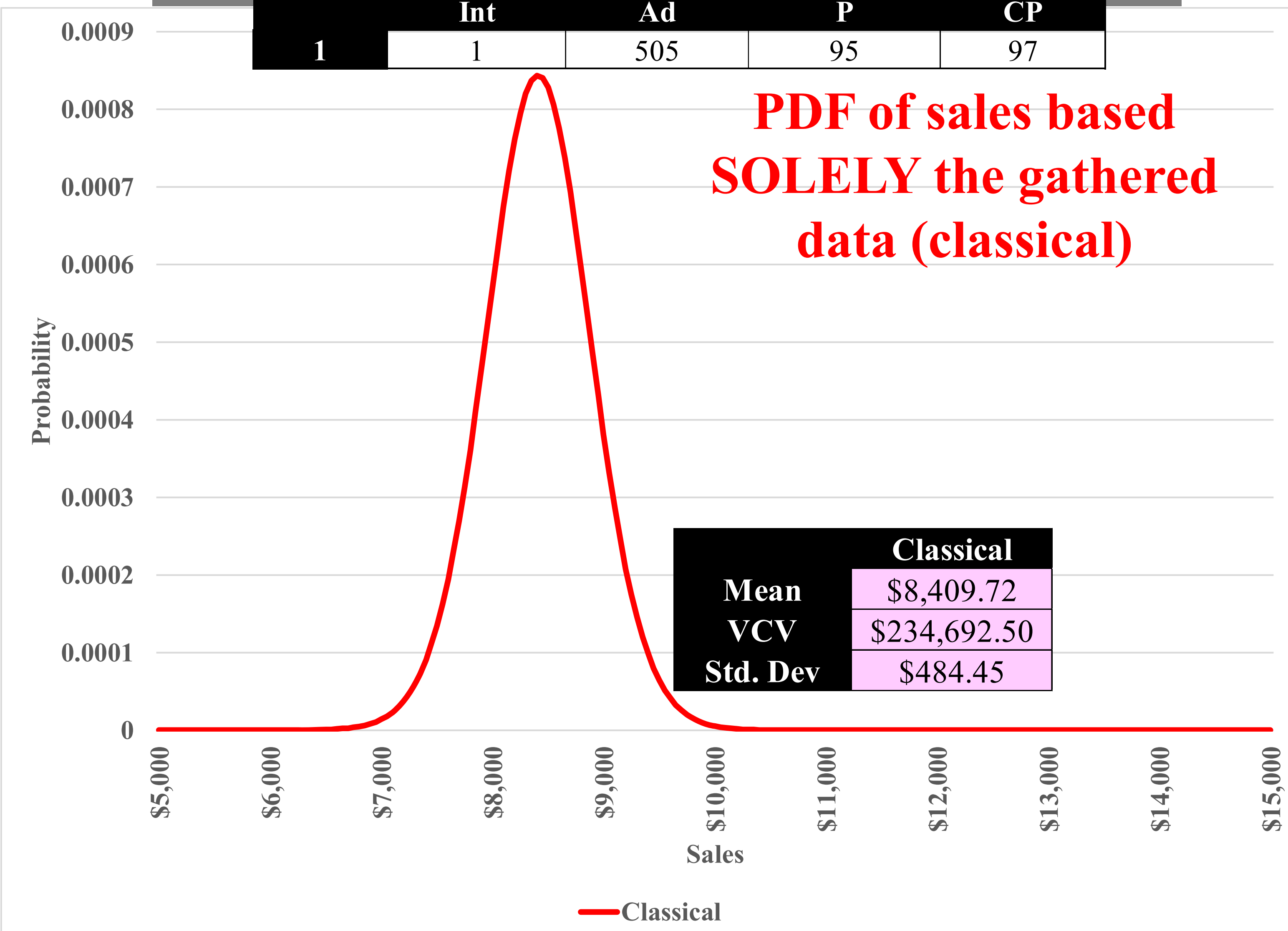
2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Predictive Distribution Based Solely on the Data

