Jim L. Sanders, MBA
real@cox.net
520-322-0088
http://realestateappraisalandlitigation.com

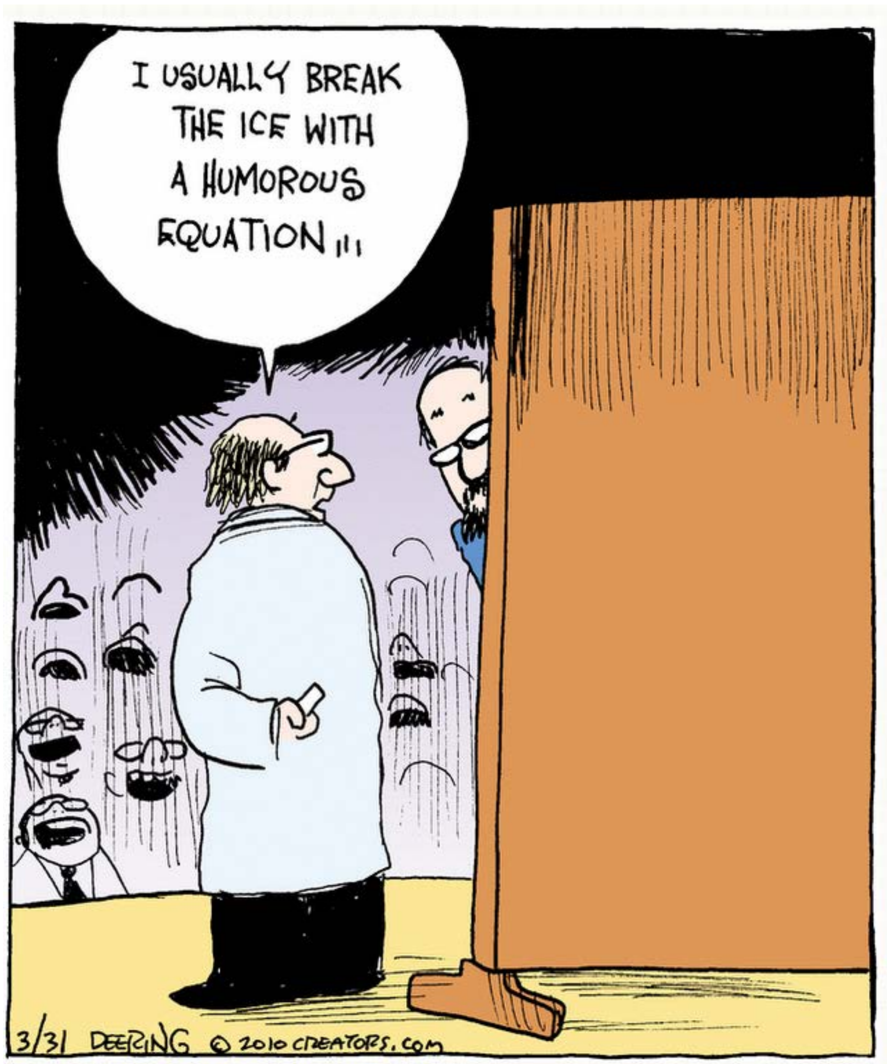
Condemnation Appraisals for Public Utilities:

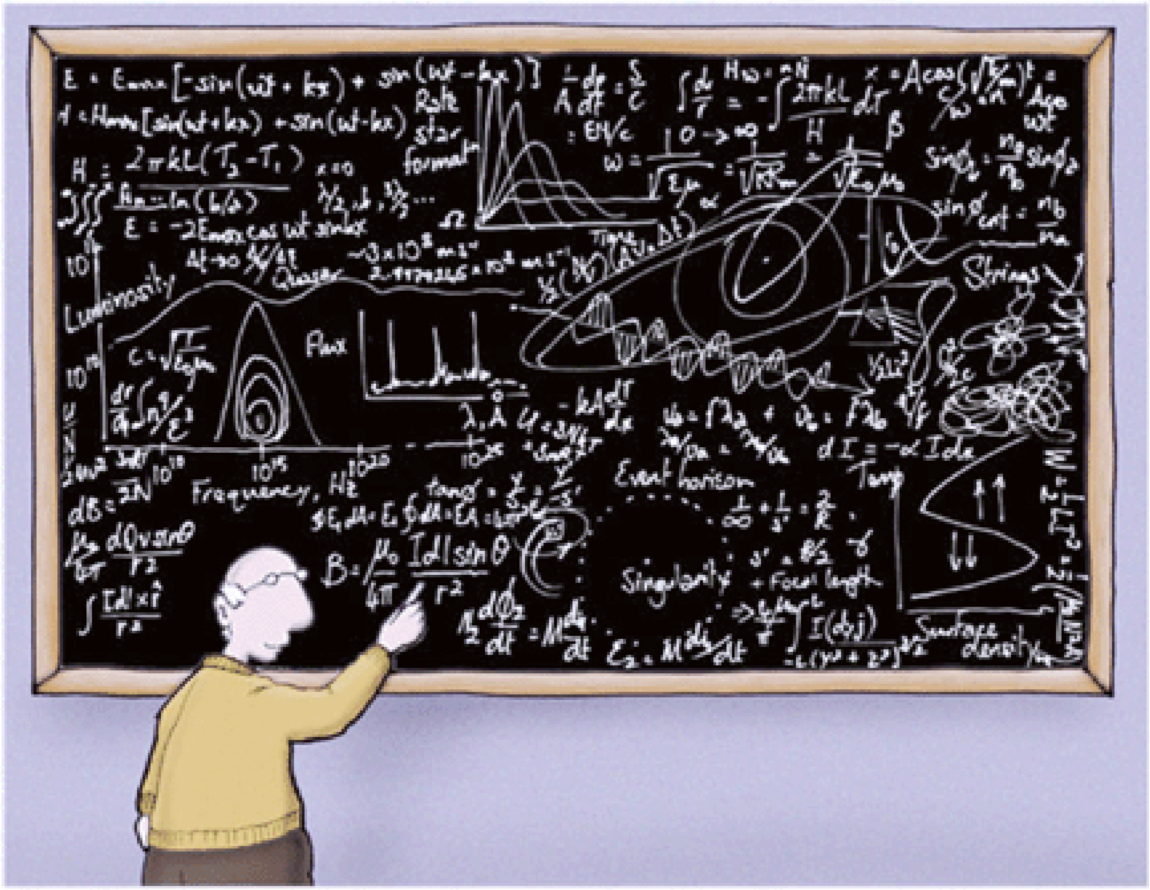
A Review and Critique of Regression Analysis

May 25, 2016

Points of Emphasis

- Regression Analysis is wonderful **TOOL** for the appraiser **if, and only if, used correctly.**
- From my observations, Regression Analysis is NOT being used correctly but **GIGO** (garbage in, garbage out).
- **Using Regression Analysis DOES NOT make science.** Science is about independent parties being able to replicate results, then examining the model, data, results, etc.
- Thus, independent parties **MUST have access to the dataset** used.
- There is an **ongoing problem** in science with studies that cannot be replicated, even with the data.
- For powerline cases, a **"Decision Science"** way of looking at data probably works better than "Science" analysis. This implies looking at the "most probable" standard rather than a standard of "beyond a reasonable doubt."
- This can be accomplished using **Bayesian Statistics** rather than Frequentist Statistics.
- How is all of this relevant to **Daubert**?


I USUALLY BREAK THE ICE WITH A HUMOROUS EQUATION...
3/31 DEERING © 2010 CREATORS.COM



STATISTICS MADE EASY

“Explanations exist;
they have existed for all time;
there is always a well-known solution to every human problem —
neat, plausible, and wrong.”
- H. L. Mencken

I. Does regression analysis add scientific validity or reliability to an appraisal?

No access to data?

1. No reliability
2. No validity

Note: After I examined data for Wisconsin land study discussed later, I have tried to get underlying data for many studies including FOIA requests and have been denied.

No data, No cigar

- **Transparency and Reproducibility are key ingredients of good science, and Require that data and methods, including computer code, be made available.** http://www.nature.com/ngeo/focus/transparency-in-science/index.html

- Reproducibility is regarded as one of the foundations of the entire scientific method, a benchmark upon which the reliability of an experiment can be tested.

- The basic principle is that, for any research program, an independent researcher should be able to replicate the experiment, under the same conditions, and achieve the same results. https://explorable.com/reproducibility

Data-Access Practices Strengthened

In our continued drive for reproducibility, *Nature* and the Nature research journals are strengthening our editorial links with the journal *Scientific Data* and enhancing our data-availability practices. We believe that this initiative will improve support for authors looking for appropriate public repositories for their research data, and will increase the availability of information needed for the reuse and validation of those data (November 19, 2014). http://www.nature.com/news/data-access-practices-strengthened-1.16370

