What Experiment Did We Just Do?

Counterfactual Error Statistics and Uncertainties about the Reference Class

Kent W. Staley[*]

Arkansas State University

Introduction

Among experimenters you will sometimes hear it said that one should avoid looking at one's data prior to deciding how to analyze those data. This "no peeking" rule is probably broken fairly often in practice, but many experimenters will at least pay it lip service. The question for philosophers of science is whether the rule has a sound epistemic rationale. If there is such a rationale, that is to say, if there is some epistemic damage done by peeking at experimental data prior to a determination of the testing procedure, then a question of both philosophical and practical interest is whether there is some way to repair that damage.

I aim, in the present essay, to aim to accomplish two things: First, I will argue that there is a sound rationale for the "no peeking" rule for many experimental contexts. Second, I will present a method to use in some contexts in which that rule has been violated, in order to retain the ability to draw reliable experimental conclusions from the data in question. I will do both of these things by drawing on the conceptual resources of the error statistical theory of experimental inference (Mayo 1996). Not least among the

[*] I am grateful to Douglas Allchin and Peter Lewis for helpful comments on an earlier version of this paper, and to the participants in Deborah Mayo's 1999 NEH Summer Seminar at Virginia Tech for comments on an embryonic version.

reasons for doing this is to see whether the error statistical theory can provide sensible advice on problems encountered by experimenters in practice — a necessary condition for the adequacy of any philosophical account of experimental inference. To illustrate the problem of peeking and a method by which to alleviate its negative epistemic consequences, I will discuss one aspect of the search for the top quark at the Fermi National Accelerator Laboratory by the Collider Detector at Fermilab (CDF) collaboration.[1]

I will conclude with some brief comments on the implications of the "no peeking" rule for the objectivity of the evidential relationship, in light of the error statistical theory. It will strike some as odd, given the error statistician's commitment to the objectivity of evidential relationships, that whether given data are evidence for a given hypothesis, or the extent to which they are such, might depend on what the experimenter knows, and when she knows it. Why *that* kind of information should make any difference, when the question about the evidential relationship is supposed to be about what is true, independently of what the investigator thinks, requires explanation.

II. The error-statistical model of experimental inference.

Briefly, the core evidential principles of the error statistical model of experimental inference (Mayo 1996, see especially 178–87) can be expressed in terms of the Severity Requirement (SR) and the Severity Criterion (SC):

[1] The no-peeking requirement may call to mind the "novelty" requirement, concerning which philosophers of science have produced a vast literature. For a discussion of the relationship of this issue to novelty, which also discusses the top quark search, see (Staley 1996).

SR: An experimental result E constitutes evidence in support of a hypothesis H iff:

(1) E fits H, and

(2) H passes a severe test with E.

SC: A hypothesis H passes a severe test T with outcome E just in case the probability of H passing T with an outcome such as E, given that hypothesis H is false, is very low.

These principles entail that in order to determine whether a given experimental outcome E is evidence for a given hypothesis H, one needs to address what we might call the Severity Question (SQ).

SQ: How often would a result such as this occur in an experiment such as this, assuming that the hypothesis is false?

From SR and SC, it follows that if the answer to SQ is that such a result would occur fairly often, then the hypothesis has not passed a severe test, and the result does not constitute evidence for that hypothesis.

III. Error Statistics at Work: The Search for the Top Quark

In the early 1990s, about 450 physicists comprising the Collider Detector at Fermilab (CDF) collaboration, sought evidence of the existence for the top quark. Of the six kinds of quark postulated in the “standard model” of the elementary particles and forces, the top quark was the last to be experimentally confirmed. The experiment itself was remarkably complex, and I will discuss only a small fragment of the work that went into substantiating the first claim of evidence for the top quark’s existence, published as (Abe, Albrow et al. 1994).

CDF used a detector to examine the products of proton-antiproton collisions. If the top quark did exist, then every once in a while one of several “signatures” would show up in their data — indications that a top quark had been produced and then decayed into other particles. One such signature would involve a high transverse-momentum lepton (either an electron or a muon), three or more high energy “jets” of strongly-interacting hadrons, and another electron or muon with low transverse momentum — a “soft” lepton. The search for events bearing this signature was called “soft lepton tagging” or SLT.

