"Constructive Skepticism" Volume 3 – Notebook #I: Model Risk

Chapter 8: The "Practical Significance" of "Spinach" and "Invisible Gorillas"

Previous chapters showed the problematic presence of "*Spinach*" [*Things we think unquestionably true but look ambiguously false after asking a few questions*] in research papers, and its organic growth from Model Risks such as:

- "*Statistical Illusions*" associated with the Measurement Problem,
- "*Roughness*" associated with the Preference Problem,
- "*False Reconstructions*" associated with "*Dimension Reduction*" & "*Scaling Bias*", and
- "*Misdirection*" from testing models and hypotheses.

The last chapter closed with the Validation Score Card based on Retirement Planning, and Hypothesis Testing "*Forms*", also shown below.

At this point in this notebook, and in addition to the "*Spinach*" analogy to sort the "*Real Story*" from the "*Good Story*", readers have also seen "*Tools, Checklists & Processes*" that address these problems and model risks:

- The Business Ecosystem Template for a holistic view of the "*Terrain*" (Introduction),
- The Template for Reading Research Papers for a holistic view of the "*Process*" (Chapter 1),
- The Matrix of "*Metiers*" and "*Professional Deformation*" to see the structural nature of "*Perspectives*" on the part of readers, and researchers (Chapter 2),
- The "*Wedge-Shaped Cylinder*" analogy to see the "*Willful Ignorance*" from model "*Dimension Reduction*" (Chapter 4),
- The table of "*Axioms, Assumptions & Hypotheses*" in Financial Economics to see the "*Boxes within Boxes*" limitations of models, and their recommendations (Chapter 4),
- The "*Mini-Fig*" analogy to see the impact of model "*Scaling Bias*" (Chapter 6), and
- The Archery "*Form*" analogy for Hypothesis Testing and Retirement Planning Processes to create the Validation Score Card for "*Forms*" in retirement planning research papers (Chapter 7).

This chapter regroups derivative calculations from Hypothesis Testing into "*Confusion Matrices*", that we call "*Targets*"", to validate the models in a retirement planning "*Form*". These "*Targets*" enable the formal introduction of the "*Invisible Gorilla*" analogy [*Things we do not see, but stand right in front of us*].

These "*Tools, Checklists & Processes*" provide context, ask questions, and deliver answers when reading research papers. They help us focus our attention, form an opinion,

and make predictions in fields filled with the "*False Reconstructions*" and "*Misdirection*" of the "*Attention Merchants*".

Validation Score Card for "*Forms*" in Retirement Planning Research Papers
Hypothesis Testing "*Target*": "*Confusion Matrix*" for Decision Making with the "*Modern Synthesis*"

Retirement Planning Model Steps by Sources of Validity	The Unknown Significance from the Absence of Hypothesis Testing	The "*Practical Significance*" of the ***Bayesian*** Updating of Individual Beliefs & Clinical Ambiguity	The "*Evidential Significance*" of the sample data (Likelihood Ratio)	The "*Statistical Significance*" of a Single Hypothesis against "*Randomness*" (***Fisher***)	The "*Statistical Significance*" of a Choice Between Two Hypotheses (***Neyman-Pearson*** or "*Modern Synthesis*")
Human Capital					
Life Trajectory					
Beliefs, Values & Expectations					
Household Income & Composition					
Social Capital					
Social Security					
Pensions					
Family					
Business Ecosystem					
Culture & Policy					
Taxes and Inflation					
Social Programs					
Financial Capital					
Account Vehicles					
Financial Assets					
Tangible Assets					
Interest Rates					
Credit Rating					
Market Expectations					
Discount Rates					
Consumption					
Budget & Forecasts					
Debt Management					
Absorbing Barriers					
Income Threshold					
Expense Threshold					
Risk Capacity					
Recommendations					
Risk Allocations (Exposures, Hedges, Insurance Contracts, Leverage, Reserves)					
Account Locations,					
Asset Allocations					
Product Selections					

"Confusion Matrix" for the "Modern Synthesis" of Hypothesis Testing

As shown below, the Table of Outcomes for Hypothesis Testing ("*Confusion Matrix*") with the "*Modern Synthesis*" reflects several quantitative measures of statistical doubt. This table summarizes the statistical doubt in answering one of the five questions in the Validation Score Card for "*Forms*" in Retirement Planning Research Papers.