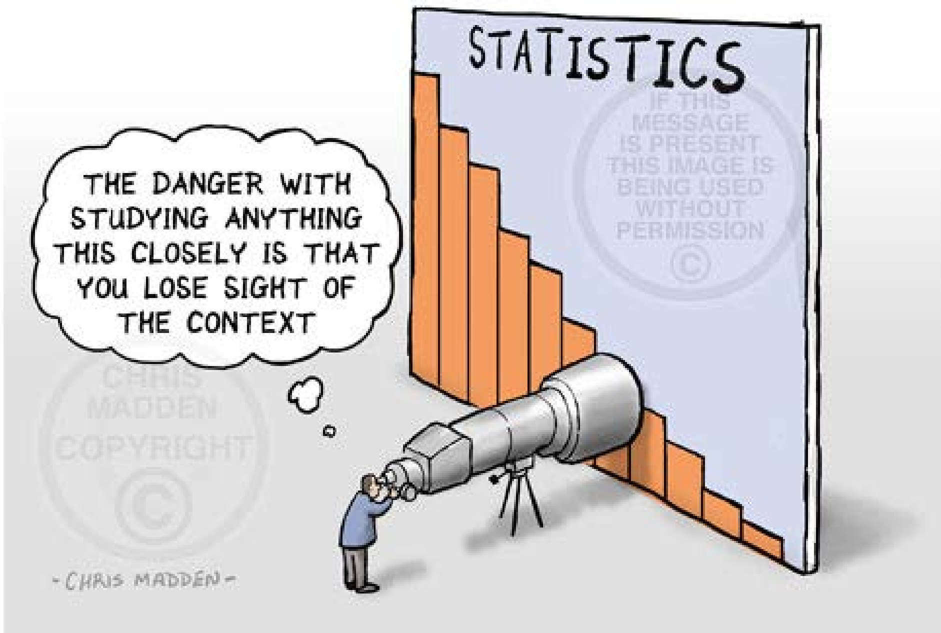
A. Using statistics does not make study scientific (Implications for Daubert?)

B. To reproduce studies using observational data, one must have the data

C. Most peer reviewers in non-sciences do not get access to data (this is true for appraisal and real estate journals). So, is peer reviewed meaningful under Daubert without the data for powerline cases?

II. What can be done if one gets the data that was analyzed using regression analysis?

Danger Issue:



Sometimes the proper instrument is a telescope and other times a microscope.

A. GIGO (Garbage In, Garbage Out)

GIGO (Garbage In, Garbage Out) is a concept common to computer science and mathematics: the quality of output is determined by the quality of the input. So, for example, if a mathematical equation is improperly stated, the answer is unlikely to be correct. Similarly, if incorrect data is input to a program, the output is unlikely to be informative.
http://searchsoftwarequality.techtarget.com/definition/garbage-in-garbage-out

B. This suggests two issues:

1. A regression analysis model is a simple algebraic equation. If the model is incorrectly specified, GIGO.
2. In almost all cases for real estate data, it is important to have data with the same highest & best use. One does not compare a 0.8 acre lot on a bluff overlooking a river that sells for over $200,000 per acre with a 655 acre farm that sells for approximately $5,000 per acre.

Note: For reasonable results, one should have similar scale of data on all dimensions (acres, price per acre, building area, etc.) For example, one does not compare a 1,200 square foot shoe store with a 15,000 square foot big box even though both are "retail". Also, one does not compare either to a 4,500 square foot fast feeder. Why? Retail is not a highest & best use. Type of retail may be.

- The statistics program will do calculations even if these concepts are violated enhancing the "probability" of GIGO.
- Summary statistics of the data should always be provided. This means summaries of the raw numbers (number of acres, price per acre, etc.) and not summaries of the statistics (i.e. range of 1-2.8 standard deviations from the mean of acres. What does that tell you?).

"Remember, statistics are in the eye of the manipulator."

III. If you have the study but not the data

- Must have a basis for understanding a regression analysis far beyond what can be discussed in my allotted time.
- The study should not be relied upon because it cannot be replicated.
- The reported results can be examined for the following:

1. **Are summary statistics of every variable provided?**
2. **Does the data appear to contain competitive properties?**

   Example: I just attended Condemnation summit in Arizona where engineer/appraiser discussed results of large study in Kansas. The dataset for retail properties contained properties ranging in size from approximately 1,200 square feet to over 15,000 square feet and had many different types without variable identifying these different types (e.g. fast food, big box, stand-alone shoe store).

3. **Are the model specifications discussed** (did modeler use data to predict sales price, natural log of sales price, etc., why where variables chosen)?
4. **Does the author discuss the analyses to suggest the assumptions of regression analysis were not violated (discussed later)?**
5. **Do the results make sense?**

   For example, if I have farms ranging from 20 to 600 acres, does it make sense to estimate an absolute value loss (if any) or should one expect a loss that is a function of price or size?

IV. If you have the study and the data

A. Re-run the model with the data. (This is better if you have statistics program, but can marginally be done using Excel).

