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

Chapter 7: What Can Long Bow Meditative Archery Do for Hypothesis Testing?

Part B: The "Forms" of Retirement Planning and Hypothesis Testing

Previous chapters showed the 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 (i) "*Statistical Illusions*" associated with the Measurement Problem, (ii) the "*Roughness*" associated with the Preference Problem, and (iii) the "*False Reconstructions*" associated with "*Dimension Reduction*" & "*Scaling Bias*". This chapter moves the discussion from model description to model testing, and the resulting problem of "*Misdirection*", and concludes with a Score Card to evaluate retirement planning papers.

Following ***Ole Peters***' frequent observations that one should look for concrete analogies in order to explain abstract "*Processes*":

- Part A of this chapter started by observing Hypothesis Testing through the lens of Longbow Meditative Archery. This gave us the analogy of the "*Form*" as a "*Process*" for model building, and a list of "*Purpose*" questions for Hypothesis Testing, as we seek ways to validate retirement planning models.
- Part B takes the measure of a messy state of historical and current affairs affecting statistical "*Theories, Methodologies & Methods*", "*Tools, Checklists & Processes"* and their matching "*Axioms, Assumptions & Hypotheses*". This review connects the "*Forms*" of retirement planning with the "*Forms*" of Hypothesis Testing in order to score (i) The lack of significance, (ii) The "*Practical Significance*", (iii) The "*Evidential Significance*", and (iv) the two types of "*Statistical Significance*" of retirement planning research papers.

The "*Forms*" of Hypothesis Testing: "*A Mess Full of Stuff*"

Chapter 7 - Part A: The Analogy showed how reading ***Richard Royall***'s 1997 book titled "*Statistical Evidence, A likelihood paradigm*" inspired the connection between Longbow Meditative Archery and Hypothesis Testing: ***Royall*** quantifies questions about "*Purpose*" in order to differentiate "*True Shots*" from bad shots, and lucky shots when it comes to determining the validity of research results. ***Royall*** brings clarity by differentiating the questions from the "*Forms*". One "*Form*" does not fit all questions. This connection leads to matching these questions with specific "*Forms*", i.e. statistical "*Processes*" that return specific answers. This matching of questions about "*Purpose*" with specific statistical "*Processes*" includes:

- 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 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
- 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***

These questions show that Hypothesis Testing does not look like the clear-cut program that one would hope to see in order to trust the science, and the "*Evidence-based*" research findings. Testing hypotheses for significance can take many "*Forms*", ask many questions, and provide a range of answers that lead to more confusion rather than more clarity. In the 1980s, when I was working on "*Task Environment*" analysis for Expert Systems in the AI Section of Arthur D. Little, ***Helen Ojha*** summed up these issues with the following quip: "*Life is a Mess, Full of Stuff*".

Hypothesis Testing as the Tuning of a Radio

The analogy of tuning a radio to find a signal brings clarity to this mess of statistical "*Theories, Methodologies & Methods*", "*Tools, Checklists & Processes"* and matching "*Axioms, Assumptions & Hypotheses*". The radio analogy helps us make the distinction between:

- "*Statistical Significance*" as a numerical tuning in the mathematical world, analogous to tuning the radio dial in order to find a specific frequency,
- "*Strength of Evidence*", or "*Evidential Significance*" ranging from weak & gibberish to strong & understandble as a semantic "*Perception*" of the listener, analogous to fine-tuning on the selected radio frequency, and
- "*Practical Meaning*", or "*Practical Significance*" as pragmatic value statement from the listener, analogous to understanding the information and content in the signal carried by the radio frequency.

Combining ***Royall's*** questions with this radio analogy gives us the following mapping for Hypothesis Testing:

- "*Practical Meaning*", or "*Practical Significance*" as pragmatic value statement from the listener, analogous to understanding the information and content in the signal carried by the radio frequency:
    - 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

- "*Statistical Significance*" as a numerical tuning in the mathematical world, analogous to tuning the radio dial in order to find a specific frequency:
    - 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

- "*Strength of Evidence*", or "*Evidential Significance*" ranging from weak & gibberish to strong & understandble as a semantic "*Perception*" of the listener, analogous to fine-tuning on the selected radio frequency:
    - 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***

Anders Hald

Anders Hald's 2007 book titled "*A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713 – 1935*" benchmarks the pace of adoption for best practice statistical "*Theories, Methodologies & Methods*", "*Tools, Checklists & Processes"* and their matching "*Axioms, Assumptions & Hypotheses*" in research papers. "*Parametric Statistical Inference*" seeks to find a value for an unobserved parameter based on the presence of sample observations, and ***Hald*** provides examples of the slow, and imperfect diffusion speed of statistical innovations in the real-world. Statistical Inference started with ***Laplace*** in 1774, and experienced two revolutionary re-conceptualizations in its developmental history. "*Parametric Statistical Inference*" started to achieve best practice status in 1956 – a development history of nearly 200 years, and as we will see below this is still a work in progress.