This table applies to the "*Modern Synthesis*" for "*Statistical Significance*" between two Hypotheses, and its color scheme matches the order of the chart shown in Chapter 7 – Part B and titled: "*The Distribution of Observables and Hypotheses on the "Measurement Scale" based on Statistical Samples from a Population of Interest"* with the "*Null Hypothesis*" in red font, and the "*Alternative Hypothesis*" in green font, as shown below:

Table of Outcomes from Hypothesis Testing: Comparing the α, and the p-values Probabilities of False Positives ("*Spinach*"), False Negative ("*Invisible Gorillas*"), "*Sensitivity*", "*Specificity*", and "*Statistical Power*"

Hypothesis Testing	H(o) True	H(o) False
Test is Positive Reject H(o) p-value < α	False Positives (FP) Type I Error: Fail to Reject α = P(rejecting H(o) | H(o) true) "Spinach"	True Positives (TP) Correct Inference **1-β =** **P(Rejecting (H(o) | H(o) false)** (Sensitivity)
Test is Negative Fail to Reject H(o) p-value > α	True Negatives (TN) Correct Inference 1-α = P(failing to reject H(o) | H(o) true) (Specificity)	False Negatives (FN) Type II Error: Fail to See **β =** **P(failing to Reject H(o) | H(o) false)** "Invisible Gorillas"
Color Codes: Null Hypothesis Distribution Alternative Hypothesis Distribution	P(NotB|A) = TN / [FP + TN] = 1-α	Statistical Power P(B|A) = TP / [TP + FN] = **1-β**

The first column shows the results of the test based on the "*p-value*" described in Chapter 7, and its relationship to the critical value α:

- Test can be positive with a "*p-value*" lower than α, thus rejecting the "*Null Hypothesis*"
- The test can be negative with a "*p-value*" higher than α, thus failing to reject the "*Null Hypothesis*".

The second column shows, and names the consequences of rejecting, or failing to reject the "*Null Hypothesis*" when it happens to be true:

- In the case of the test rejecting the "*Null Hypothesis*", this creates a False Positive, also called Type I Error that Fails to Reject, that we call "*Spinach*"
- In the case of the test failing to reject the "*Null Hypothesis*", this creates a True Negative, a correct inference called "*Specificity*"

The third column shows, and names the consequences of rejecting, or failing reject the "*Null Hypothesis*", when the "*Alternative Hypothesis*" is true

- In the case of the test rejecting the "*Null Hypothesis*", this creates a True Positive, a correct inference called "*Sensitivity*"
- In the case of the test failing to reject the "*Null Hypothesis*", this creates a False Negative, also called Type II Error that Fails to See, that we call "*Invisible Gorillas*"

Thus, this table that summarizes answers that quantify statistical doubt in answers to the following Validation Score Card question:

- *What is the Statistical Significance" of a Choice Between Two Hypotheses based on* ***Neyman-Pearson****'s method or the "Modern Synthesis"?*

It also shows the the tradeoffs between "*Sensitivity*" and "*Specificity*" as well as "*Spinach*" and "*Invisible Gorillas*" [*An expression developed by* ***Christopher Chabris & Daniel Simons*** *to illustrate optical illusions in human perceptions such as* ***Edgar Rubin****'s "Figure" & the "Ground" Vase.*]

Simplifying this table by deleted the right column, and eliminating "Type I/Type II" errors summarizes answers to the Validation Score Card question: *What is the Statistical Significance of a Single Hypothesis against "Randomness" (Fisher)?*

Answering **Royall's** Other Questions

At this point, we have a "*Confusion Matrix*" that answers two out of ***Richard Royall***'s four questions:

- What specific Hypothesis Testing "*Process*" should I use to reveal the pros-&-cons for a single hypothesis?
  - Measuring "*Uncertainty*" with ***Fisherian*** Hypothesis Testing ("*p-values*")
- What specific Hypothesis Testing "*Process*" should I use to support a selection between two or more hypotheses?
  - Making a Decision with ***Neyman-Pearson*** Hypothesis Testing

How can we summarize statistical doubt to answer ***Royall's*** two other questions:

- What specific Hypothesis Testing "*Process*" should I use to change my subjective belief about validated vs. random results based on specific observations?
    - Updating "*Belief*" with ***Bayesian*** Hypothesis Testing
- What specific Hypothesis Testing "*Process*" should I use to evaluate the strength of the evidence in order to update beliefs, measure pros-&-cons against a single hypothesis, or make a decision between alternatives?
    - Measuring Strength of Evidence with ***Likelihood Ratios***

We can summarize answers to these last two questions in a different but related table as shown below:

Table of Probabilities for Diagnostic Classification Using "Natural Frequencies" for an Example with a Population Sample N = 1,000, a Condition Prevalence of 1%, a True Positive Rate of 80% and a False Positive Rate of 10%