B. Compute summary statistics.

Example:

Descriptive Statistics: SalePr, PPA, WetAcres, Wood_Acres, Open_Acres, Total Acres

Variable	Total Count	Mean	TrMean	StDev	Minimum	Q1	Median	Q3	Maximum
SalePr	385	105639	89401	123137	6000	42500	79900	129200	1289500
PPA	385	3621	2551	10467	325	1197	1916	3292	192771
WetAcres	385	3.753	2.194	9.277	0.000	0.000	0.000	2.800	86.226
Wood_Acres	385	27.49	22.24	46.08	0.00	1.73	18.00	37.00	552.81
Open_Acres	385	22.79	18.39	33.27	0.00	1.60	10.34	34.04	274.40
Total Acres	385	54.03	47.56	59.65	0.83	20.07	40.14	77.98	655.32

Thus, one can see the acreage varies from 0.83 acres to 655.32 acres; sales price varies from $6,000 to $1,289,500. This is a major problem.

C. Get the model output and confirm results.

Example:

Regression Analysis: SalePr versus S_2002, S_2003, ...

```
SalePr = - 38478 + 41400 S_2002 + 34888 S_2003 + 40929 S_2004 + 57818 S_2005 + 73203
S_2006 + 93323 S_2007 + 145305 S_2008 + 1131893 TRANSITION_4 + 22073 AGRICULTURE_4
+ 1963 Wood_Acres + 1394 Open_Acres - 1261 WetAcres + 74153 Govt_purchase + 40886
Brown - 37589 Clark + 112184 Dane + 35917 Rock - 2588 Online
```

Predictor	Coef	SE Coef	T	P	VIF
Constant	-38478	9115	-4.22	0.000	
S_2002	41400	12863	3.22	0.001	1.841
S_2003	34888	11148	3.13	0.002	1.999
S_2004	40929	9795	4.18	0.000	2.842
S_2005	57818	9794	5.90	0.000	3.105
S_2006	73203	9717	7.53	0.000	2.965
S_2007	93323	15467	6.03	0.000	1.884
S_2008	145305	30925	4.70	0.000	1.307
TRANSITION_4	1131893	48477	23.35	0.000	1.076
AGRICULTURE_4	22073	6625	3.33	0.001	1.664
Wood_Acres	1962.84	72.98	26.90	0.000	1.994
Open_Acres	1394.49	88.42	15.77	0.000	1.525
WetAcres	-1261.3	339.2	-3.72	0.000	1.745
Govt_purchase	74153	18279	4.06	0.000	1.054
Brown	40886	11462	3.57	0.000	1.304
Clark	-37589	7332	-5.13	0.000	1.198
Dane	112184	16632	6.75	0.000	1.357

```
Rock              35917     15740    2.28  0.023  1.215
Online            -2588      5808   -0.45  0.656  1.051


S = 46677.9   R-Sq = 86.3%   R-Sq(adj) = 85.6%
```

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	18	5.02504E+12	2.79169E+11	128.13	0.000
Residual Error	366	7.97451E+11	2178827599		
Lack of Fit	363	7.96709E+11	2194791381	8.88	0.047
Pure Error	3	741630000	247210000		
Total	384	5.82249E+12			

