IPRPD
International Journal of Business & Management Studies
ISSN 2694-1430 (Print), 2694-1449 (Online)
Volume 03; Issue no 07: July, 2022

Inferential Decision Support Systems: Pedagogic Enhancements for Delivering Statistical-components in an Auditing Course

Edward J. Lusk[1]

[1] *Emeritus Professor: [Statistics] The Wharton School, University of Pennsylvania, Pennsylvania USA Emeritus Professor: [Accounting] School of Economics & Business, SUNY: Plattsburgh, Plattsburgh, USA Emeritus Chair: [Economics] International School of Management: Otto-von-Guericke, Magdeburg, Germany*

Abstract[i]

Context *Statistical testing to arrive at decision-making-intel useful in setting the risk-level for the audit-client has been a PCAOB best-practices-staple for many years. Initially, such inferential-profiles required auditing students to be able to calculate all the component-parts of the inferential tests—e.g., Means, Correlations, Standard Deviations, and α-Rejection Regions to mention a few—in forming the required inferential-intel. Recently, the pendulum has swung to the other extreme; now there are software platforms that accept data and spew-out results with little contextual guidance.* ***Both are the Bain of Pedagogic-sensibilities in the STEM-context. Deliverable*** *In this research report, a normative pedagogic Decision Support System [DSS] is offered that: (i) is initialized* ***Ex-Ante*** *by the* ***a-priori*** *specification of the nature of the particular audit-context so as to form an indication as to the risk-level of the audit-client and so to justify the related audit-testing needed—this is termed the* ***Anchoring-Phase****, (ii) then the random-sampling results are collected and are made available to the auditor—this is termed the* ***Ex-Post-Phase****, (iii) in this* ***Ex-Post****-phase the inference-platform selected* ***Ex-Ante*** *is often changed—this is termed the* ***Conditioning-Phase****, and (iv) then the inference-results are presented by the DSS and used in the calibration of the risk-level of the audit.* ***Research Question*** *Does the* ***Ex-Post Conditioning-Phase*** *result in a change in the inferential-protocol selected by the auditor in the* ***Ex-Ante****-Phase? If so, this would be considered a violation of the logic of the standard inferential model that is based upon the* ***Ex-Ante*** *selection of the inference-platform. In a debrief with the students using the DSS, the DSS-results-profile was used as the instructional-platform to better understand the correct application of the inferential-profile that is used to inform the auditor. A second instructional aspect is that the DSS is VBA-programed to provide numerous guidance alerts that are focused on inappropriate testing aspects that can be elected but that are inconsistent with the standard inferential logic. Discussion of these aspects seem to enhance understanding of inferential testing. Additionally, a few of the "programmatic issues" that are currently inherent in using Excel[2019]™ to create certain inference-profiles are addressed. The DSS corrects for these Excel-issues.* ***This DSS is offered as an illustration of the pedagogic benefits of inferential training enabled by using course specific software modeled on the requirements of the inferential model****. The DSS, a VBA open-access modular platform, is available as a download without cost or restrictions on its use either academically or professionally.*

Keywords: Audit Risk-level Calibration, Inferential Conditioning, Inferential Debriefing, DSS Alert Guidance
JEL (Classification): M42

1. Introduction

1.1 Overview

This research report began circa 2002 when I held a Chair at the Otto-von-Guericke Universität [OVG][International Masters Program] Magdeburg, Germany. This was the year that the Public Company Accounting Oversite Board: PCAOB[ii] was created by the Dodd/Frank Act [Title IX]: The Sarbanes-Oxley Act of 2002 under HR[3763 [30 July 2002:Public Law 107-204: 107th Congress[USA]]. The tipping-point that resulted in the creation of the PCAOB was the defalcations engineered by Andy Fastow[iii], and many others, of the *Enron* Corporation. Subsequent investigations into the veracity of the financials of listed companies brought to light the wide-spread fraudulent activities of: (i) many corporations during the 1990s in the USA, as well as (ii) the shocking complicity in these fraudulent endeavors of the Public Accounting LLPs charged with audit-assurance of the trading markets. As a historical side-bar the report of Segal, Mansa & Reeves (2021) is a must read in any Audit & Assurance or Accounting Information Systems Course. The PCAOB issued AS2 that offered guidance in the required conduct of the audits of traded organizations under the PCAOB's licensure contract with ANY firm offering assurance re: the "veracity" of the information offered by firms traded on exchanges. One of the themes running through AS2 and continuing to AS5 and the many interpretative discussions issued by the PCAOB, AICPA & SEC over the years is that the use of statistical inferential models to aid in the ***judgmental-assessment*** of the ***risk-level*** of the audit client so as to determine the level of testing needed in each unique audit. There are two authoritative references that focus on the critical nature of such ***judgmental-assessment*** in the conduct of the audit.

The first is given by the AICPA: Generally Accepted Audit Standards [GAAS (Note 2)]:

GAAS: GS1 The auditor must have adequate technical training and proficiency to perform the audit.

GAAS: SFW2 The auditor must obtain a sufficient understanding of the entity and its environment, including its internal control, to assess the risk of material misstatement of the financial statements whether due to error or fraud, and to design the nature, timing, and extent of further audit procedures.

To coordinate with the GAAS, ***the second*** authoritative source mentioned above is the PCAOB provides important elaborations and guidance relative to AS5[15Dec2017] wherein it is emphasized that the judgement of the auditor is a critical feature underlying the PCAOB-Audit.

PCAOB rule set: AS 1001: Responsibilities and Functions of the Independent Auditor:
"Page:5: 05. In the observance of the standards of the PCAOB, the independent auditor must exercise his judgment in determining which auditing procedures are necessary in the circumstances to afford a reasonable basis for his opinion. His judgment is required to be the informed judgment of a qualified professional person."

PCAOB rule set: AS 1010: Training and Proficiency of the Independent Auditor:
"Page: 9: 03. In the performance of the audit which leads to an opinion, - - -. *The junior assistant, just entering upon an auditing career, must obtain his professional experience with the proper supervision and review of his work by a more experienced superior. The nature and extent of supervision and review must necessarily reflect wide variances in practice. The engagement partner must exercise seasoned judgment in the varying degrees of his supervision and review of the work done and judgments exercised by his subordinates, who in turn must meet the responsibilities attaching to the varying gradations and functions of their work."*

PCAOB rule set: AS 2305: Substantive Analytical Procedures:
"Page9:09. The auditor's reliance on substantive tests to achieve an audit objective related to a particular assertion may be derived from tests of details, from analytical procedures, or from a combination of both. The decision about which procedure or procedures to use to achieve a particular audit objective is based on the auditor's judgment on the expected effectiveness and efficiency of the available procedures. For significant risks of material misstatement, it is unlikely that audit evidence obtained from substantive analytical procedures alone will be sufficient." (See paragraph .11 of AS 2301, The Auditor's Responses to the Risks of Material Misstatement.)

PCAOB rule set: AS 2315: Audit Sampling:
"Page.207:01. Audit sampling is the application of an audit procedure to less than 100 percent of the items within an account balance or class of transactions for the purpose of evaluating some characteristic of the balance or class. This section provides guidance for planning, performing, and evaluating audit samples."

"Page.207.02 The auditor often is aware of account balances and transactions that may be more likely to contain misstatements. He considers this knowledge in planning his procedures, including audit sampling. The auditor usually will have no special knowledge about other account balances and transactions that, in his judgment, will need to be tested to fulfill his audit objectives. Audit sampling is especially useful in these cases."

"Page.207.03 There are two general approaches to audit sampling: nonstatistical and statistical. Both approaches require that the auditor use professional judgment in planning, performing, and evaluating a sample and in relating the evidential matter produced by the sample to other evidential matter when forming a conclusion about the related account balance or class of transactions. Either approach to audit sampling can provide sufficient evidential matter when applied properly. This section applies to both nonstatistical and statistical sampling."

These AICPA:GAAS and PCAOB guidance principles suggest and rationalize that: (i) the judgmental decisions of the auditor are the basis of the Assurance Audit, and (ii) reliance on statistical testing is needed to form a reasoned judgmental assessment of the risk-level of the audit so as to satisfy the PCAOB best-practices mandate under Sarbanes-Oxley. The important implication of this is:

The Basis of the collection of Audit Evidence is (i) a Random Sample of Sufficient size from the Population of Sensitive AIS accounts under Audit Scrutiny, and (ii) that the inferential testing protocol is formulated by the auditor <u>before</u> the sampling profile is collected, then, and only then, can the inferential results be relied upon to create risk-calibration that is: relevant and reliable given the usual inferential guidelines.

