Statistical significance testing (SST) is a hypothesis testing tool, the purpose of which is to identify universally true effects. SST's secondary and closely related objective is that of generalizing sample-based insights onto a larger population. Although principally a theory development method, significance testing has in recent years been adopted to promotional program measurement where it gained quick acceptance as the impact validation standard.

Operationally, SST utilizes any of the known distribution statistical difference tests, such as *F*, *t*, or *χ*2 to compare observed effects to expected effects with the purpose of distinguishing between spurious and persistent relationships, as shown below in Figure 8.10.

While the statistics utilized in significance testing (i.e., the above referenced *F*, *t*, or *χ*2) are themselves methodologically sound, their program measurement applications tend to outstretch their usability limits leading to misapplications and misinterpretations. Some of it is due to simple user error, but a considerable share of SSTs misuse can be attributed to fundamental lack of fit between *theory testing* and typical *business objectives*.



*Figure 8.10* Types of Statistical Significance Tests

Although rarely compared “side-by-side,” theory testing and applied knowledge creation processes differ on some very important dimensions. Perhaps most importantly, theory testing aims to uncover universally true knowledge claims, while marketing analytics focus on the

identification of sustainable competitive advantage. It follows that significance testing is used as a *sample-to-population* generalization tool for scientific theory building purposes, and as a *now-to-future* or longitudinal replicability tool for applied knowledge creation. This is a critical distinction as it gives rise to one of the more common SST application errors discussed later in this text.

Another common SST misapplication stems from its dependence on sample size. Sample size and the likelihood of detecting statistical significance are highly correlated, so much so that at a moderately large sample size even inordinately trivial differences can become statistically significant, while not being statistically significant at a smaller sample size (everything else being the same). For a variety of reasons that are not important at this moment, theory testing research typically utilizes small sample sizes leading to limited sample size distortion. The opposite, however, is true for most applied business endeavors which depend on large scale (i.e., large sample size) for business viability, resulting in a considerable sample size distortion.

Expected precision of estimates is yet another (albeit more subtle) theory testing vs. applied business knowledge-creation distinction. In short, while theory development is primarily concerned with the identification of universally true relationships and less so with the exact quantification of the magnitude of effects, business analyses are almost single-mindedly focused on quantifying program-specific incrementality. It is a matter of pragmatism: The goal of business actions, such as promotions, is to benefit a particular organization only; hence it is of little concern to business analyses if a particular relationship is not generalizable to other users. In fact, from the standpoint of a particular organization, the lack of cross-user generalizability is actually a preferred outcome.

Putting the above pieces together suggests that when applied to a large-scale database analytical initiatives, statistical significance testing is of questionable value for three key reasons: First, extremely small treated vs. control differences are likely to be found statistically significant even if their magnitude renders them practically inconsequential, which will then give rise to the previously discussed statistical vs. practical significance divergence, ultimately leading to SST misapplication. Second, significance testing does not support future replicability generalizations, which means that we cannot use the results from today's test as basis for forming expectations regarding tomorrow's rollout; again, an issue of central importance to promotional program measurement. Third, treatment attributable incrementality cannot be expressed as an exact quantity, which although not a show-stopper is still less than ideal, particularly when the range of effects encompasses both positive and negative values.

Those are not trivial differences. Significance tests are computationally relatively straightforward and highly suggestive of normative applicability limits. At the same time, the goals of the theory building and practical applications-focused analyses are oftentimes quite different. The interaction between the significance tests’ applicability limits and the different (i.e., theoretical vs. practical) applications of those tests are sufficient to question the wisdom of unqualified significance testing usage in business applications. SST's sample size dependence (i.e., the likelihood of a given relationship being deemed “statistically” significant increases as the sample size gets larger, everything else being the same), inability to support longitudinal conclusions (i.e., offering an objective quantification of the probability of future replicability of current relationships) or the basic incommensurability of scientific and business objectives (i.e., seeking universally true generalizations vs. future replicability) all highlight the dangers of blind SST reliance by business analysts.

Faced with these shortcomings of an otherwise key methodological element, analysts grew accustomed to drawing a line of demarcation between the *statistical* and practical *significance*. In effect, it has become a commonplace in applied marketing analytics to expressly differentiate between the “statistically significant results we accept” (i.e., the results that are deemed both statistical and practically significant) and the “statistically significant result we do not accept” X–Y dependence can be tested as shown below:



*Figure 8.11* Distinguishing between Correlation and Dependence