	Int	Ad	P	CP
1	1	505	95	97



	Classical
Mean	$8,409.72
VCV	$234,692.50
Std. Dev	$484.45

2018, Dale M. Nesbitt, All Rights Reserved

Predictive Distribution Based Solely on the Data

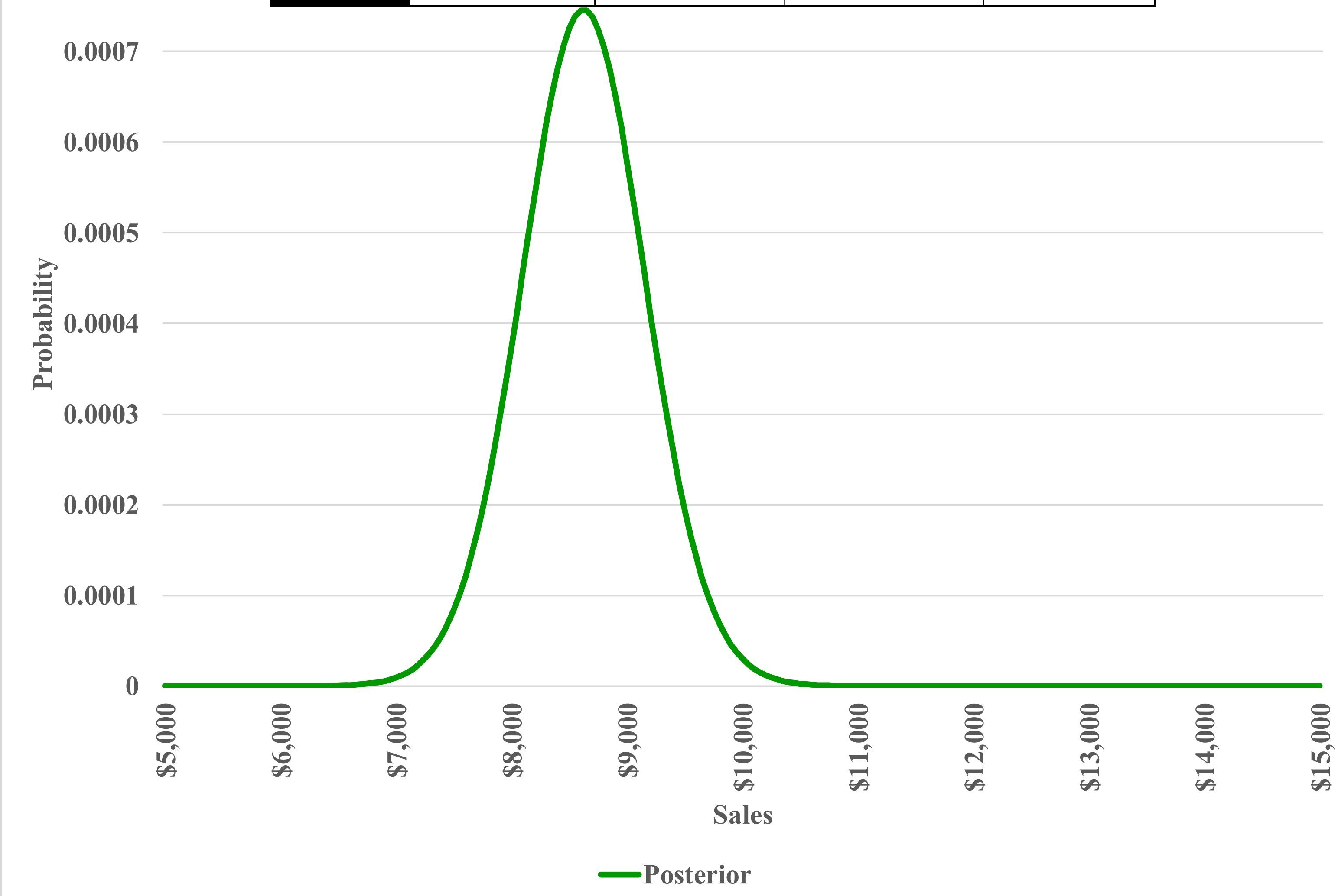
	Int	Ad	P	CP
1	1	505	95	97

	Classical