1.2. Research Plan Given this context, following is the plan of this research report:

I. A clarification of the purposes of this research report,

II. Using an illustrative example of risk-setting calibration in an audit context, offer a detailed presentation of the standard inference-protocol,

III. Discuss four statistical decision examples taken primarily from statistical-texts that used a Decision Support System [DSS] in a classroom setting to collect experiential indications on two modes of interference calibration: ***Ex-Ante*** & ***Ex-Post***,

IV. Profile these ***Ex-Ante*** and the ***Ex-Post*** inference elections of the OVG-students to create, for the first time, information on the possible ***bias*** introduced by observing the sampled information, and

V. Elaborate selected extensions of the DSS as a Pedagogic tool.

2. Clarification of Intention: To Enhance the Inferential Common Sense

2.1 Inference Issues There are two operational issues in play in the audit-context:

The First is that in setting the risk-level of the audit, inferential testing is one of standard tools in the panoply of the auditor. However, experiential evidence garnered over the years suggests that the application of inferential testing in the audit-context has drifted off the usual and standard guidelines that underly the assumptions of inferential testing models. Thus, ***the Second*** issue to be addressed is to introduce, in the auditing course, a protocol that motivates the ***correct*** usage of inferential testing so as to ***correctly*** inform the auditor re: inferential testing results so that the audit risk is ***correctly*** calibrated that, of course, leads to the ***correct*** commitment of audit testing to rationalize the two-opinions: the COSO & the Standard Opinion that the Financials are fair representations of the results of operations over the audit year in conformity with the GAAP-lens. Note the epistemological and ordered linkage of the term ***correctly***. Of course, it would be almost impossible to arrive at the ***correct*** level of the commitment of audit-resources to collect reliable information of the client's risk level—the last-link in audit-testing—if the initial inferential calibration is not ***correctly*** configured.

Thus, the approach in this research report is to first address the initialization of the audit testing and risk setting stage; the usual academic presentation of the inference-modeling protocol focuses on the computational components needed to form the p-value [or α-rejection region] of the False Positive Error [FPE] derived from the ratio of the Difference of: The *Ha* [The Alternative of Interest] vis-à-vis that of the *Ho* [Null] <u>to</u> The Standard Error derived from the randomly sampled information. These "computational" foci do not address the required order of the creation of the information-logistic where the inferences are required in risk-setting. Most audit texts and auditors are not alerted that there are two-phases to correctly use the standard inference protocol:

> ***Phase I*** The auditors ***Ex-Ante*** [before sampling information is collected and profiled] using their ***a-priori*** experiential-intel indicate: (i) the nature of *Ho* [the test against benchmark often noted as the Null], (ii) the statistical testing platform, and (iii) the FPE-risk [p-value] that is the decision-making driver for calibrating the level of audit risk. This information-set is the ***population-screen*** in that it is produced without knowledge of the sampling-profile. In this context, if these Phase I parameters are used to create the inferential profile re: the risk-level of the audit-client then the audit InCharge [IC] is testing the unique experience that the InCharge brings to the audit ***not conditioned on the result of a sampled-profile***. To be clear, if the auditor used a random sample to create the testing protocol, then the IC losses the credibility of testing the Experience and Judgment that the IC is charged to use by the PCAOB in executing of the audit.
>
> ***Phase II*** The auditor ***Ex-Post*** [after the audit-sampling information is collected and profiled] creates the inferential-results using the ***Ex-Ante*** parameters to create the audit-decision-making intel that was the driver of the inferential risk-setting study. ***Inferential Alert*** Using the sampling results as the driver and the rationale for ***re-parameterization the Ex-Ante profile*** then is evidence that the ICs are rejecting their judgmental and experiential-intel, in a capricious way or

> perhaps to better align their "judgement" with the observed sampling profile. This is flawed inferential-execution. The FPE[p-value] has no inferential meaning if the inferential protocol is developed using the sampling results—*end of homily*.

In this context, for my auditing-courses, the following was conversationally offered:

> *It is not permitted to collect a sample of sufficient size randomly drawn from a population well delineated and then to use that sample profile to form the testing-protocol. The reason for this is: (i) that biases the inference calibration in that usually the mean of the sample becomes an* ***anchor-value*** *that all but guarantees that the inferential testing will be directional, and (ii) [ignoring the equality event] using the expectation that 50% of time the mid-point of the* [*1-*$\alpha/_2$]Confidence Interval will be < than μ *[the true mean of the population], and 50% of time the mid-point of the* [*1-*$\alpha/_2$]Confidence Interval will be > than μ*, and the divergence around* μ *will be symmetric only in expectation, this means that a particular sampled mean will not be a reliable indication of the directional or non-directional calibration needed to arrive at a consistent p-value re: the False Positive Error[FPE]. In practice, it is often the case, that the auditor is aware of the sampled-profile and tacitly uses this information to create the inferential protocol.* ***This is to be avoided****.*

2.2 The Audit-context: Benchmarking the Audit-Risk

To better understand the critical organizational logistic of the inferential testing frame in the audit-context as it pertains to the calibration of the risk-level following is an example presented to the students in the AIS-course at the School of Business & Economics: SUNY: Plattsburgh, USA Spring 2019. The context presented to the course was offered as follows[iv]:

2.2.1 Ex-Ante Information

Context: In the audit context, the risk-level of the audit needs to be established; this is one of features of the licensure-test used by the PCAOB in their evaluation-review of Audit LLPs. As implied above there are many versions of protocols that can be useful in collecting relevant inferential-information important in setting the risk-level for the certification audit.

Assume that the IC has decided to use a benchmarking-protocol to collect a risk-indication of the audit. Specifically, the auditor will be using the *Bloomberg Market Navigation Platform*™ to collect information on firms that have been coded as ESG[ISS[1] or ESG[ISS[10].

Institutional Shareholder Services™ [ISS] is a well-established organization that has been reporting their assessment of the Corporate Governance Risk [CGR] for firms traded on active stock exchanges.

ISS offers a well-expressed, articulated, and focused set of protocols to develop the richness as well as the nuances of CGR. Lusk & Wells (2021) note:

> *A singular distinction for ISS is that circa 2014 they have been integrated as a third-party data provider into the renowned Environment, Social, and Governance [ESG©]-platform offered by Bloomberg™ Professional Service. - - -. The ISS-screening Pillars Following are details from the ISS Methodology Guide (2020)ii regarding their four screening or evaluation pillars: [Bolding Added]: ISS ESG Governance Quality Score (GQS) is a data-driven scoring and screening solution designed to help institutional investors monitor portfolio company governance. At both an overall company level and along topical classifications coveringBoard Structure, Compensation, Shareholder Rights, and Audit & Risk Oversight, scores indicate relative governance quality supported by factor-level data*

ESG Fund Rating - ISS (issgovernance.com, p.5), ISS [note: **Bolding** added]

> *Factors used to assess risk-related concerns for a given company in each market are based on the same principles that form the foundation of ISS' global benchmark voting policy. A score in the* ***1st decile*** *indicates higher quality and relatively lower governance risk, and, conversely, a score in the* ***10th decile*** *indicates relatively lower quality and higher governance risk. These scores provide an at-a-glance view of each company's governance risk relative to their index and region. The individual factor breakdown takes a regional approach in evaluating and scoring companies, to allow for company-level comparisons within markets where corporate governance practices are similar*

2.2.2 ISS:Inferential Design

In this context, the four-ordered steps in the client-risk-setting-protocol are:

1. ***Ex-Ante*** create a standard inferential testing protocol detailing (i) the *Ho*[Null], (ii) the *Ha*[Rejection Option], and (iii) note the p-value that rationalizes rejection of the *Ho*[Null], then and only then
2. Take a random sample of firms that have been ISS-rated as ESG[ISS[1] or ESG[ISS[10],
3. For each of these firms record their latest Current Ratio [CR] [Current Assets / Current Liabilities], and
4. For the audit client, if their CR is in the 95% Confidence Interval of the ESG[ISS[1]-firms then the risk level indication will be ***Low***; if the CR is in the 95% Confidence Interval of the ESG[ISS[10]-firms then the risk level indication will be ***High***. Otherwise, the ISS-measure will not be used in the risk-calibration of the audit-client.

The ISS-example will illustrate a critical aspect of the inferential model, called ***Conditioning***. To present the conditioning aspect, which is a point of emphasis of this research report, we will assume an actual example that was used in an academic setting.