Using ***Hald's*** historical timeline, this work in progress developed as follows:

Starting with the development of Binomial Statistical Inference:

- ***James Bernoulli***'s 1713 Law of Large Numbers for Binomial distribution.
- ***DeMoivre***'s 1733 Normal approximation to the Binomial, and its generalization
- ***Bayes's*** 1764 Posterior Distribution of the Binomial Parameter and his rule for inductive inference

It continued with Direct Probabilities, and then moved to Inverse Probabilities:

- ***Laplace's*** 1774 Statistical Inference from Inverse Probability
- ***Gauss's*** 1809 derivation of the Normal distribution and method of the least-squares
- ***Edgeworth's*** 1908 and 1909 Genuine Inverse Method, and the equivalence of Inverse and Direct Probability in Large Samples
- ***Laplace's*** 1810 Central Limit Theorem

It transitioned from Inverse Probabilities to Frequentist Error Theory:

- ***Chauvenet's*** 1863 Frequentist Theory
- ***Pearson's*** 1990 Chi-Test for Goodness of Fit
- ***Galton's*** 1869 investigation s of Regression
- ***Gosset's*** Student's t Distribution

And transitioned again from Frequentist Error Theory to the Fisherian Revolution of "*Statistical Significance*":

- ***Fisher's*** 1912 Absolute Criterion
- ***Fisher's*** 1922 Parametric Model, and Criteria of Estimation
- ***Fisher's*** most important book, "Statistical Methods for Research Workers" reached its 14th, and posthumously published edition in 1970.

This historical development moved from idealized large sample theories to the practical reality of working with small sample sizes. ***Ronald Fisher*** recognized the problems that come from working with small samples, including: Unknown, non-Normal and uncontrolled factors. Small samples make it difficult to separate the signal of interest from the unknown, non-Normal and uncontrolled factors. ***Fisher*** realized that conceptualizing these factors in the form of "*Randomness*" made it possible to apply the mathematics of probability theory. ***Fisher's*** "*Statistical Significance*" quantifies the plausibility of chance as an explanation for observations and correlations based on small samples.

Moving on from ***Hald***'s book, and coming back of ***Royall***'s book, ***Fisher*** introduced the "*Null Hypothesis*", "*Significance Testing*", and "*p-values*" in 1925, and ***Jerzy Neyman & E.S. Pearson*** introduced "*Hypothesis Testing*", "*Type I Error*" and "*Type II Error*" in 1933. During their lifetime, these researchers argued back and forth about who had the better ideas. ***Royall*** compares ***Neyman-Pearson*** Hypothesis Tests with (***Fisherian***) Significance Tests ("*p-value*" procedures) as follows:

Neyman-Pearson Hypothesis Tests	(***Fisherian***) Significance Tests ("*p-value*" procedures)
Tests that answer the following question of "*Purpose*": How to choose one of two specified hypotheses, H1 and H2, on the basis of an observation X = x	Tests that answer the following question of "Purpose": For a single hypothesis, H, to measure the evidence against H represented by an observation X = x
Element 1: Two hypotheses (families of probability distributions) H1 and H2	Element 1: One hypothesis, H, called the "Null Hypothesis"
Element 2: A test function $\delta(x)$ that specifies which hypothesis to choose when X = x is observed: if $\delta(x) = 1$ we choose H1, if $\delta(x) = 2$ we choose H2	Element 2: A real-valued function t(x) that gives an ordering of sample points as evidence against H: t(x1) > t(x2) means that x1 is stronger than x2 as evidence against H
Element 3: Result is a decision or action, "Choose H1" or "Choose H2"	Element 3: Result is a number, the significance level, or "*p-value*", interpreted as a measure of the evidence against H; the smaller the "*p-value*" the stronger the evidence.

In the end, a so-called "*Modern Synthesis*" combined these ideas in a complex, yet incomplete, formalization.