Mean	$8,409.72
VCV	$234,692.50
Std. Dev	$484.45

2018, Dale M. Nesbitt, All Rights Reserved

PDF Over Sales Combining Prior and Data Using Bayes

	Int	Ad	P	CP
1	1	505	95	97



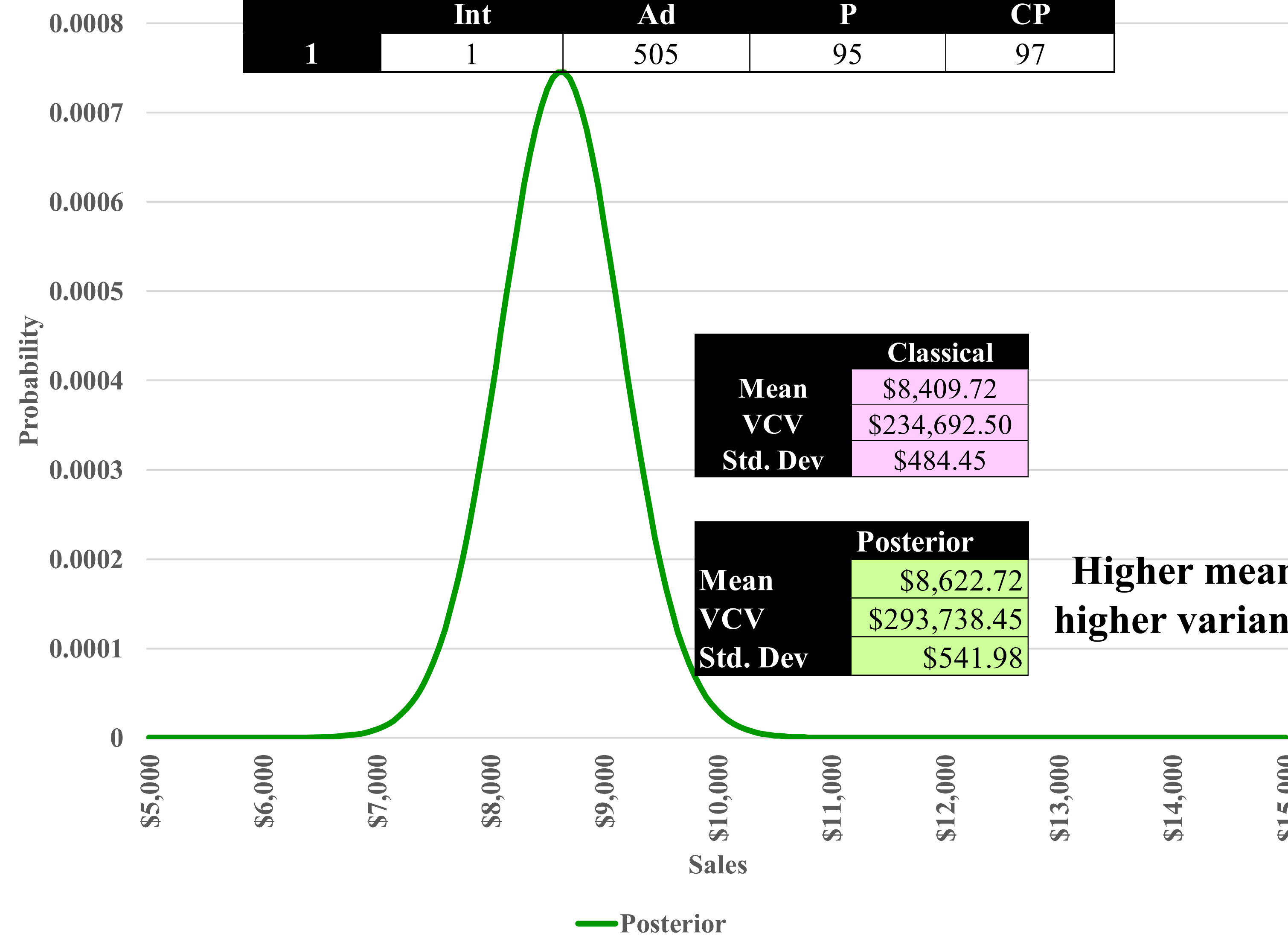
2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