Condition (+, -) x-axis		Positive Likelihood Ratio: LR+			Diagnostic Odds Ratio
Test (+,-) x-axis		8.00			36 to 1
Prevalence TP+FN 10		TPR / FPR			LR+ / LR-
No Cond. FP+TN 990					
F1-Score	True Positive Rate: TPR		False Positive Rate: FPR		Balanced Accuracy: BA
13.68%	TP / CP	80%	FP / (N-CP)	10%	40.00%
2 * [(PPV*TPR) / (PPV+TPR)	**Sensitivity,** Hit Rate				(TPR+TNR) / 2
Harmonic Mean Precision & Sensitivity	Probability of Detection, Power, Recall		Fall Out		
Positive Predictive Value: PPV	**True Positives: TP**		**False Positives: FP**		False Discovery Rate: FDR
7.48%	**8**		**99**		92.52%
TP / (TP+FP)	Correct Inference		Type I Error, False Alarm		FP / (TP+FP)
Precision	High Sensitivity Result,Rule In		Spinach		
False Omission Rate: FOR	**False Negatives: FN**		**True Negatives: TN**		**Negative Predictive Value: NPV**
0.22%	**2**		**891**		**99.78%**
FN /(FN+TN)	Type II Error, False Rejection		Correct Rejection		TN / (FN+TN)
	Invisible Gorillas		High Specificity, Result,Rule Out		
Threat Score (TS)	False Negative Rate: FNR		True Negative Rate: TNR		**Accuracy: ACC**
7.34%	20%		90%		89.90%
TP / (TP+FN+FP)	FN / CP		TN / (N-CP)	TN / (FP+TN)	(TP+TN) / N
Critical Success Index (CSI)	Miss Rate		**Specificity,** Negative Selectivity		
Population N 1000	Negative Likelihood Ratio: LR-				Matthews Correl. Coeff.: MCC
No Condition 99%	FNR / TNR	0.22			0.82%
Condition Prevalence					[(TP*TN) - (FP*FN)] /
CP 1% 10					SQRT(TP+FP)(FP+FN)(TN+FP)(TN+FN)

This table uses ***Gerd Gigerenzer***'s "*Natural Frequencies*" instead of ***Bayesian*** Probabilities, and provides the formulas. The end results between these two formalizations arrive at the same values. However,"*Natural Fequencies*" make it easier for readers to "*See for Yourself*", and reproduce the calculations.

Starting with the center of this "*Confusion Matrix*", we can see how the four quadrants in the middle of this table match the four quadrants from the previous "*Confusion Matrix*":

- True Positives
- False Negatives
- False Positives
- True Negatives

However, observant readers will notice that the left/right columns are reversed between the two "*Confusion Matrices*". This order will be changed in future versions of this notebook so that both "*Confusion Martrices*" have the same orientation.

These four quadrants highlighted in yellow show the "*Natural Frequency*" calculations derived from the input information, the size of the sample, the prevalence of the condition, the true positive rate, and the false positive rate of the test.

The cells with calculation shown in grey represent derivative calculations based on the input information and "*Natural Frequencies*". These include:

- The False Negative Rate, and the True Negative Rate
- The Positive Predictive Value, and the Negative Predictive Value
    - These answer the question about updating one's subjective belief
- The Positive Likelihood Ratio, and the Negative Likelihood Ratio
    - These answer the question about evaluating the Strength of Evidence.

This table shows additional calculations, and maps them logically in the context of the ealiers results. This gives us a map to place summary statistics in context, and to see what other results we may logically expect, or calculate on our own.

"*Targets*" for the Validation Score Card

We call these "*Confusion Matrix*" tables "*Targets*" because we can make an explicit match between the core quadrants with archery targets as follows:

- Shooting arrows in the center of the target with a tight grouping reflects "*Precision" & "Accuracy*" thus matches the top right quadrant of "*Sensitivity*".
- Shooting arrows loosely all over the target reflects imprecision & inaccuracy, thus matches the bottom right quadrant of "*Invisible Gorillas*".
- Shooting arrows in a specific side of the outer bands of the target with a tight grouping reflects precision & inaccuracy, thus matches the top left quadrant of "*Spinach*".
- Shooting arrows in a specific side of the outer bands of the target with a loose grouping reflects imprecision & accurac, thus matches the bottom left quadrant of "*Specificity*".

Evaluating "*Forms*" of Retirement Planning with these "*Tools, Checklists & Processes*"

Retirement planning research papers come in many "*Forms*" ranging from product-focused papers that address issues of portfolio management, to client-focused papers that address issues ranging from human capital, social capital, financial capital, consumption, etc.

Recent examples include the following paper, and monograph:

- ***Blanchett, David*** (2022), *Redefining the Optimal Retirement Income Strategy*, Financial Analyst Journal, Volume 79, Number 1
- ***Idzorek, Thomas M. & Kaplan, Paul D.*** (2024), *Lifetime Financial Advice: A Personalized Optimal Multilevel Approach*, CFA Institute Research Foundation

Readers can use the "*Tools, Checklists & Processes*" to "*See for Yourself*". Alternatively, they can wait for the reading notes for these and other papers that will appear in future Notebooks.