2.2.3 Un-Conditioned Ex-Ante: Assumptions made by the IC

Given the ISS-scoring, it seems reasonable that the firms with the ***best*** ISS-rating on Corporate Governance Risk[CGR]—i.e., the ESG[ISS[1] firms with ***Lowest*** CGR overall—will have the best systems of Internal Control over Financial Reporting [ICoFR]. If the audit client has a CR-profile that is in the 95% Confidence Interval of the CR of the ESG[ISS[1]-firms in the random sample, it is not unreasonable to use that as a positive indication of ***Low Risk*** of the client's system of ICoFR. Additionally, given the ISS-scoring it seems reasonable that the firms with the ***worst*** rating on Corporate Governance Risk—i.e. the ***Highest*** CGR—will have the overall the least adequate systems of ICoFR. If the audit client has a CR-profile that is likely in the 95% Confidence Interval of the CR of the ESG[ISS[10]-firms, it is not unreasonable to use that as a positive indication of ***High Risk*** of the client's system of ICoFR.

2.2.4 Linking ISS-scoring with ICoFR & The CR Profile & Audit Risk Calibration

With this as the ***Ex-Ante*** Un-Conditioned experiential mindset, the IC further assumes that the population average of CRs of the ESG[ISS[1]-firms will be Greater [larger in value] than the population average of CRs of the ESG [ISS[10]] firms. *Rationale*: *It seems reasonable that firms with better ICoFR with respect to CGR will make wiser decisions on the allocation of corporate resources that overall will result in higher CR-average-profiles than firms with less adequate ICoFR with respect to CGR. In this case, thus the IC* ***proffers*** *the following inferential testing-frame*:

Ex-Ante Un-Conditioned Hypotheses Relative to the Current Ratio [CR]
Ex-Ante Expectation : Mean [CR of ISS[1]] > Mean [CR of ISS[10]]

Ho**: Case{A} :** ***Mean [CR of ISS[1]] = Mean [CR of ISS[10]]*****:Null**

Note: ***Consistency Failure*** if the Mean [CR of ISS[1]] is **<** the Mean [CR of ISS[10]] then the directional ***Null p-value is >50% and so NOT Rejected. No further testing.***

Ha1 : Case{B} : CR of ISS[1] ≠ CR of ISS[10] Not expected or tested

Ha2 **: : If the Null is rejected then Ha is: Case{C} : CR of ISS[1] > CR of ISS[10] for which the directional p-value is evaluated by the IC.**

Ha3 : Case{D}: CR of ISS[1] < CR of ISS[10] Not expected or tested

This is the CORRECT inferential decision-screen. In fact, these are ONLY four cases in the inferential model protocol. However, for comparison, let us assume that the Audit-Staffer profiled the following ISS-random sample and uploaded it to the IC-audit link.

Ex-Post Test Profile:
ISS[1]-Group
Sample Size **10**
Mean **2.15**
Standard Deviation **1.14**
ISS[10]-Group
Sample Size 1**6**
Mean **2.82**
Standard Deviation **1.57**

The IC ***happens*** to look this ISS-sampling profile over. Assuming that at this point, the IC creates the inferential screen for the client-risk profile and the IC ***inappropriately*** uses the above ISS-sampling profile to form the inferential testing framework. In this case, the inferential judgment of the IC will be ***conditioned—changed/effected/biased*** by the above sampling-profile resulting in the following testing profile:

2.2.5 Re-Linking the ISS-scoring with ICoFR & The CR Profile & Audit Risk Calibration

With this as the ***Ex-Post Conditioned*** experiential mindset, the IC now assumes that the population average of CRs of the ESG[ISS[10]-firms will be Greater [larger in value] than the population average of CRs of the ESG [ISS[1] firms. *Rationale*: *It seems reasonable that firms with better ICoFR with respect to CGR will make wiser decisions on the allocation of corporate resources and that since higher risk requires higher return that overall ISS-Firms with lower risk will have lower return on allocated resources resulting in lower CR-average-profiles than firms with less adequate ICoFR with respect to CGR that are able to select risker projects. In this case, the re-proffered or re-spun inferential logic will be:*

Ex-Post Conditioned Hypotheses Relative to the Current Ratio [CR]

Ex-Post Expectation: Mean [CR of ISS[10]] > Mean [CR of ISS[1]]

Ho: **Case{A} : *Mean [CR of ISS[1]] = Mean [CR of ISS[10]]*:Null**

Note: ***Consistency Failure*** if Mean [CR of ISS[1]] > Mean [CR of ISS[10]] then the directional ***Null p-value is >50% and so NOT Rejected. No further testing.***

Ha1 : Case{B} : CR of ISS[1] ≠ CR of ISS[10] Not expected or tested

Ha2 : Case{C} : CR of ISS[1] > CR of ISS[10] Not expected or tested

***Ha3* : Case{D}: : If the Null is rejected then Ha is: Case{D} : CR of ISS[10] > CR of ISS[1] for which the directional p-value is evaluated by the IC.**

The critical point is that if the IC fixed the testing parameters in the ***Ex-Ante*** Phase I ***and then used them***, as is consistent with the assumptions of inferential testing, the ***correct result would be***:

Ex-Ante Inferentially ***Correct*** Profile:

Ex-Ante Expectation : Mean [CR of ISS[1]] > Mean [CR of ISS[10]]

Ho: **Case{A} : *Mean [CR of ISS[1]] = Mean [CR of ISS[10]]*:Null**

Note***: Consistency Failure*** if the Mean [CR of ISS[1]] is < the Mean [CR of ISS[10]], then the directional ***Null p-value is >50% and so NOT Rejected. No further testing.***

Ex-Post Random Sampled-Profile: Mean [CR of ISS[1]] = 2.15; Mean [CR of ISS[10]] = 2.85

In this case, the directional FPE[p-value] will be > than 50% as the central-tendencies are reversed from the IC's ***Ex-Ante*** expectation. For, example:

$$t_{df=23} = [2.15 - 2.85]/0.533 = -1.257$$

Directional p-value =T.DIST.RT($-$1.257, 23) = **0.889 ≡ 88.9%**

2.2.6 FPE Decision

The IC's expectation was not directional founded as the Mean [CR of ISS[1]] = 2.15 < the Mean [CR of ISS[10]] = 2.85; thus, the Null-***risk*** is >50% and so rejecting the *Ho*[Null] is not warranted. ***Decision: There is evidence that the ISS-screen is NOT likely to be a valid screen for calibrating the audit-risk.***

However, if the testing decision-election was conditioned by the sampling profile, ***an abhorrence to inferential testing***, and the IC changed the inferential context as noted above, ***the incorrect or re-spun results would have been***:

Ex-Post Inferentially ***Conditioned Incorrect*** Profile:

Ex-Post Expectation : The Mean [CR of ISS[10]] > The Mean [CR of ISS[1]]

Ho: **Case{A} : *Mean [CR of ISS[10]] = Mean [CR of ISS[1]]*:Null**

Note: ***Consistency Failure*** if the Mean [CR of ISS[10]] is < the Mean [CR of ISS[1]], then the directional ***Null p-value is >50% and so NOT Rejected. No further testing.***

Random Sampled-Profile Mean [CR of ISS[1]] = 2.15; Mean [CR of ISS[10]] = 2.85

***Ha3* : Case{D}: : If the Null is rejected then Ha is: Case{D} : ISS[10] > ISS[1] for which the directional p-value is evaluated by the IC.**

In this case, the FPE[p-value] for the directional test of Case{A}is 22.1% as the central-tendencies are in the expected ordered direction. For, example:

$$t_{df=23} = [2.85 - 2.15]/0.533 = 1.257$$

Directional p-value =T.DIST.RT(1.257, 23) = **0.221 ≡ 22.1%**

FPE decision: The Null-risk is <25% and so rejecting the *Ho*[Null] is in the reasonable-zone. ***Decision***: There is *suggestive* evidence that the ISS-screen may likely be a valid screen for calibrating the audit-risk with respect to CGR in the audit-context.

2.2.7 Simple Summary

Here it is clear that using the correct ***Ex-Ante*** calibration casts doubt on the reliability of using the ISS-taxonomy to create risk-intel for the audit-client. Simply, the ***Ex-Ante*** expectation upon which the IC will calibrate the risk-level is NOT founded in that the mean of the CRs of the ISS[1] group is not larger than those of the ISS[10]. Simply, the IC would do well to not rely on the ISS-polar groups as a reliable measure the nature of which speaks to the risk-calibration of the audit-client. However, ***if*** the IC creates the ISS-calibrations expectation using the sampling results, then. in this case, there is not unreasonable evidence that the ISS-firms could logically provide logical risk-calibration for the audit-client.

To reinforce the temptation of using the Condition-intel of the sampling-profile to set up the inferential context, the same ISS-information set that was presented to the SUNY-students in the AIS-course. These results are presented in Table 1.