PDF Over Sales Combining Prior and Data

	Int	Ad	P	CP
1	1	505	95	97



	Classical
Mean	$8,409.72
VCV	$234,692.50
Std. Dev	$484.45

	Posterior
Mean	$8,622.72
VCV	$293,738.45
Std. Dev	$541.98

Higher mean, higher variance

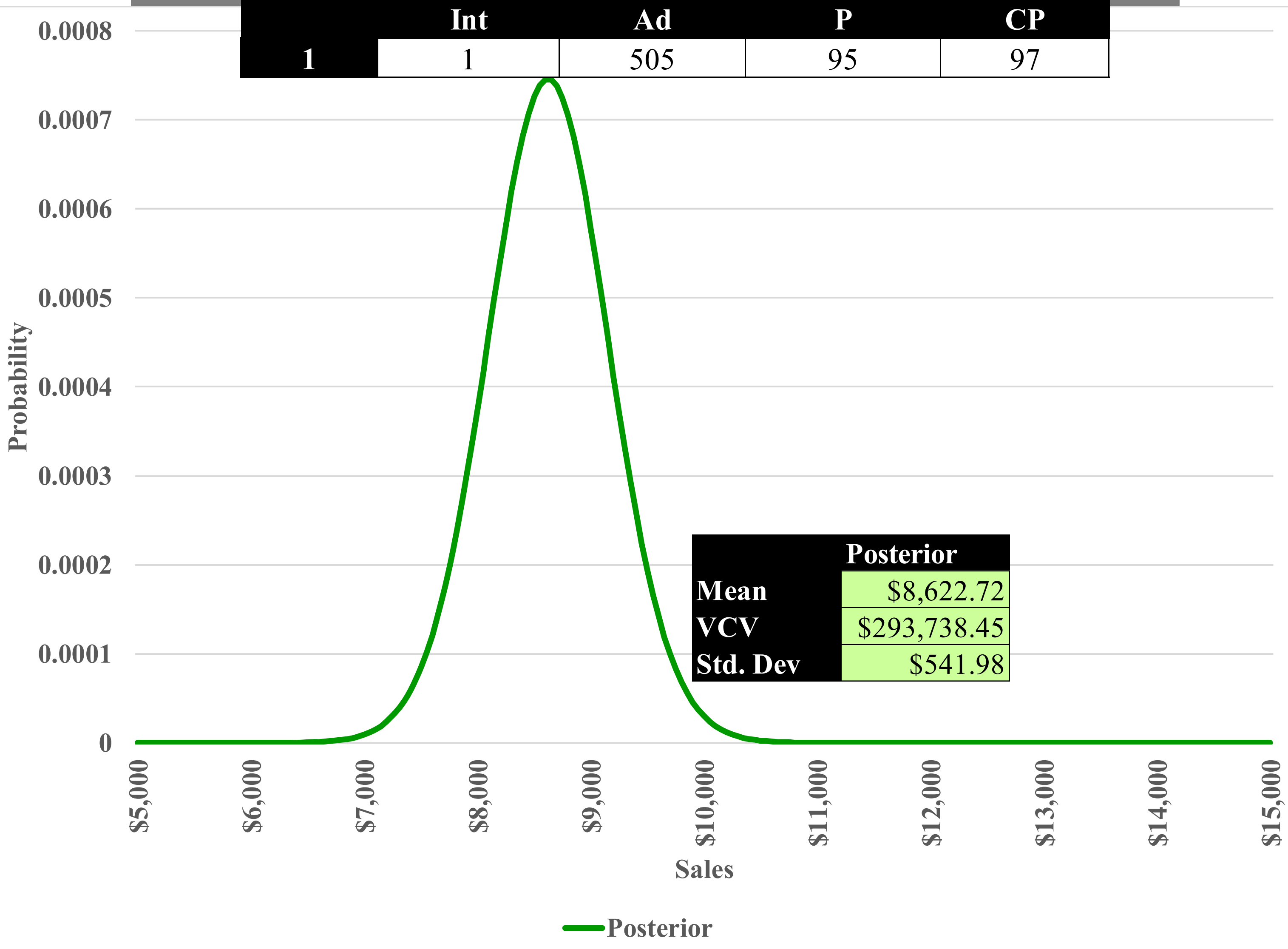
2018, Dale M. Nesbitt, All Rights Reserved