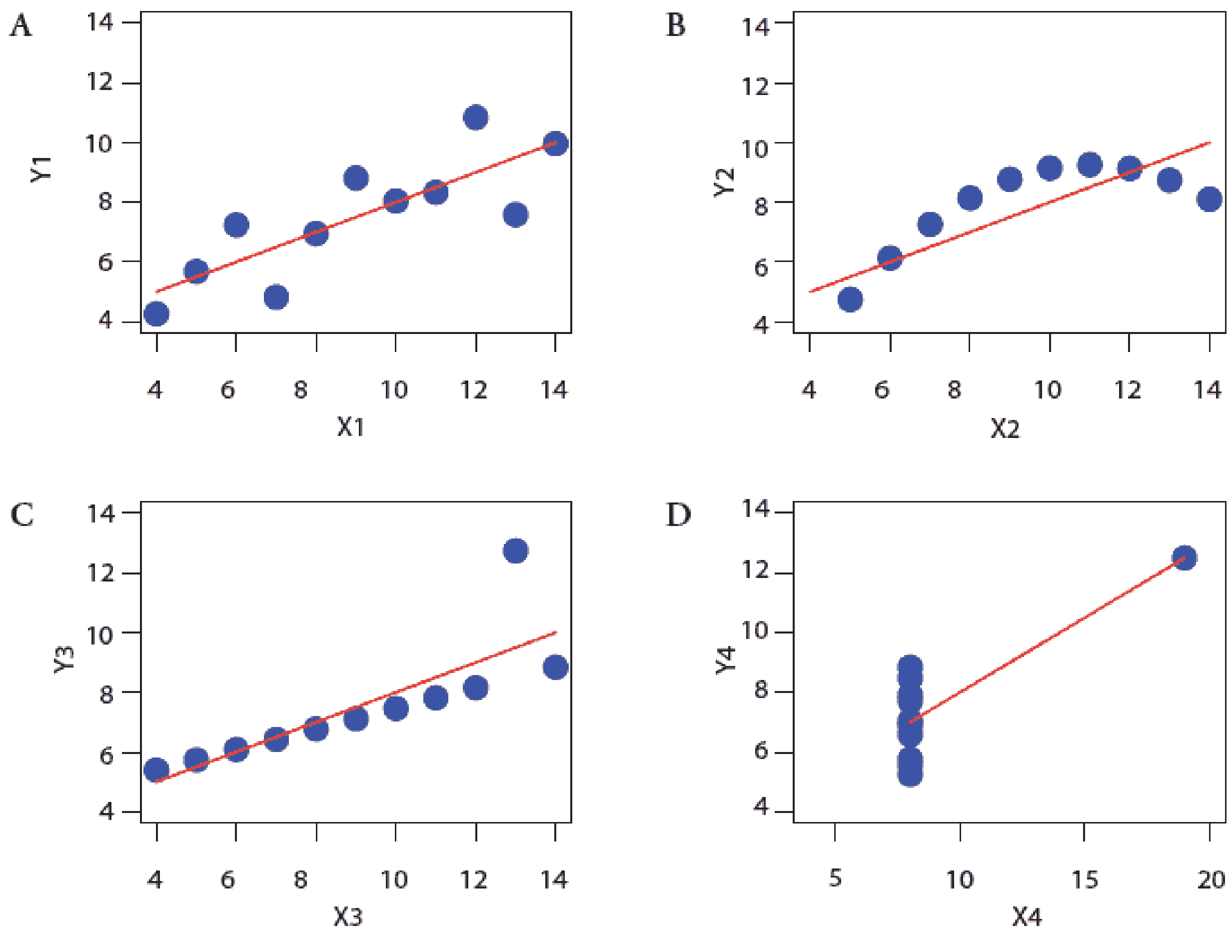
D. Make sure you get the residuals for completing residual analyses.
Note: *If you do not already know what residuals are, you are not ready to go on.*

E. Test the residuals to make sure that a serious violation of the assumptions of regression analysis does not exist that would invalidate the study.

- **Be familiar with Anscombe's Quartet – Graph the data.**

e.g. From https://www.aasv.org/shap/issues/v15n5/v15n5editor.htm

Figure 1: Scatterplots of data from four different sources and the least squares regression line illustrating the "best" linear relationship between the independent and dependent variables (data adapted from Anscombe, 1973).



Linear regression models assume:

- That the residuals are normally distributed.
- That each observation is independent of the others.
- That there is a **linear** relationship between the independent and dependent variables.
- That the variance of the dependent (outcome) variable does not change with the value of the independent variable.

More details about the assumptions of linear regression models may be found elsewhere. The major assumptions need to be evaluated, and fitting the best final model requires much more than simple one-step specification of a model and interpretation of summary statistics. **It is an iterative process** in which outputs at one stage are used to validate, diagnose, and modify inputs for the next stage. **Small violations of assumptions usually do not invalidate the conclusions. However, a large violation will substantially distort the association and lead to an erroneous conclusion.**

F. Are there serious violations?

Unusual Observations

Obs	S_2002	SalePr	Fit	SE Fit	Residual	St Resid
6	0.00	1289500	1289500	46678	-0	* X
35	0.00	102800	-1599	8619	104399	2.28R
72	0.00	340000	475050	23644	-135050	-3.36RX
82	0.00	400000	280879	16810	119121	2.74R
83	1.00	186200	241675	18079	-55475	-1.29 X
85	0.00	74000	117985	18311	-43985	-1.02 X
86	0.00	615000	496990	21392	118010	2.84RX
87	0.00	390000	367823	27906	22177	0.59 X
88	0.00	185000	210446	18048	-25446	-0.59 X
89	0.00	155000	195369	18349	-40369	-0.94 X
90	0.00	235000	220101	18026	14899	0.35 X
91	0.00	235000	272764	28046	-37764	-1.01 X
99	0.00	130000	238075	7119	-108075	-2.34R
107	1.00	120000	218552	11127	-98552	-2.17R
116	0.00	132000	38143	8977	93857	2.05R
131	0.00	358000	174701	6145	183299	3.96R
134	0.00	320000	219696	8365	100304	2.18R
150	0.00	936989	1021328	27514	-84339	-2.24RX
151	0.00	172000	59564	5750	112436	2.43R
153	0.00	1228500	1060596	29340	167904	4.62RX
154	0.00	200000	88428	7290	111572	2.42R
180	0.00	524900	215633	9505	309267	6.77R
202	0.00	290000	238490	18910	51510	1.21 X
219	0.00	322200	373704	21449	-51504	-1.24 X
224	0.00	75100	164924	19550	-89824	-2.12RX
225	0.00	50000	147633	10040	-97633	-2.14R
226	0.00	50000	144027	9260	-94027	-2.06R
228	0.00	27400	73088	19368	-45688	-1.08 X
229	0.00	54000	18272	21430	35728	0.86 X
241	0.00	381000	387731	19864	-6731	-0.16 X
242	0.00	250000	272265	19413	-22265	-0.52 X
259	0.00	330000	156205	6674	173795	3.76R
262	0.00	152000	236846	19151	-84846	-1.99 X
276	0.00	400000	202157	22144	197843	4.81RX
279	0.00	45000	171605	17502	-126605	-2.93R
283	0.00	229700	214113	30185	15587	0.44 X
333	0.00	162000	61874	13696	100126	2.24R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.