Benchmarking Audit Risk with the ISS {Polar Groups ISS[1] & ISS[10]

Phase I Test	**Ex-Ante Case {A}**	**If for Case{A} Ho is rejected then Case {B, C or D} is Selected**		
Phase II if needed		**Case {B}**	**Case {C}**	**Case {D}**
Ex-Ante Election	***100%[3%]***	***35%***	***30%***	***32%***
Inference Action	If Ho is Not Rejected: End of Analysis	Ha1 ISS[1] ≠ ISS[10]	Ha2 ISS[1] > ISS[10]	Ha3 ISS[1] < ISS[10]
SUNY: Final Profile	**N/A***	**12%**	**0%**	**88%**

Table 1 ISS[1] v. ISS[10], n=32 *Not Tracked

Source: Author Collected

2.2.8 Discussion

ISS-Audit Risk Setting Recall that the IC is interested in the question: "***Is there evidence that there is a difference between the CRs of the firms classified as ISS[1]—Low CG-Risk v. ISS[10]—High CG-Risk?***" The ISS ***Ex-Ante anchoring***-profile is interesting. Initially, the ISS-classification model, presented above, was discussed with the SUNY-students. They were asked if they, as the IC, expected that the ISS-classification could provide useful risk-calibration intel in the audit-context. This is the Case{A} stage of Phase I of the ***Ex-Ante-stage***. Only 3% of the SUNY-students, individually acting as the IC, seem to have assumed there would NOT be a difference in the CR-profiles as between the ISS[1] & the ISS[10] firms. The 97% that did assume that there would be a CR-profile difference seemed to be basing that opinion on (i) the belief that the ISS-categorization was rationale and objective and so the CRs would follow logically the polar ISS-categorization groups, (ii) others felt that the ISS[10]-firms where there was less "control" would have been able to "manage" their system of ICoFR and thus create a better CR-profile, and (iii) some felt it could go either way. This produced "basically" a uniform split over the Cases {B[35%], C[30%] & D[32%]}—***the anchoring-effect***. Then, the actual profile was revealed to the students. This ***conditioning*** information produced a dramatic re-organization of the inferential-testing-profile. After the sampling profile was made available to the students, then in the ***Ex-Post*** phase BUT before the p-values were calculated, the students dramatically changed their expectations to: Cases {B[12%], C[0%] & D[88%]}. This is a ***shocking*** indication of the strength of the ***Conditioning*** effect *vis-à-vis* that of the Ex-Ante ***Anchoring*** Expectations. ***Summary The students focused on the ISS-mean CR relationships: [ISS[1]:2.15] & [ISS[10]:2.85] go from a more or less uniform Anchoring-expectation over the three Cases: {B, C or D} to the vast majority expecting that Case{D[88%]} would be the inferential case of interest as a Conditioning over-ride to their Ex-Ante expectations.***

With this observation it is clear that the temptation to inappropriately use the sampled-profile to condition the ***Ex-Ante*** inferential testing profile needs to be addressed. This SUNY-experience was the last in a series of inferential tests conducted and so used as an indication of currency. These inferential tests were first conducted when I held the OVG: Chair. At this time the audit world was change by the PCAOB as noted above. In the Topics of Current Interest section of my courses *circa* 2002, I introduced the importance of the PCAOB, the AS2: audit rules, risk calibration and inferential testing. Also given the decision latitude that I had as the Chair, I created four inferential vignettes that offered inferential testing on *Common Knowledge* topics drawn largely from texts used at Wharton. To demo, the temptation to Condition the inference protocol, I used one or two of these to illustrate the Conditioning effect. This was a valuable "quasi-experimental design" to collect "impact" data on conditioning that we used to profile the then developing audit rules that were used to get the Markets in the USA under control and restore confidence in the trading world. In the next section of this research report, these OVG-testing vignettes are presented and their Conditioning aspects detailed and discussed.

3. The Set of Penultimate Inferential Exercises: First Tested at the OVG

Following are the four exercises that were developed, tested and used to sensitize the OVG-students to the critical importance of executing the inferential model as it is intended so as to develop the useful intel. For exposition, these example will be stated as there were given. They will be summarized to better underscore the apparent natural tendency to condition the ***Ex-Ante*** inferential parameters using the sample profile. Also, not all four of these were given prior to discussing the then developing PCAOB-audit-risk setting context. Usually, one of these four were discussed as the penultimate experience.

3.1 Sample Profile of March Madness

Source: **Tamhane, A. & Dunlop, D**. (2000). Statistics & Data Analysis, Prentice Hall 2000: ISBN:0-13-744426-5[**pp. 303-304; slightly modified**].

3.1.1 *Ex-Ante* Context [March Madness]

A Division 1 Women's Basketball coach lost an important March Madness game due to an inability of the players to make Free Throws. In the next season, she introduced a set of practice protocols designed to improve the percentage of Free Throws made. She used the percentages from the previous season as the test-against benchmark. This historical benchmark was X%. During the practice pre-season-games, she collected information on the percentages of Free Throws made after execution of her Free Throw Training Protocols. **{PoI}** ***Note that in the Ex-Ante phase there are no sampling results presented and additionally there are no measured values noted. The reason for this is critical to the execution of the inferential model to avoid the conditioning effect.***

3.1.2 Expectations Ex-Ante Phase

Initially, the students using their *a-priori* experiential-intel considered the following information:
Case {A}Does this inferential testing context seem as though that there is likely to be a rejection of *Ho: There is not an important impact due to the proposed training*?
If you answered that Training *could have an effect*—i.e., rejected *Ho* which of the following seems to be the likely case:
Case{B}Training could have a positive impact <u>or</u> it could have a negative impact on the percentage of Free Throws made,
Case{C}Training will likely have a positive impact on the percentage of Free Throws made, or
Case{D}Training will likely have a negative impact on the percentage of Free Throws made.

Their answers were recorded and are presented in the **Ex-Ante Elections** row of the March Madness De-Brief Profile Table 2

Ex-Post Phase Then the students were given the following sampled information:

Previous to Free Throw Training: **Population A**
Previous Season: March Madness: ***Ho*** = **70%**
After Free Throw Training: **Population B**
Trials **400** Free Throws: Successful Free Throws **300**

Hypotheses:
Ho Case{A}: Training Performance = Previous Performance [*Null*]
$\boldsymbol{H_{a1}}$ Case{B}: Training Performance ≠ Previous Performance
$\boldsymbol{H_{a2}}$ Case{C}: Training Performance > Previous Performance
$\boldsymbol{H_{a3}}$ Case{D}: Training Performance < Previous Performance

Then the students recorded their Ex-Post selections *for the above four cases. These are recorded in the* ***Ex-Post Elections*** *section of* March Madness De-Brief Profile Table 2

3.1.3 Debrief of March Madness

Following is an example of an Inference Profile Table that served as the discussion platform at the OVG to enhance the understanding of the nature of inference calibration and evaluation.

OVG Trial March Madness Free Throw Training

Phase I Test	**Case {A}**	**If for Case{A} Ho is rejected then Case {B, C or D} is Selected**		
Phase II if needed		**Case {B}**	**Case {C}**	**Case {D}**
Ex-Ante Election	***100%[25%]***	***38%***	***58%***	***4%***
DSS p-value	**2.91%**	**2.91%**	**1.46%**	**>50%**
Inference Action	If Ho is Not Rejected: End of Analysis	Ha1 Training Performance ≠ Previous Performance	Ha2 Training Performance > Previous Performance	Ha3 Training Performance < Previous Performance
OVG: *Ex-Post* Profile	**25%[2%]**	**3%**	**94%**	**1%**

Table 2 March Madness De-Brief Profile, n = 87
Source: Author Collected

3.1.4 Summary Profile of Table 2

Discussion Recall, there are two phases: The ***Ex-Ante*** & The ***Ex-Post*** to the inferential testing that are enabled by the DSS-protocol.

Ex-Ante The students using their ***a-priori*** experiential-intel recorded their ***Ex-Ante*** Elections by considering the information profile: **Context[**MM**]** *sans* any measured values. These are recorded in the ***Ex-Ante Election*** row. {***PoI***} ***The Ex-Ante Elections are the Selections of the Nature of the inferential testing frame that will BEST create the most meaningful information—independent of the actual p-value for evaluating the FPE.*** For Case {A}, it was suggested that everyone record their ***Ex-Ante a-priori*** expectation THAT, given their expectation of the ***non-directional*** test of Case {A}, the Null[*Ho*] will NOT being rejected thus ending the analysis. Thus, 100% was recorded; in the **[]s** is the percentage of individuals that expected that the Null[*Ho*] would not be rejected—specifically 25%. Additionally, these students were re-queried and thus asked the following question:
Assume that your expectation was not correct and you ***now believe that Ho should have been rejected****; in this case, which of the Cases {B, C or D} would you have selected?*
Thus, in Row[3]: Cols [3, 4 & 5] the sum of the those elections will sum to 100%.