PDF Over Sales Combining Prior and Data

Decision Analysis

Dale M. Nesbitt

	Int	Ad	P	CP
1	1	505	95	97



	Posterior
Mean	$8,622.72
VCV	$293,738.45
Std. Dev	$541.98

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Posterior that Combines Prior and Data

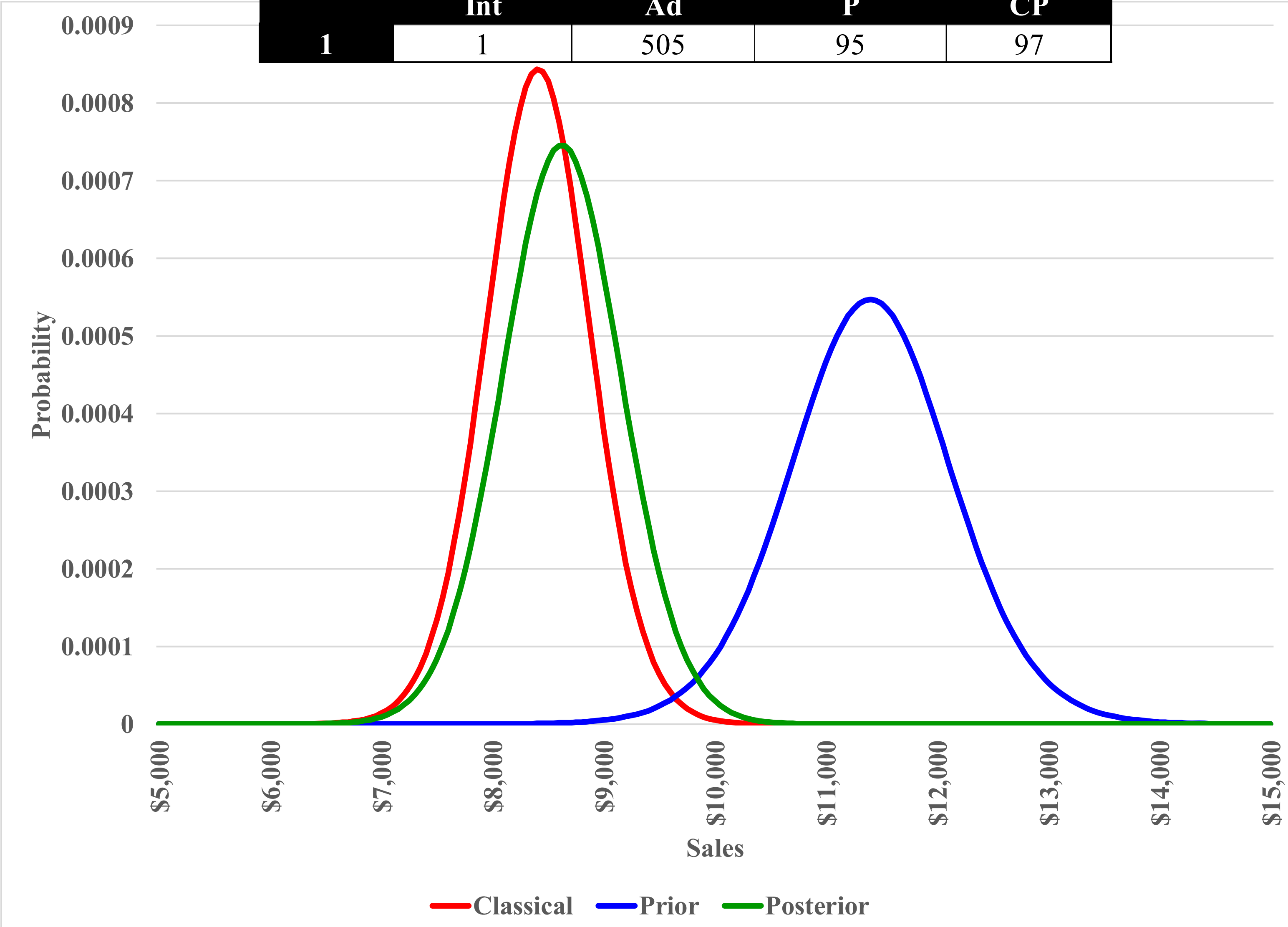
	Int	Ad	P	CP
1	1	505	95	97

	Posterior
Mean	$8,622.72
VCV	$293,738.45
Std. Dev	$541.98

2018, Dale M. Nesbitt, All Rights Reserved

What Do They Look Like on the Same Axis?