The "*Modern Synthesis*" of Hypothesis Testing

From a statistical point of view, we live in "*Populations*", and experience life in "*Groups*". We can only measure data from "*Samples*", and for specific "*Variables*" in order to validate two types of "*Inferences*": Extending differences between "*Measures of Central Tendency*" from "*Samples*" to "*Groups*" or "*Populations*", or asserting levels of association between "*Variables*". Researchers can apply Hypothesis Testing to both types of "*Inferences*".

The validation of "*Effects*" in research papers tends to focus on the "*Modern Synthesis*" for Hypothesis Testing, a set of "*Theories, Methodologies & Methods*" that quantifies statistical doubt between two hypotheses in order to justify an inference from sample data. These hypotheses include the research's hypothesis, called the "*Alternative Hypothesis*", and a "*Null Hypothesis*", also called the "*Null Model*".

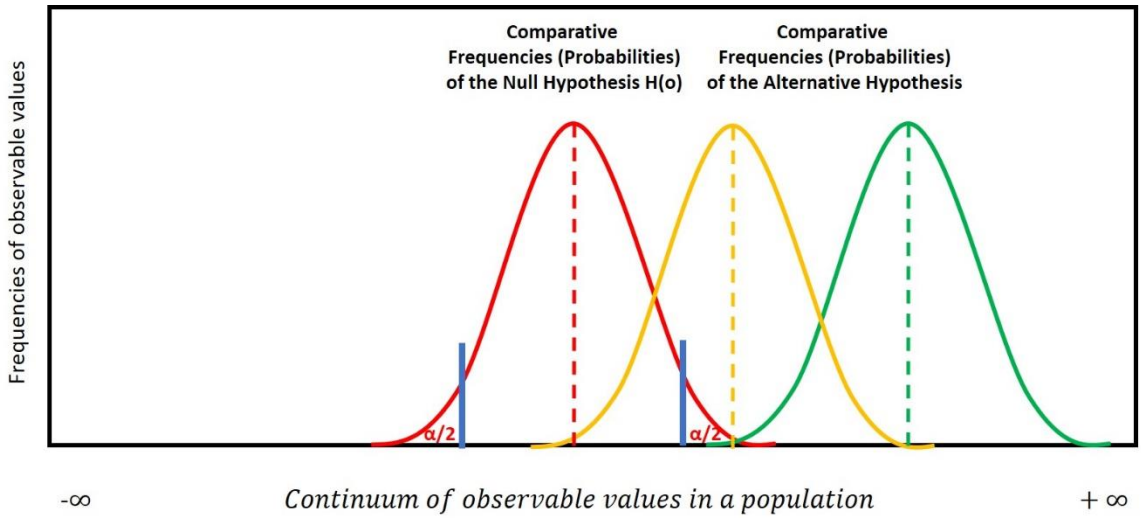
In their 2002 book titled "*Statistical Inference*", and at the beginning of Chapter 8 on Hypothesis Testing, ***George Casella & Roger L. Berger*** start with the following definition: Definition 8.1.1 "*A hypothesis is a statement about a population parameter.*"

Additionally, they describe Hypothesis Testing procedures, or Hypothesis Tests as rules that specify for which sample values the decision is made to accept the "*Null Hypothesis*", or the "*Alternative Hypothesis*" as true. Finally, they specify Hypothesis Tests with a test statistic as a function of the sample, such as the sample mean. For instance, a Hypothesis Test might specify that the "*Null Hypothesis*" should be rejected if the mean is greater that a specific value.

The following chart illustrates this "*Process*" as follows:

- The X-axis of the chart shown below uses a "*Measurement Scale*" from negative infinite to positive infinity to quantify observables in the statistical populations of interest.
- Its Y-axis represents the frequency of observation for the relevant observables on the X-axis.
- These axes provide a framework to visualize the use of the distribution of certain random variables such as the Normal Distribution, the Z-distribution, the t-distribution, the F-distribution, and the Chi-Square distribution.

Chart: The Distribution of Observables and Hypotheses on the "Measurement Scale" based on Statistical Samples from a Population of Interest



The distributions shown on the Chart have "*Normal*" shapes with two equal tails. The area under theses curves sums up to 1, and represent the probability of drawing an observable value on the "*Measurement Scale*" in a random selection process.

- The distribution in red-font represents the probability distribution, and position on the "*Measurement Scale*" of the "*Null Hypothesis*", centered on an un-interesting mean value for the research program.
- The distribution in green-font represents the probability distribution, and position on the "*Measurement Scale*" of the "*Alternative Hypothesis*", centered on the hypothesized mean value of the results of the research.