Ex-Post Then, after the OVG-students were given the actual sample profile—i.e., with the measured values of the sample, but ***before*** they computed the p-value, they were asked the following:

> **Case {A}**Does this inferential testing context seem as though that there is likely to be a rejection of *Ho: There is not an important impact due to the proposed training?*
> If you answered that Training could have had an effect which of the following seems to be the likely case:
> **Case{B}**Training could have a positive impact or it could have a negative impact on the percentage of Free Throws made,
> **Case{C}**Training will likely have a positive impact on the percentage of Free Throws made, or
> **Case{D}**Training will likely have a negative impact on the percentage of Free Throws made.
> Their answers were recorded and are presented in the **Ex-Post Elections** of the March Madness De-Brief Profile Table 2.

{**PoI**} The correct p-values are recorded in Row [4] shaded. ***These were NOT given to the students; rather they generated them using a DSS developed for course***. The March Madness context was given as a Quiz to prepare the students for a term-test. For the quiz, only 25% of the OVG-students actually considered Case{A} and recorded their expectation at the ***Ex-Post*** stage. Perhaps, the students wanted to conserve quiz test-time and so eliminated the suggestion to first consider Case{A}. Of these 25%, about 2% decided not to reject the Null of *Ho*. Thus, 98% went to the Phase II-stage to test Cases:{B, C or D}. For that group, the percentages in Row[6] are the testing activity of the 98% who did reject *Ho* in the ***Ex-Post*** analysis.

The summary of this recording activity at the ***Ex-Post*** stage is:

> ***Case [B]***: Many of the OVG-students changed their ***Ex-Ante a-priori*** opinion after seeing the sample-profile and left Case {B}. This resulted ***Ex-Post*** in **3%** of the OVG-students electing Case{B} as their decision-election for testing using the DSS.
> ***Case [C]:*** Many of the OVG-students changed their ***Ex-Ante a-priori*** opinion after seeing the sample-profile and joined Case {C}. This resulted ***Ex-Post*** in **94%** of the OVG-students electing Case{C} as their decision-election for testing using the DSS.

Case [D]: A few of the OVG-students changed their ***a-priori*** opinion after seeing the sample-profile and left Case {D}. This resulted ***Ex-Post*** in **1%** of the OVG-students electing Case{D} as their decision-election for testing using the DSS.

The approximate conditioning change from the ***Ex-Ante*** to the ***Ex-Post*** was **62.1% [[94% – 58%]/58%].**

3.2 Single Population: Real-Valued Variates Automobile Tires Warranty

Source: Ott, R.L. (1993). *An Introduction to Statistical Methods and Data Analysis*. 4th Ed: [ISBN:0-534-93150-2] Duxbury Press: [**pp: 237-238**][v]

3.2.1 Ex-Ante Given Information: Context

A tire manufacture that sells passenger tires advertises that their tires will allow a standard passenger car to travel at least *X-miles* before the tires will fail the usual uniform state inspection standards often noted as "The Lincoln-Head Penny Frontier Test". A consumer magazine, that offers verification information on such manufacture-claims, decides to conduct a test of the *X-miles* claim. ***To be clear***: It is assumed that *X-miles* is the high-end of the gold standard in the USA for these tires to PASS the tire-wear state-standard.

3.2.2 *Ex-Post* Given Information: Inferential Design

The Consumer Magazine randomly selected ten (10) test-tires from different Retail stores in three cities. The tires are tested to the point of Failure--i.e., where the test-tires would ***not have passed the State Inspection test***. The magazine selects the Higher Limit of the Manufacture's Claim of 42,000 miles. This is the usual likely conservative testing frontier.

Hypothesis The Null to be tested is: ***Ho*** μ= 42,000 miles. Logically, this is a directional test to see if the manufactures-claim to avoid failing the state-testing of standard driving is warranted. In this context, rejecting **Ho** will ONLY occur if the population of actual tire-wear is LESS than 42,000 Miles. The Question is "How much less than 42,000 Miles provides convincing p-value evidence that the Manufactures-claim is not likely to be founded?”

In this case, if ***Ho*** is rejected then ***Ha*** is accepted that the manufacture's testing-claim is ***not*** validated/founded at some FPE-risk developed from the p-value.

Test Profile
Sample Size 10
Ho = 42,000 or more miles
Ex-Post Test: Mean: 41,000 miles
StDev = 3.59011

Hypotheses
Ho Case{A} : Ho μ= 42,000[Null]
Ha1 Case{B} : H_{a1} $\mu \neq$ 42,000
Ha2 Case{C} : H_{a2} $\mu >$ 42,000
Ha3 Case{D} : H_{a3} $\mu <$ 42,000

OVG Trial Tire Warranty Verification Test

Phase I Test	**Case {A}**	*If for Case{A} Ho is rejected then Case {B, C or D} is Selected*		
Phase II if needed		**Case {B}**	**Case {C}**	**Case {D}**
Ex-Ante Election	***100%[15%]***	***42%***	***3%***	***55%***
DSS p-value	40.1%	40.1%	>50%	20.1%
Inference Action	If Ho is Not Rejected: End of Analysis	Ha1 $\mu \neq$ **42,000**	Ha2 $\mu >$ **42,000**	Ha3 $\mu <$ **42,000**
OVG: Final Profile	**30%[6%]**	**4%**	**2%**	**88%**

Table 3 Automobile Tire Warranty Testing Profile, n = 112
Source:Author Collected

3.2.3 Tire Warranty Summary

In this case, 15% of the OVG-Students initially decided that *Ho* would have a non-directional p-value that would not suggest that *Ho* would not be the state of nature. These students were re-queried resulting in the profile in row 3. Then, ***Ex-Post***, after the sample profile is presented to the OVG-students, their performance profile is: 30% actually considered Case {A} and 6% expected to fail to reject that the Null[*Ho*] was the state of nature. ***Ex-Post*** the

conditioning effect—before any p-values were computed—was profound resulting in 88% winding up selecting Case {D}. **This change was 60.0% [[88% – 45%]/45%]**.

3.3 Two Populations : Frequency Variates Test Performance Females v. Males

Source: Test Dataset from Introductory Undergraduate Statistics ! & II Courses at the Wharton School of the University of Pennsylvania

3.3.1 Ex-Ante Given Information: Context

We collected over three semesters the scores earned for the Introductory Undergraduate Statistics I & II Courses at the Wharton School of the University of Pennsylvania. USA. The dataset was for only one of the Professors. There were X students who completed these courses. The test-question posed was:

Is there a difference in the final grades with respect to Gender?

3.3.2 Ex-Post Given Information

The measurement was the percentage of students that received a B or Higher as the Final Note for the semester. The Profile of the student Pool accrued was:

Female Students [F]
n= 274
Grade[B or Higher] 129 Students
Percentage: 47.1% [129/274]

Male Students [M]
n= 309
Grade [B or Higher] 135 Students
Percentage: 43.7% [135/309]

Hypotheses

Ho Case{A}: Female Performance = Male Performance [Null]
H_{a1} Case{B} : Female Performance ≠ Male Performance
H_{a2} Case{C} : Female Performance > Male Performance
H_{a3} Case{D} : Female Performance < Male Performance

OVG Trial Wharton Gender Profile

Phase I Test	**Case {A}**	**If for Case{A} Ho is rejected then Case {B, C or D} is Selected**		
Phase II if needed		**Case {B}**	**Case {C}**	**Case {D}**
Ex-Ante Election	***100%[50%]***	***40%***	***5%***	***55%***
DSS p-value	**20.6%**	**20.6%**	**10.3%**	**>50%**
Inference Action	If Ho is Not Rejected: End of Analysis	Ha1 Female Performance ≠ Male Performance	Ha2 Female Performance > Male Performance	Ha3 Female Performance < Male Performance
OVG: Final Profile	**10%[5%]**	**15%**	**79%**	**1%**

Table 4 Statistic Course Grades Gender Profile n = 98
Source: Author Collected

3.3.4 Wharton Gender Analysis Summary

In this case, ***Ex-Ante*** 50% of the OVG-Students initially decided that *Ho* would have a non-directional p-value that would not suggest that *Ho* would not be the state of nature. These students were re-queried resulting in the profile in row 3. Then, ***Ex-Post***, after the sample profile is presented to the OVG-students, their performance profile was: 10% actually, considered Case {A} and 5% expected to fail to reject that the Null was the state of nature. ***Ex-Post*** the ***conditioning effect*** was also profound resulting in 79% winding up selecting Case {C}. **This change was 1,480% [[79% – 5%]/5%].**

3.4Two Populations : Real-valued Variates Sheep Tapeworm Infestation

Source: Ott, R. L. (1993**)**. *An Introduction to Statistical Methods and Data Analysis*. 4th Ed: [ISBN:0-534-93150-2] Duxbury Press: [**pp: 267-270**]

3.4.1 Ex-Ante Given Information: Context

Tapeworm infestation is impossible to control for free-range grazing Sheep without a medical intervention. The only possible pro-action is to treat the incidence of Tapeworm infestation with specific drugs. The Inferential Question then addresses the Effectiveness of the Treatment.