	Int	Ad	P	CP
1	1	505	95	97



Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

This Is Profound—Posterior is "Mix" of Prior and Likelihood

- Below is the mean (and mode) of the posterior
- It is a very special "matrix weighted average" of the prior and likelihood.
- This is so, so, so intuitive when you think of prior times likelihood and think of these terms in the exponent.
- It allows an arbitrary number of variables in your linear model.

$$Q^{-1}Q\beta_0 = \beta_0$$

$$\beta^* = \left(X^T X + Q\right)^{-1}\left(X^T y + Q\beta_0\right)$$

$$\left(X^T X\right)^{-1} X^T y = \overline{\beta}$$

2018, Dale M. Nesbitt, All Rights Reserved

Statistics Gives You a Continuous Curve CONDITIONAL on the Inputs

Decision Analysis

Dale M. Nesbitt

- You are not going to be using influence diagram software unless you discretize the inputs as well as the outputs given the inputs.
- It is a big job, but well worth it to get a really sophisticated, mutually relevant answer
- The pdfs are “influenced” in the sense of Howard and Abbas; probabilities depend on decisions.

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Which One Would You Want to Use?

Obviously the Bayesian posterior

2018, Dale M. Nesbitt, All Rights Reserved

Nesbitt, There Is No %^$&%*$ Way I am Programming Statistics!

- I am using fricking Excel regression if I do this.
- How can I garner the requisite information out of Excel?
- You cant; we have the software to do it.
- This software is really important, and we will give it to you.

Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

Excel Ignores Small Sample Size Adjustment

Decision

ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regression	3	137289637.6	45763212.55	217.1229936	2.41E-21			
Residual	32	6744669.357	210770.9174		6744669	R		
Total	35	144034307			32	ν_C	210770.9174	s_C^2

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*	*Head Calculated*	
Intercept	2199.342251	3839.735609	0.572784815	0.570793512	-5621.94	10020.62774	-5621.943242	10020.62774	3965.662	1.032796
Ad	15.04660288	1.172569433	12.83216367	3.67174E-14	12.65816	17.43504865	12.6581571	17.43504865	1.211025	1.032796
P	-503.7640378	28.3435642	-17.7734894	3.84442E-18	-561.498	-446.0300868	-561.4979888	-446.0300868	29.27311	1.032796
CP	499.6712512	30.55929246	16.35087762	4.29051E-17	437.424	561.9184929	437.4240094	561.9184929	31.5615	1.032796

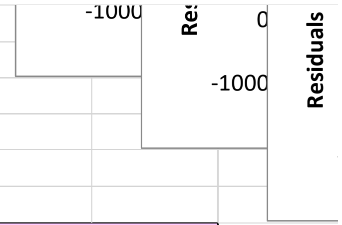
$\frac{\nu_C}{\nu_C - 2}$ 1.066667

$\sqrt{\frac{\nu_C}{\nu_C - 2}}$ 1.032796

Our SE is higher by

$\sqrt{\frac{\nu_C}{\nu_C - 2}}$

Multiply Excel by this factor



Dale N

- They need to calculate variance/covariance matrix and predictive density. They don't.

2018, Dale M. Nesbitt, All Rights Reserved

Classical Statistics Is the Bayesian Formulation but with a "Diffuse Prior"

- The model coefficients β are uniformly distributed between –a and a, with a going to infinity.
- The logarithm of the uncertainty coefficient σ is uniformly distributed between ln(1/a) and ln(a) with a going to infinity.
- This is **abject, utter, complete, blockheaded prior ignorance**.
  - You might as well get the prior from a St. Bernard or a banana slug.
  - Even politicians have a better prior than this!
  - Nesbitt's Maxim No. 2: I NEVER WANT TO BE THAT DUMB, (AND I DON'T BELIEVE ANYONE ACTUALLY IS).