The "*Modern Synthesis*" compares the position of the means, shown in the chart as the vertical, dashed lines. The distance between the mean of the "*Null Hypothesis*" and the mean of the "*Alternative Hypothesis*" shows the difference between the probability distributions around those means.

The value, α, represent a key step in the testing process, a user-defined probability that defines the acceptable probability for an incorrect rejection of the "*Null Hypothesis*", thus rejecting it when it is true. In this example, the user-defined selection of the probability, α, divides between the two tails (see "α/2" in the top chart), and maps to specific numbers called "*Critical Values*" on the "*Measurement Scale*" as shown by the short vertical bars in blue-font. The value of α is commonly and arbitrarily set at 0.05 (5%) in Social Science research.

Statistical tests for data created from measurements on a scale (i.e. "*Scaled Variables*") require the presence of assumptions that include normally distributed data, independence of observations, and homogeneity of variances. Departures from these assumptions weaken the "*Statistical Meaning*" of these tests.

For instance, the t-test used to analyze the "*Statistical Significance*" of the difference between the "*Means*" of two "*Groups*" from a "*Sample*" [e.g. the "*Null Hypothesis*" vs. the "*Alternative Hypothesis*"] combines variance and sample size [i.e. The square root of variance divided by sample size] to create a distribution of "*Standard Errors*". The relevant t-distribution comes from a family of distributions, and is selected as a function of sample size expressed as "*Degrees of Freedom*". As sample sizes increase the t-distribution becomes more and more "*Normal*". Finally, "*p-values*" represent calculated probabilities based on the t-test, and related to the research observables. This statistical test becomes significant, meaning rejection of the "*Null Hypothesis*", when the "*p-value*" is lower than α.

Note that sample data must meet the following assumptions for a valid use of the t-test as a "*Method*" to make inferences about the difference between the "Means":

- Sample data from the Population comes from a continuous "*Random Variable*"
- Samples prove representative of their matching population
- Samples reach large enough sizes to detect differences between averages
- Observations in each sample have the property of independence from one another
- Sampled observations conforms to a normal distribution
- Samples display the same amount of variation abour their means

These assumptions enable the use of the Central Limit Theorem for the test, and the "*p-value*" from the Normal Distribution.

Thus, "*p-values*" measure of how much signal-looking noise, one could expect to see in "*Randomness*" before it ceases to look like "*Randomness*". Assuming the "*Null Hypothesis*" is true, a large "*p-value*" shows the probability that a repeated experiment would get a test statistic as or more extreme than the first result. A small "*p-value*" would suggest that we observed a surprising outcome, and discredit the "*Null Hypothesis*".

"*p-values*" do not provide a true or false statement about theory-free, "*As-if*" models represented by an "*Alternative Hypothesis*". They do not measure the size of an "*Effect*", the "*Power*" of a study, the "*Evidential Significance*", or the "*Practical Significance*" of a result. "*p-values*" should mark a step, not the end, in a "*Process*" to ask questions, and to qualify answers in support of a model based on a theory other than "*Randomness*".

The "*Alternative Hypothesis*" in green-font shows a mean value (dashed-line) to the right of the right-hand side "*Critical Value*". This would show a "*p-value*" lower than α. We would then reject the "*Null Hypothesis*" because we are now more confident that the sample observations come from the "*Alternative Hypothesis*" (green distribution)" rather than the "*Null Hypothesis"* (red distribution). On the other hand, the distribution in yellow-font shows that distributions can overlap with the distribution of the "*Null Hypothesis*" making it harder to differentiate with "*p-values*", up to the point of not-rejecting the "*Null Hypothesis*". Thus, rejecting the "*Null Hypotheses*" does not imply proving the "*Alternative Hypothesis*".

Despite its apparent complexity, and sophistication this process remains incomplete theoretically as well as practically. This leaves researchers with the problem of interpreting the "*Meaning*" of rejecting or not rejecting hypotheses. It also leaves researchers with the problem of making judgements about the "*Meaning*" of the test statistics such as the means, the variances, and the distributions based on their match or mis-match with the assumptions that must be met for a valid use of the "*Tools*".

As the reader can "*See for Yourself*", the "*Modern Synthesis*" for Hypothesis Testing has turned into a complicated and confusing exercise. Problems include:

- The double-negative logic of its "*Process*"
- The conflation of ***Fisherian*** testing evidence for-or-against "*Randomness*" for a single hypothesis with ***Neyman-Pearson*** decision-making between two hypotheses
- The need for user-selected types of test distributions
- The need for user-defined "*Critical Values*" with matching probabilities, and
- The reliance on user-calculated test statistics to determine "*p-values*".