3.4.2 Ex-Post Given Information

Inferential Design: Two populations of Sheep were randomly selected from grazing stock at a randomly selected farms. Fourteen [14] Sheep were randomly selected from the herd-populations. They were then randomly assigned to one of TWO Groups: { **Group A:Treatment using Drug X** & **Group B: Control NO Application of Drug :X**}. The Drug was applied to the Seven Sheep in Groups A. During the trial, one of the Group B-Sheep was hit by a tractor and so removed from the study. After the suggested time, the 13-Sheep were slaughtered and two-technicians independently counted the Tapeworms. The accounting was reconciled.
The Sample Profile [Tapeworm Infestation Counts]

Group A Treated Sheep
Sample Size **7**
Mean **9.000**
Standard Deviation **6.218521**

Group B Non-Treated Sheep [Controls]
Sample Size **6**
Mean **40.1**
Standard Deviation **16.067669**

Hypotheses
Ho Infestation: Case{A} : Treated Sheep = Non-treated Sheep [Controls] Null
$\boldsymbol{H_{a1}}$ Infestation: Case{B} : Treated Sheep ≠ Non-treated Sheep [Controls]
$\boldsymbol{H_{a2}}$ Infestation: Case{C} **:** Treated Sheep > Non-treated Sheep [Controls]
$\boldsymbol{H_{a3}}$ Infestation: Case{D} **:** Treated Sheep < Non-treated Sheep [Controls]

OVG Trial Sheep Tapeworm Infestation

Phase I Test	**Case {A}**	**If for Case{A} Ho is rejected then Case {B, C or D} is Selected**		
Phase II if needed		**Case {B}**	**Case {C}**	**Case {D}**
Ex-Ante Election	***100%[1%]***	***5%***	***1%***	***94%***
DSS p-value	**0.42%**	**0.42%**	**>50%**	**0.21%**
Inference Action	If Ho is Not Rejected: End of Analysis	Ha1 Treated Sheep ≠ Non-Treated Sheep	Ha2 Treated Sheep > Non-Treated Sheep	Ha3 Treated Sheep < Non-Treated Sheep
OVG: Final Profile	**5%[1%]**	**1%**	**0%**	**98%**

Table 5 Test of Treatment v. Control n=67
Source: Author Collected

3.4.3 Sheep Infestation Summary

In this case, the 1% of the OVG-Students initially decided that *Ho* would have a p-value that would not suggest that Ho would not be the state of nature. These students were re-queried resulting in the profile in row 3. Then, ***Ex-Post***, after the sample profile is presented to the OVG-students, their performance profile is: 5% actually, considered Case {A} and 1% expected to fail to reject that the Null was the state of nature. ***Ex-Post*** the ***conditioning effect*** was rather low resulting in 98% winding up selecting Case {D}. **This change was 4.3% [98% – 94%]/94%].**

3.5 Overall Summary **The very interesting aspects of these four OVG-studies are displayed in Table 6.**
OVG Trial Summary

	March Madness	**Tires Tested**	**Wharton Gender**	**Sheep Infestation**
Ex-Post Case {A}	**25%**[2%]	**30%**[6%]	**10%**[5%]	**5%**[1%]
Conditioning Change Percentage	62.1%[<0.0001]	60.0%[<0.0001]	1,480.0%[<0.0001]	4.3%[11.8%]
p-value Confusion	1%	2%	1%	0%

Table 6 Collected Highlights from the Four OVG-Trials ***Source: Author Collected***

Discussion of Overall Summary

The take-aways from these OVG-studies offer clarity regarding the way that inferential statistical needs to be "controlled" to align the actual-results with the assumptions that underly the *expected* meaning of the results.

Codex Clarification For Table 6 In the Conditioning-row the p-value is noted in the **[]s**. For example, for the OVG March Madness test, the largest change was for Case {C}: where the inferential context was that training would have a positive-effect. Initially—i.e**.,** ***Ex-Ante***—58% of the OVG-students selected Case{C} as the *Ha* relative to the *Ho*[Null]-test. However, in the ***Ex-Post*** context many re-posited their selection; this was likely due to the ***Conditioning effect*** of the sampled-profile. The ***Ex-Post*** result was that 94% of the students selected Case{C} as the *Ha* for the *Ho[Null]*-test. The p-value of the ***Ex-Ante*** *Ho[Null]* of 58% as the test-benchmark for the ***Ex-Post*** result of 94% has a p-value < 0.0001.

The three summary points of de-briefing interest are:

3.5.1 Consideration of the Case{A}Ho[Null] for the Ex-Post Phase

There are two instances where it is recommended that the nature of the *Ho[Null]*-test is considered: The ***Ex-Ante*** Phase where only the general context is articulated and the ***Ex-Post*** where the sample-profile is known. The intention is to have the analyst reflect on the inferential ***gestalt—Will the random sample likely offer evidence that there is an effect relative to Ho?*** By reflecting on the testing decisions made at the ***Ex-Ante*** & ***the Ex-Post*** phases, the analyst can learn about the quality of the information that has been collected re: ***Does the sample profile aid in probing the inferential questions that are of interest?*** Such ruminations, considering the ***Ex-Ante*** and then the ***Ex-Post***, are a critical aspect of the inferential protocol. What we learn from the above OVG-inferential tests is that in the ***Ex-Ante*** phase most all the students recorded their expectations re: the veracity of *Ho and the Case that would be offered as the Ha*. However, in the ***Ex-Post*** phase, when time was a consideration, many fewer students recorded their reflections. This likely occurred as in the ***Ex-Post*** phase there was a quiz in play. Specifically, only at most 30% of the time did analysts, before they computed any p-values, render an expectation as to their assessment of their testing profile of *Ho vis-à-vis* Cases {B, C or D}. ***Implication*** *This does not bode well for expecting that when time is an important consideration, as it always is, that analysts will use these* ***Ex-Ante*** *&* ***Ex-Post*** *Case{A} Ho[Null]-reflections to enrich their inferential profiles and the related analytics.*

3.5.2 The Conditioning-effect dominated the Anchoring-effect

Recall ***Ex-Ante*** the OVG-students indicated their analytic-preference among the three cases: Case{A} v. Cases {B, C or D}. This is the ***anchoring***-event. However, when the sampled information was known, the OVG-cadre dramatically changed their inferential election except in the case where their ***a-priori*** election was an "obvious" prevision of the result of the sampling-profile as it seemed to have been in the Sheep Infestation example. This is very troubling as this is a bias in using the sample as the "GPS"-for electing the inferential- protocol. ***Implication:*** *The conditioning effect is relatively ubiquitous. When this is the case, it is not clear as to the meaning of the p-value as produced by the standard statistical model. Interestingly, in the case of the Sheep-Infestation in the debrief session, most of the OVG-students indicated that their* ***Ex-Ante*** *choice was Ha3 Case {D} as all drugs go through extensive testing and so are by definition most always relatively effective. Thus, a one-directional-test is actually the only logical choice and that direction is that of: Ha3: Case D. As noted in Table 5 this had the smallest conditioning effect. The largest effect was for the Wharton Gender scenario. In this case, the debrief focused on the experiential context for the OVG-cadre that were mostly students from the Euro-zone, China & the Balkans. They assumed, based upon their experiential-context, that the least likely case would be Case{C} where* ***Ha2*** *Case{C}: Female Performance > Male Performance. However, this* ***Anchoring*** *result, was overridden by the* ***Conditioning*** *effect in the* ***Ex-Post*** *where the sampling results were posted.*

3.5.3 A Positive Effect

The only encouraging information is that the logic of the p-value seems to be understood in that where the p-value was >50%, rarely—e.g., about 1% of the time—was this Case selected by the OVG-Students in the ***Ex-Post*** context before the p-values were known. In the debrief sessions, we discussed that when the p-value is >50% that only happens when the Mean of the sample is not aligned directionally with the expectation. Most of the students seemed to understand this.