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

Let’s Have a Plebiscite

- Who LIKES the diffuse prior?
- Who thinks anyone is really **that** dumb or **that** agnostic?
- Is anyone in the class **that** dumb? (Let the TA’s know.)
- Who thinks that represents anything close to reality?
- Who thinks that represents anything close to objectivity or transparency?

2018, Dale M. Nesbitt, All Rights Reserved

How Many Times Have You Heard Some Regression Person Say….

- Oh, that cant be right. The price elasticity should be negative. (Duh…)
- Oh, that cant be right. A should be more important and have a bigger coefficient than B.
- Oh, that cant be right. The R^2 is too small.
- Oh, that cant be right. A and B cant be that correlated (i.e., have that high a covariance).
- This ain't abject ignorance; this is either bias or problem knowledge! They need to be in the prior.
- Oh, oh, multicollinearity.
- We need more data; there isn't enough variation.

Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

Our Classical Solution Is Different from the Excel Solution (Say What?)

ANOVA

	df	SS	MS	F	ignificance F			
Regression	3	137289637.6	45763212.55	217.1229936	2.41E-21			
Residual	32	6744669.357	210770.9174		6744669	R		
Total	35	144034307			32	ν_C	210770.9174	s_C^2

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%	Head Calculated	
Intercept	2199.342251	3839.735609	0.572784815	0.570793512	-5621.94	10020.62774	-5621.943242	10020.62774	3965.662	1.032796
Ad	15.04660288	1.172569433	12.83216367	3.67174E-14	12.65816	17.43504865	12.6581571	17.43504865	1.211025	1.032796
P	-503.7640378	28.3435642	-17.7734894	3.84442E-18	-561.498	-446.0300868	-561.4979888	-446.0300868	29.27311	1.032796
CP	499.6712512	30.55929246	16.35087762	4.29051E-17	437.424	561.9184929	437.4240094	561.9184929	31.5615	1.032796

$$s_C^2\left(X^TX\right)^{-1}$$

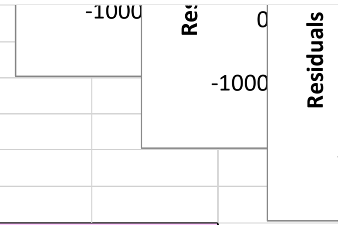
$$s_C^2\left(X^TX\right)^{-1}\frac{\nu}{\nu-2}$$

$\frac{\nu_C}{\nu_C-2}$ 1.066667

$\sqrt{\frac{\nu_C}{\nu_C-2}}$ 1.032796

Our SE is higher by $\sqrt{\frac{\nu_C}{\nu_C-2}}$

Multiply Excel by this factor



They use $s_C^2\left(X^TX\right)^{-1}$ for the variance covariance matrix within Excel. We use the right answer $s_C^2\left(X^TX\right)^{-1}\frac{\nu}{\nu-2}$

Excel assumes normal rather than Student's t, which is technically incorrect

2018, Dale M. Nesbitt, All Rights Reserved

Decision Analysis

Dale M. Nesbitt

It's a Good Thing You Only Pay About $100/yr for Excel!