Further, in the second half of the 20th Century, its software-driven & mechanical application by researchers, as a short-cut solution to this complexity and required assumptions led to a pervasive irreproducibility of test results, research findings & prescriptive recommendations in the Social Sciences. One may be able to master the "*Process*", but bad habits, and especially habits automated with software defeat the value of good exercise.

This brings us back to ***Richard Prum***'s lament about using "*Null Hypotheses*" that truly reflect "*Randomness*" in biology. ***Prum*** see a habit, a structural bias in his area of research where "*Null Hypotheses*" do not reflect "*Randomness*", but instead reflect an implicit assumption of purposeful, objective "*Evolutionary Fitness*". Thus, creating pre-ordained conclusions of purposeful fitness, instead of remaining open to the possibility of random, subjective "*Beauty Contests".*

How many established hypotheses, models, and theories filled with purposeful, evolutionary fitness cover up realities akin to ***Prum's*** "*Beauty Contests*" in retirement planning research, and how can we sort these out in order to keep the rest?

The "*Forms*" of Retirement Planning

Retirement planning requires the development of models that work together in ways similar to the "*Form*" of Longbow Meditative Archery described in Chapter 7, Part A. Combining models of clients in the context of their individual goals, and emotions as well as their ecosystem flows, property and liabilities enables the creation of retirement planning "*Forms*" built-up from models that range from the prescriptive based on ensemble expectations, to the predictive based on individual growth rates. These include:

- Models for an individual's Human Capital including their ability to work, household composition, life trajectory & expected longevity
- Models for an individual's Social Capital including expected Pension/Social Security benefits, and the Social Factors that shape the context of their life including Family Values, Social Culture, Business Cycles, Government Policy, Taxes, Inflation, and social programs such as Medicare
- Models for an individual's Financial Capital including account balances and risk exposures, as well as forecasts of capital market expectations and discount rates
- Models for an individual's budget, and how it changes their life trajectory, including the management of debt
- Models for an individual's risk exposures ranging from subjective feelings about risk (Risk Perception, Risk Tolerance, Risk Aversion, Risk Utility) to measurable risk exposures (Risk Capacity as the difference between the net present value of their assets and liabilities in the household balance sheet), including the management of options, hedges and insurance contracts
- Models for an individual's beliefs and expectations about absorbing barriers over the retirement horizon including phases of personal activity, personal bankruptcy, end-of-life, and bequests

Selections from these models create a specific "*Form*" for retirement planning. These "*Forms*" also have a choice of Targets to benchmark expected success ranging from the optimization of cash-flow forecasts over the retirement horizon to the management of Risk Capacity as a current dollar measure in the household balance sheet.

Finally, advisors can apply a personal value-judgement about these "*Forms*", and view the successive model steps as a sum of "*Random Variables*" with diverging outcomes over time, or as an average of "*Random Variables*" with converging outcomes over time. The former creates a value judgement of "*Safety-First*", and the latter creates a value judgment of "*Probability-First*".

Scoring Papers on Retirement Planning

Combining the "*Forms*" of Hypothesis Testing with the "*Forms*" of Retirement Planning creates a Score Card to see if models presented in research papers have:

- Unknown significance from the absence of Hypothesis Testing
    - Which research papers amount to unfounded "*Narrative Engineering*", and can be ignored?
- "*Practical Significance*" from ***Bayesian*** updating of individual beliefs, and clinical ambiguity.
    - Which research paper provide test results that we can use to update our beliefs, and reduce ambiguity?
- "*Evidential Significance*" based on the sample data (***Royall's*** Likelihood Ratios)
    - Which research papers rest on empirical "*Fee of Clay*", and create fragile findings likely to fail reproducibility?
- "*Statistical Significance*" based on a single hypothesis tested against "*Randomness*" (***Fisher***)
    - Which research paper test their preferred theory against the possibility of "*Randomness*"?
- "*Statistical Significance*" based on a choice between two hypotheses with either the ***Neyman-Pearson*** method, or the "*Modern Synthesis*" method.

The following table closes this Chapter to brings these ideas together in the form of a Score Card that readers can use to "*See for Yourself*".

Score Card for "*Forms*" in Retirement Planning Research Paperw

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					