4. The Interactive Pedagogic Inferential Evaluation Platform[IPIEP]: Decision Support System[DSS]

4.1 Overview

It is clear and disturbing that **Conditioning** is endemic in the creation of the final-version of the inferential testing platform. Simply, when the analyst is aware of the sampling-profile there is clear evidence that the sampling-profile overrides the ***Ex-Ante*** a-priori experimental judgment of the analyst. These conditioning changes in the nature of the inferential testing program are most troubling as they bias the inferential discovery phase of statistical testing by "aligning" it with the observed sampling profile. Rather than bemoan or curse this conditioning-temptation, it is better take a positive action to identify instances where such divergence from the standard inferential model are highlighted and offer "*tough-love guidance*". The question is how to create such a *Balanced Scorecard*™ feedback-loop. In the OVG-courses, a VBA-DSS was developed that queried the auditor and where there seemed to be divergence from the standard inferential modeling protocol, the DSS gave an VBA-Alert indication. Hopefully such interactive feedback will "Un-Condition" the Inferential Conditioning and over time the correct inferential information will be the norm.

Consider now the **IPIEP:DSS. {PoI} The IPIEP is a DSS that is available only as a download at no cost and there are no restrictions on it use.}**

4.2 IPIEP:DSS : Points of Interest

The IPIEP:DSS is a slight-variant of the DSS that was used in the academic settings. All the VBA-modules are open access and so code-modifications are possible. The IPIEP:DSS is interactive and all parameters that are entered activate a VBA-MsgBox that indicates what was just entered and usually indicates what the next DSS-action is in the activation-queue.

4.2.1 Ex-Ante Phase-Before the Sampling-Profile is Known

Query-Set A The first IPIEP:DSS query addresses the Conditioning-Bias. A *VBA[Yes/No]-codex* asks if the analyst has knowledge of the details of the sample-profile. If the *Yes*, the IPIEP:DSS indicates that there can be Conditioning biasing issues and the IPIEP:DSS ***will be terminated***. If *No*, the analysis continues.

Query-Set B In this case, if Query A indicates that the analyst ***does not have knowledge of the sample-profile***, then the analyst is asked to assume that the *Ho[Null]* will be tested. If the analyst expects that *Ho[Null]* is the state of nature—no expected Effect— then an Alert indicates that the IPIEP:DSS ***will terminate***. If the analyst indicates that there is likely an Effect relative to the *Ho[Null]*, then the IPIEP:DSS indicates that there are three cases to be tested. Using the **March Madness** example, the three cases, only one of which is to be selected, are:

Ha**[Case{B}]**Training could have a positive impact or it could have a negative impact on the percentage of Free Throws made,

Ha**[Case{C}]**Training will likely have a positive impact on the percentage of Free Throws made, or

Ha**[Case{D}]**Training will likely have a negative impact on the percentage of Free Throws made.

After the analyst selects the Case that will be tested relative to *Ho* and that Case is recorded as the ***Ex-Ante expectation***.

4.2.2 *Ex-Post* Phase-After the Sampling-Profile is Known

Query-Set C In the ***Ex-Post*** Phase, the details of the sampling-profile are provided to the analyst. The next query asks, with knowledge of the Sample Profile, does the *Ho[Null]* seem to be the likely State of Nature. If *Yes*, then that is recorded in the ***Ex-Post*** section and the DSS ***terminates***. If the *Ho[Null]* seems likely to be rejected, then a query is launched to ask which of the three cases will be tested: Cases {A, B or C}. ***Note at this juncture the FPE[p-values] have not been produced by the IPIEP:DSS as this may influence the Case-selection to be tested.*** IF the case to be tested noted in the ***Ex-Ante*** Phase differs from the case noted in this ***Ex-Post*** Phase, then the IPIEP:DSS notes that this ***could be*** an indication of a Conditioning-Bias. If the same case is tested in the ***Ex-Ante*** and this ***Ex-Post*** Phases, then the IPIEP:DSS indicates that this is a positive inferential indication. In either case, the IPIEP:DSS continues and the IPIEP:DSS is parameterized by the analyst.

Query-Set D In this final-stage, the sampling-profile is entered in the IPIEP:DSS; after the parameters are entered a non-directional p-value is displayed. This feature was not part of the OVG- or SUNY-tests. For the third time, the analyst is asked if the *Ho[Null]* is rejected. If the *Ho[Null]* is not rejected the analysis is ***terminated***. If the *Ho[Null]* is rejected, after considering the p-value, then the Case to be tested is entered and the related inference profile is presented by the IPIEP:DSS. The case actually tested is compared to the case selected at the ***Ex-Ante*** stage. If they are not the same, an Alert is offered that this may be evidence of a Conditioning Bias and the inferential results ***may be questionable***. Otherwise, the IPIEP:DSS offers that these results are likely in conformity with the nature of the inferential model. Note that there are six-instances where the IPIEP:DSS terminates or offers Alerts as to the creation of questionable inferential results. This is one of the positive features of the IPIEP:DSS—it has been designed to be very sensitive to the Conditioning bias.

In addition, if the analyst enters a Case that is inconsistent with the Sample-profile, to wit, the means of the sample and those expected are not aligned in order, then the IPIEP:DSS color-fills the mean-values and the cells where the Case-tested is indicated with a light-rose tincture. In this case, no p-values are produced and an indication of the inconsistency that would lead to a p-value that is > 50% is noted.

4.3 Summary These IPIEP:

DSS Alerts and termination loops are intended to create guidance so as to aid and guide the auditor in using inference models in the risk-setting context in the manner in which they were designed to be used.

5. Summary & Outlook

5.1 Summary

In this research report, an alert is offered, and defended experimentally, that Conditioning is not part of the usual inference modeling protocols and thus should be avoided. In the context of this study, the simple take-away message, *that should be obvious but apparently is not*, is:

> ***In the logistics of inferential testing most often the collection of the sampling information is the progenitor-event to the delineation of the inferential-protocol. In an academic context for courses in auditing or generalized decision-analytics where the inferential preparation of the students is introductory statistics, it is critical that their experiential a-priori Ex-Ante judgmental be the driver of the inferential-analytics; using the sampled information to create the inference-protocol is an anathema to inferential credibility and over-time will likely render decision-making less effective than if the inferential models were correctly calibrated. It is hoped that the IPIEP:DSS model can be used to create an inferential-culture that uses the sampling-profile after the inferential protocol has been detailed using the experiential judgment of the analyst.***

5.2 Outlook

5.2.1 Tracking Practice Reality

In the future, it would be useful to track the nature of the inferential-logistic, the p-values, and the related decisions made in actual inferential cases where managerial decisions are made. This information would be most instructive and over time could be used as a practical guide to p-value impact-frontiers *vis-à-vis* the application of the correct testing frame.

5.2.2 Tracking Academic Reality

If the IPIEP:DSS is used in the delivery of introductory courses, statistics or auditing, it would be useful to create a data-capture module to understand the important concept of Conditioning and if the IPIEP:DSS, in fact, delivers on the expectation that it will move decision-makers in the direction of reducing the instances of Conditioning.

5.2.3 Extensions

Finally, if possible, inferential models that are variants of the standard models, the ilk of which are treated by: Shadish, Cook & Campbell (2002) would be challenging but useful additions to the inferential delivery in STEM-courses. Also, there are other techniques that can be used. A most promising but relatively overlooked variant is Necessary Conditions Analysis [NCA]. See Dul (2016), Dul (2020) and Dul, van der Laan & Kuik (2020). The NCA-models are refinements of the inferential world that was effectively developed by R.A. Fisher.

Technical Appendix [TA]

A few idiosyncratic comments on the nature of the IPIEP:DSS.