2018, Dale M. Nesbitt, All Rights Reserved

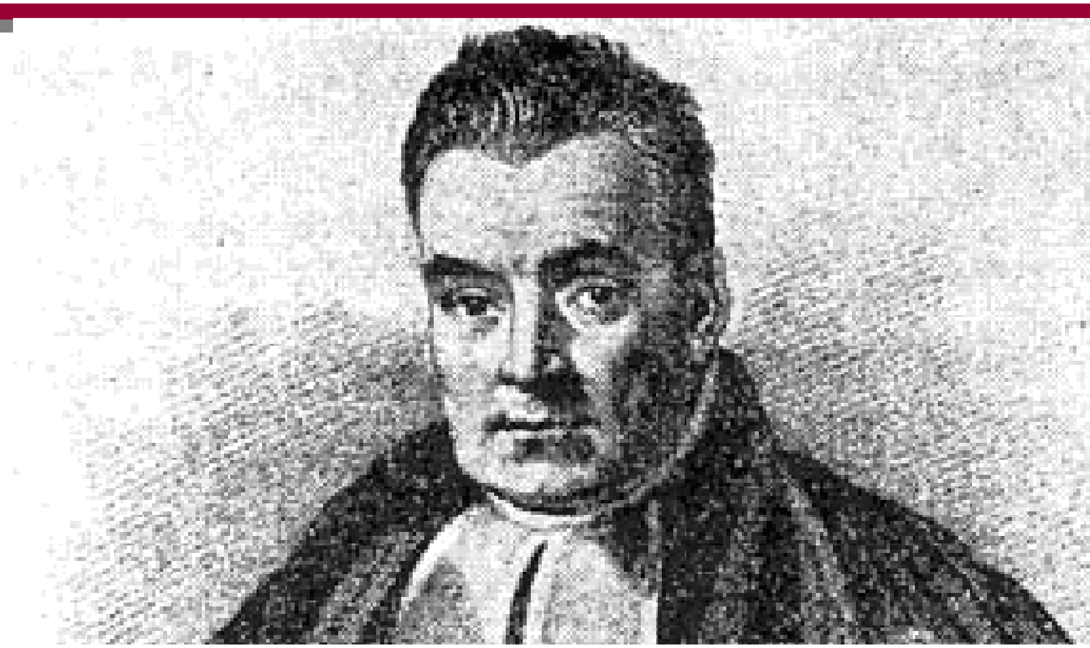
Decision Analysis

Dale M. Nesbitt

History

2018, Dale M. Nesbitt, All Rights Reserved

The Reverend Thomas Bayes



- English statistician, philosopher and Presbyterian minister, known for having formulated a specific case of the theorem that bears his name: Bayes' theorem.
- Bayes never published what would eventually become his most famous accomplishment; his notes were edited and published after his death by Richard Price.

Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved

People Have Pilgrimages to Bayes' Grave

- Bayes' solution to a problem of inverse probability was presented in "An Essay towards solving a Problem in the Doctrine of Chances" which was read to the Royal Society in 1763 after Bayes' death.
- He is interred in Bunhill Fields Cemetery in London where many Nonconformists are buried.
  - "Nonconformist" or "Non-conformist" was a term used in England and Wales after the Act of Uniformity 1662 to refer to a Protestant Christian who did not "conform" to the governance and usages of the established Church of England. English Dissenters (such as Puritans) who violated the Act of Uniformity 1559 may retrospectively be considered Nonconformists, typically by practicing or advocating radical, sometimes separatist, dissent with respect to the established state church.

2018, Dale M. Nesbitt, All Rights Reserved

Bob Stibolt (former EES) Told Me About a Pilgrimage to Bayes' Tomb

- Evidently several people went to Bayes Tomb to pay homage.
- Apparently a lot of people visit it.
- It is pretty convenient to get to.
- It is in near north central London.
- It is definitely on my bucket list.

2018, Dale M. Nesbitt, All Rights Reserved

Modern Bayes Hero—the Late Arnold Zellner

- Arnold Zellner (January 2, 1927 – August 11, 2010) was an American economist and statistician specializing in the fields of Bayesian probability and econometrics.
- Zellner contributed pioneering work in the field of Bayesian analysis and econometric modeling.
- Why did Zellner, who had already launched a successful research program within the classical approach, become such a stubborn advocate of the Bayesian approach?
- He undertook a research program to evaluate the two approaches, both theoretically and in applied econometric studies.

A Nesbitt Hero

2018, Dale M. Nesbitt, All Rights Reserved

I Worked with Zellner in the 1990s

- He connected the dots from regression to Bayesian probability.
- He had absolutely no reason and no personal gain from helping me, but he did.
- I adored the guy.
- He was an absolutely delightful guy, very, very helpful and intellectual.

Decision Analysis

Dale M. Nesbitt

2018, Dale M. Nesbitt, All Rights Reserved