Durbin-Watson statistic = 1.85734

Lack of fit test

Possible interaction in variable S_2006 (P-Value = 0.006)

Possible interaction in variable Open_Acr (P-Value = 0.023)

Possible curvature in variable Govt_pur (P-Value = 0.032) ***Means non-linear**

Possible lack of fit at outer X-values (P-Value = 0.000)

Overall lack of fit test is significant at P = 0.000

The appraiser must be able to understand what the above means. Many different tests exist that are beyond this seminar.

G. What is the Decision Science way?

- Decision Science involves making the best decision under uncertainty. That is, "certainty" rarely exists when a decision has to be made. Thus, one must make the best decision possible with the information available.
- Decision Science involves looking at the "Loss Function" of a decision when making the decision. See below:

From good ole Wikipedia (I know an attorney in the audience is gnashing teeth):

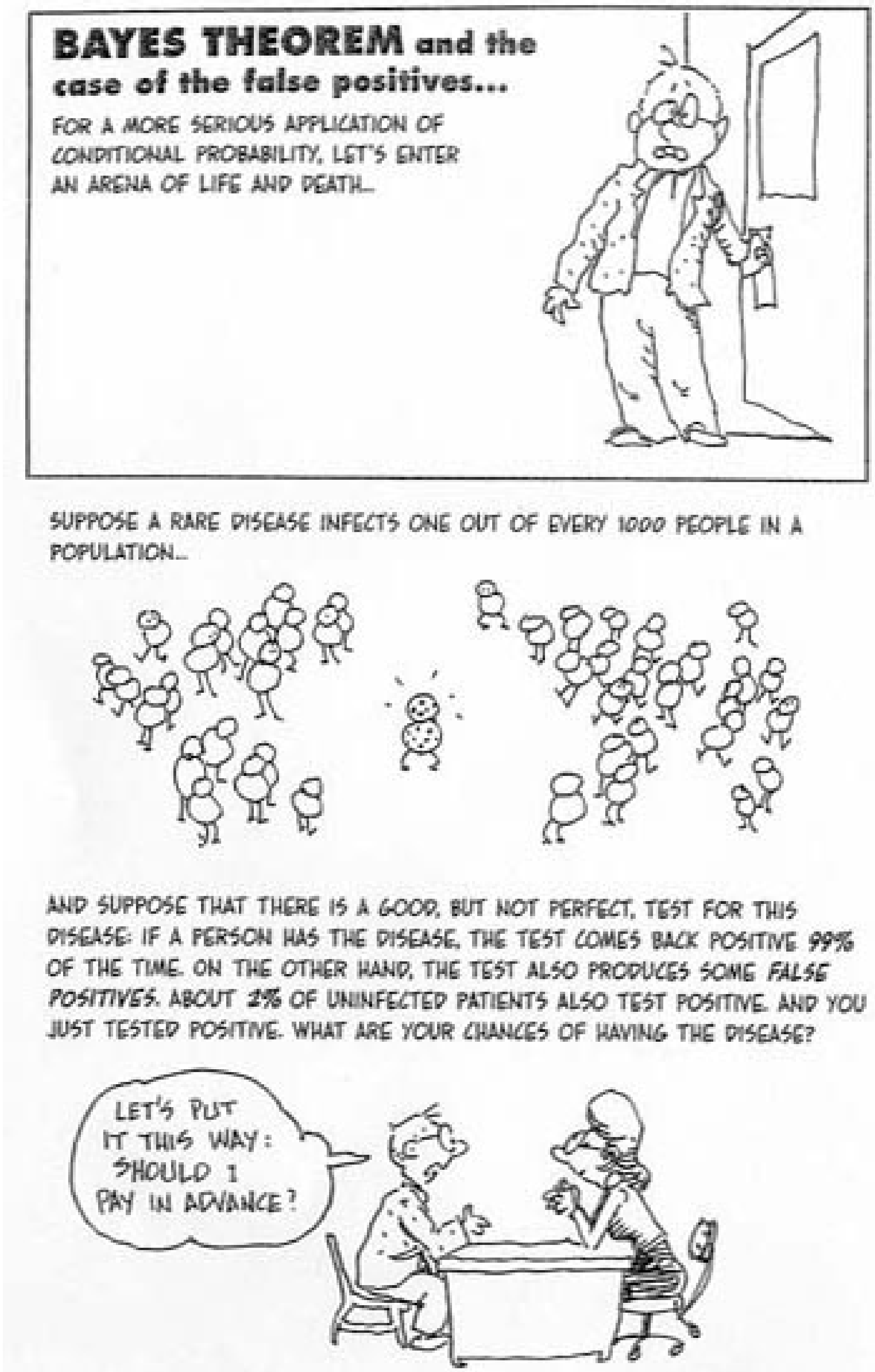
One issue in Decision Science is what is called the loss function. This is illustrative by the classic Pascal's Wager:

Pascal's Wager is a classic example of a choice under uncertainty.

Pascal's Wager is the argument that states that you should believe in God even if there is a strong chance that he might not be real, because the penalty for not believing, namely going to hell, is so undesirable that it is more prudent to take your chances with belief.

A way to deal with this issue involves what is known as Bayesian Statistics, the use of Bayes' Theorem.

Illustration of Loss Function



The above highlights a couple of issues that Bayes' Theorem can help with but is not exhaustive.

H. The property owner's appraiser must have a basis for her or his opinion that there are severance damages created by powerlines to use Bayes' Analyses.

1. Studies

2. Surveys

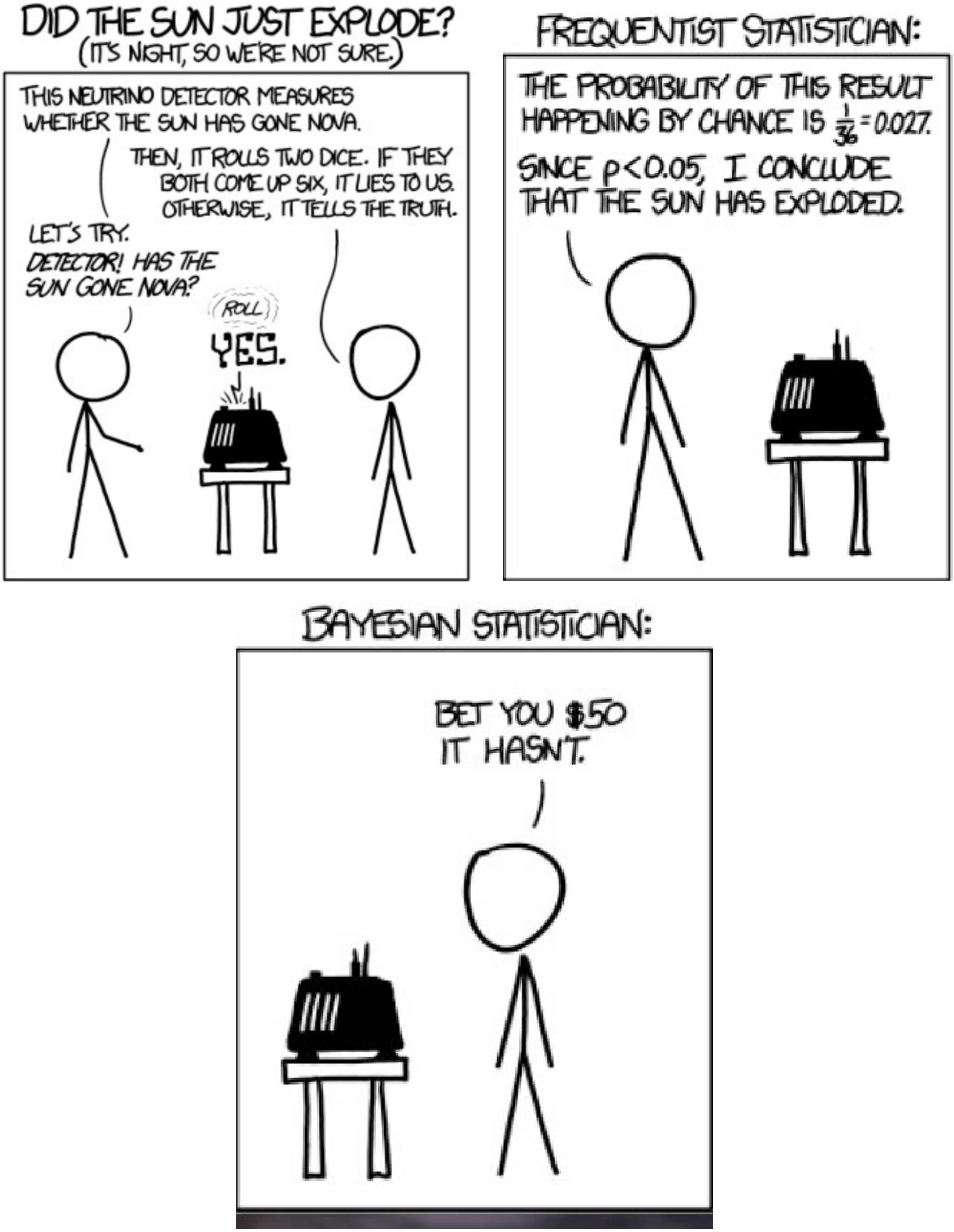
3. Common sense? I have never met anyone telling me they would want to live next to a high voltage powerline and I've asked a lot of people.