TA1 IPIEP:

DSS Test of Inferential Validity The utility of the inferential-intel depends upon the validity of the sampling plan of the Binary Population and its formation into a "Normal" approximation to the exact Binomial Distribution[n, π]. This is a very interesting transfiguration, that has a quasi-homomorphic-mapping, that is due to the mathematics of the Central Limit Theorem [CLT]. See (Tamhane & Dunlop (2000, 5.1.1[pp. 170-172])) The standard validity-check on the <u>reasonability</u> of this CLT: Normal approximation to the correct Binomial is evaluated using the IPIEP:DSS for both the Single-population and the case where there are two populations. In these two cases, the Rule of 5s and the Rule of 10s are tested in the logical-staged order, where, $[N \times (\pi)]$ & $[N \times (1 - (\pi))] > \Delta$; $\Delta = 5$ then 10.[vi] This is noted as an inferential-validity screen. If the rule of 5s is not stratified for either of the testing cases, the IPIEP:DSS is terminated. The rule of 5s is noted as the frontier-case and the rule of 10s is noted as a strong indication of the CLT-adequacy. If these Rules are satisfied, then a Continuity-Correction [CC] is used. The CC is a heuristic to better match the Normal approximation to the correct Binomial probability. For the single population this is: Left Hand Side: $P(\mathbf{X} < x) \rightarrow P(\mathbf{X} < (x + 0.5))$ & Right Hand Side: $P(\mathbf{X} > x) \rightarrow P(\mathbf{X} > (x - 0.5))$ where: $\mathbf{X}$ is the Normal approximation of the probability-variate, and x is the judgment election for testing relative to $(N \times (\pi))$. In the case for two-sampled-populations, a CC is used to bring the p-value into an approximate alignment with the p-value of the Fishers Exact test for a related classification tableau.

TA2 the FPE :

A Clarification The FPE happens when everything is going according to the expectation-plan in the population however, due to that fact the analyst ONLY knows anything about the population from a random sample from that population, it could happen that one believes that there is a problem/issue when in fact this is not the case—i.e., one observes a "possible" divergence from the population-expectation however that divergence is a FALSE indication due to the random sample. The risk of the FPE occurring is measured by the p-value of the inferential-test. Thus, the FPE-risk—the p-value—are the odds that the observation is likely to happen by random sampling chance for the size of the sample taken from the population if the population-expectation is in fact ***true***. Simply, the FPE-p-value is the gambling-odds that there is no problem but due to sampling-chance there appears to be a divergence-issue. If the FPE-p-value is low—5% or 1% this suggests that the observed divergence could have happened by random sampling chance rather than being an indication of a real structural difference but it would be a ***rare-event***. So, the risk, of rejecting the Null as the true state of nature in favor of a real difference would be low:5% to 1% or so—i.e., rejecting the Null is low-risk and so usually warranted. Simply, given the gambling-odds there is ***likely*** a divergence.

TA3 Excel as the DSS:

Can Excel be pressed into service? Do we need a DSS to create the inference profile for population frequency tests such as March Madness? Actually, assuming that one is using *Excel[v.2019]*, the answer is "Sort of " for the following reason. For frequency tests, the p-value, not using the CC, is formed as follows:

$$\boldsymbol{z_{cal}} = {}^{[\boldsymbol{\hat{\pi}} - \boldsymbol{H_o}]}\!/_{\boldsymbol{s_\epsilon}}$$

Where: $\hat{\pi}$ is the observed frequency of the random sample, H_o is the Null frequency *a-priori* posited, $s_\epsilon$is $\sqrt{[Ho \times [1 - Ho]]/n}$, and n is the number of sampled events from the reference population.

However, the Excel[*Data*[*DataAnalysis*[*AnalysisTools]*]] Platform does not have a module where these parameters $\{\hat{\pi};\ H_o; n\}$ can be entered to produce in, due course, the FPE[p-value]. Thus, the analyst will need to create a function where z_{cal} is calculated. The p-value then can be found using the Excel functions:

=(1-(NORM.S.DIST(z_{cal},TRUE)))*2 for the Non-directional inferential intel or

=(1-(NORM.S.DIST(z_{cal},TRUE))) for the Directional inferential intel.

For this reason, a DSS is critical so that the students do not have to code the z_{cal}—to wit, the IPIEP:DSS is a needed computational platform. For completeness, the same is the case for the test of proportions for frequency tests from two populations.

TA4 Excel-Welch Glitch

The Excel version of the inferential test of the Means of two-sampled populations offers the Welch-ANOVA test that assumes that there is a difference in the variances or standard deviations of the two samples. The Excel-version uses the degree of freedom[*df*] as: df = [$n_1 + n_2 - 2$] to form the p-value. Tacitly, this assumes that n_1 = n_2 and the ratio of the standard deviations is 1.0. When this is not the case, the Excel-version of the *df*-computation results in a p-value that is slightly lower than the correct computation offered in Welch(1947) & Satterthwaite(1946); the version programmed in the IPIEP:DSS that is the correct df-computation.

TA5 Inference Codex

There is an inferential feedback VBA-indication that offers generalized experiential council. Specifically, in the ***Ex-Post*** phase when the IPIEP:DSS generates the p-value, the following is displayed:

If the p-value is > 0.35, Then the VBA-message displayed is = "There is NOT SUFFICIENT evidence that the Expectation is different from the Actual result to impact decisions."

If the p-value is >= 0.01 And If the p-value is <=0.35, Then the VBA-message displayed is " There is SUGGESTIVE evidence that the Expectation is different from the Actual and so MAY impact decisions."

If the p-value is < 0.01, Then the VBA-message displayed is "There IS SUFFICIENT evidence that the Expectation is likely different from the Actual result to impact decisions."

[PoI] ***The p-value calibration of 35% as the frontier value is my experiential guideline used with the OVG-students. Others may have different cut-points, usually in the range*** **[*>25% to <50%*].**

Works Cited

Dul, J. (2016). Necessary Condition Analysis (NCA): Logic and methodology of ‘‘Necessary but Not Sufficient’’ causality. *Sage: Organizational Research Methods*, *19*, 10-52. <https://doi.org/10.1177/1094428115584005>

Dul, J. (2020). *Conducting Necessary Condition Analysis*. Sage Publications, ISBN: 978-1-52646-013-4.

Dul, J., van der Laan, E. & Kuik, R. (2020). A statistical significance test for necessary condition, *Sage: Organizational Research Methods Analysis*, *23*, 385-395. <https://doi.org/10.1177/1094428118795272>

Hildebrand, D, & Ott, R. L. (1998). *Statistical Thinking for Managers*, Cengage Learning 4th Edition: ISBN 13: 9780534204068

Lusk, E. & Wells, M. (2021). Vetting of Bloomberg’s ESG Governance ISS: QualityScore[GQS™]. *ABR*, *9*, 15-42. <https://doi.org/10.14738/abr.94.9952.>

Satterthwaite, F. (1946), An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110-114, https://doi.org/10.2307/3002019,

Segal, T., Mansa, J. & Reeves, M. (2021). Enron Scandal: The fall of a Wall Street darling; Online: <https://www.investopedia.com/updates/enron-scandal-summary/>

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Tamhane A. & Dunlop, D. (2000). *Statistics and data analysis*. Prentice-Hall: ISBN 0-13-744426-5.

Welch, B. (1947), The generalization of "student's" problem when several different population variances are involved. *Biometrika*, *34*, 28–35, <<https://doi.org/10.2307/2332510>>.

Welch, B. (1951). On the comparisons of several mean values: an alternative approach. Biometrica, 38, 330-336

[i] ***Acknowledgments*** Appreciation is due to: Mr. John Conners, Senior Vice President, Financial Counseling, West Coast Region, AYCO for his generous philanthropy which funded the establishment of the John and Diana Conners Finance Trading Lab at the State University of New York College at Plattsburgh and the Bloomberg Terminals that were instrumental in this research. Further thanks are due to: Prof. Dr. H. Wright, Boston University: Department of Mathematics and Statistics, the participants of the SUNY: SBE Workshop Series, in particular Prof. Dr. Kameliia Petrova: Dept. of Economics[Statistics] for their helpful comments and suggestions

[ii] The actual public law is most interesting and perusing it with the students is very beneficial. See: <https://pcaob-assets.azureedge.net/pcaob-dev/docs/default-source/about/history/documents/pdfs/sarbanes_oxley_act_of_2002.pdf?sfvrsn=e28707db_4>

[iii] See: <https://www.npr.org/2006/04/25/5361073/former-enron-chairman-blames-others-for-collapse>

[iv] In this presentation of the context, we had a preliminary sample-dataset taken from the Bloomberg Market Navigation Platform using the ESG-platform and the related ISS-links of the Wharton School. Subsequent, additional ISS datasets were downloaded from the Bloomberg-link of SUNY and used in this research report and that of: Lusk & Wells (2021).

[v] The actual text that was used was the Hildebrand and Ott (1998) text. However, this text is difficult to locate so Ott 4th edition was used for this research report. This text has most of the same problems that were used at the OVG and is more readily available and less costly.

[vi] The rule of 5s is well documented. See:<< https://www.statology.org/continuity-correction/>>.
The Rule of 10s is detailed by Tamhane & Dunlop (2000: 5.1.2[p.173])