I. Initial plausibility (Important if Bayes' Analyses are to be used)

Many philosophers consider the *initial plausibility* of a claim to be a factor in determining the burden of proof. This makes little differences as to who has the burden of proof but does affect the standard required for the justification to be found convincing. An extraordinary claim would require an extremely good justification. This concept is often stated as "extraordinary claims require extraordinary evidence."

All the philosophical and legal underpinnings of this issue are beyond the scope of this presentation.

One last illustration of the difference between a Bayesian Statistician and the type of statistician most common called a Frequentist.


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

References and Further Reading

Regression Analysis

GIGO (Garbage In, Garbage Out)
http://searchsoftwarequality.techtarget.com/definition/garbage-in-garbage-out

Evaluating the Assumptions of Linear Regression Models
https://www.aasv.org/shap/issues/v15n5/v15n5editor.htm

Problems with Science

Reproducibility:

Reproducibility study challenged: http://nyti.ms/1XbuXiM

"After all, reproducibility is a cornerstone of the scientific process, and in essence it allows researchers to gain confidence in others' work. What's more, sharing research artifacts allows researchers to build on others' work to avoid needless replication of research and to advance science, a process known as benefaction."
https://uanews.arizona.edu/story/ua-computer-scientists-push-for-code-sharing

Industry Sponsorship and Research Outcome: http://www.ncbi.nlm.nih.gov/pubmed/23235689

https://explorable.com/reproducibility

http://www.nature.com/ngeo/focus/transparency-in-science/index.html

Strengthening Data Access Practices http://www.nature.com/news/data-access-practices-strengthened-1.16370

P-Values:

The problem with the p-value cuts both ways. Over-interpretation of the p-value can lead to both false positives and false negatives. Dependence on a specific p-value can lead to bias as researchers may discontinue or shelve work that doesn't meet this arbitrary standard.
http://phys.org/news/2016-03-p-value-pointless.html#jCp

There has been something of a crisis in science. It has become apparent that an alarming number of published results cannot be reproduced by other people. That is what caused John Ioannidis to write his now famous paper, Why Most Published Research Findings Are False. That sounds very strong. But in some areas of science it is probably right.
http://rsos.royalsocietypublishing.org/content/1/3/140216

Bias:

"Australia's innovation agenda: embracing risk or gambling with public health?" https://theconversation.com/australias-innovation-agenda-embracing-risk-or-gambling-with-public-health-52003

False Negatives:

Ebola test http://www.washingtonsblog.com/2014/10/ebola-test-gives-false-negatives.html

False negatives https://manoa.hawaii.edu/exploringourfluidearth/chemical/matter/properties-matter/practices-science-false-positives-and-false-negatives

Peer Review:

http://www.nature.com/news/peer-review-troubled-from-the-start-1.19763

Bayesian Analysis

Bayes Decision Theory:
http://www.stat.ucla.edu/~yuille/courses/Stat161-261-Spring13/LectureNote2.pdf

Pascal's Wager: Choice under uncertainty https://en.wikipedia.org/wiki/Pascal%27s_Wager

Many philosophers consider the *initial plausibility* of a claim to be a factor in determining the burden of proof. This makes little differences as to who has the burden of proof but does affect the standard required for the justification to be found convincing. An extraordinary claim would require an extremely good justification. This concept is often stated as "extraordinary claims require extraordinary evidence."
http://bit.ly/1Tvdzjw

Questions?

Contact me at:

Jim L. Sanders, MBA
real@cox.net
520-322-0088
http://realestateappraisalandlitigation.com