“High momentum,” “low momentum,” and so on are vague terms. CDF sought to make them precise in a way that would distinguish real top quark decays (“signal” events) from background processes that might mimic this top quark signature (“background” events). This problem amounted to choosing the specific values (called “cuts”) that various particle measurements must have in order to constitute a “candidate event.” Any collision event that satisfied the cuts would qualify as a top quark candidate event -- not necessarily a top quark decay event, but a candidate for being one.

Having chosen a set of cuts, CDF could then attempt to tackle the search for the top quark by collecting data, and then trying to figure out, given the amount of data they had collected, how many candidate events they expected to find from background sources alone. The existence of the top quark would manifest itself as a significant excess in the number of candidate events beyond the expected background.

What constitutes a significant excess? Quantitative error statistics can help address this question. CDF had determined for themselves a “null hypothesis”:

H: This data sample has been drawn from a population of proton-antiproton collision events that is free of top quark production.

They also had an “alternative hypothesis”:

J: This data sample has been drawn from a population of proton-antiproton collision events that contains some top quark-producing events.

Finally, they had a “test statistic”:

$X \equiv$ the number of candidate events in the present data sample

With these elements in place, they could produce a null probability distribution for X. This distribution tells one the probability of getting various values for X, assuming that H is true, and given the particular experiment being performed. As it happened, after collecting data for about a year and half from 1992 to 1993, they had data on about 16 million collision events, and found that they had seven SLT candidate events. They estimated that they should expect approximately three candidate events from background.

Given that outcome, they then sought to calculate the probability of getting seven or more candidate events, assuming the null hypothesis H. That is, they sought to calculate the statistical significance level of the results they obtained. They found that probability to be 0.041. In other words, according to their calculations, if there were no top quark, and were they to repeat their experiment infinitely many times, they would get seven or more candidate events about 4% of the time.

The value of such a statistical significance calculation, on the error statistical approach, should be fairly obvious. The severity question SQ asks how often a result such as the one actually obtained would be found, assuming the hypothesis in question were to be false. If all goes well, significance calculations enable that question to be addressed quantitatively.

IV. Bias, Tuning on the Signal, and the Reference Class

However, things do not always go so well.

The SLT search had first been used by CDF in a data-collecting period from 1988 to 1989. During that period, no evidence for the top quark was found. However, having failed to find the top quark, CDF was able to set a lower limit on its mass (Abe, Amidei et al. 1992), since theory dictated that the lower the top quark's mass, the more frequently the particle would be produced, and the more quickly it would show up in their data. As CDF prepared to begin a new round of data-collection in 1992, some discussed the possibility of changing some of the cuts used in the SLT search. Since an SLT candidate event should have a low momentum, or "soft," lepton, a choice had to be made as to where the minimum and maximum momentum cuts should be placed for soft leptons. The minimum value had been set at 2 GeV/*c*, but for a more massive top quark, some argued, the cut should be moved to 4 GeV/*c*. Leptons with momentum in the range from 2–4 GeV/*c* were, they argued, much more likely to come from background than from top quark decays, if the top quark was fairly massive.

But this argument was not absolutely conclusive, and the two physicists who had developed the SLT search algorithm thought they had good reasons to keep the soft lepton momentum cut at 2 GeV/*c* -- not least in order to maintain continuity with the earlier search. Furthermore, those two physicists worked very independently from the rest of the group, and largely kept their deliberations to themselves. Meanwhile, new data was pouring in, and was available for examination by anyone in the collaboration .

The SLT results were eventually reported with the soft lepton cut left at 2 GeV/*c*. However, some physicists in the collaboration expressed uncertainty regarding both the timing of this choice and the way in which the choice was made. Three of the seven candidate events found by the SLT analysis went away if the analysis were done with the cut moved up to 4 GeV/*c*, yielding an apparently less significant result (see **figure 1**). Some collaboration members worried that the apparent significance of the SLT results was an artifact of a manipulation (whether intentionally deceitful or not), that created the appearance of a genuine effect out of mere background. Particle physicists consider such manipulation a sufficiently serious problem to have a special name. They call it "tuning on the signal."[2]

Consider the officially quoted significance level for CDF's SLT search: 0.041. It is true, based on the assumptions CDF was making, that if the null hypothesis were true, and were one to repeat infinitely many times an experiment using the same detector, using the same cuts, collecting the same amount of data, and so on, one would get as many as seven candidate events or more only 4.1% of the time.

However, if we know that the cuts used in this case were chosen in such a way as to exaggerate the apparent significance of the results, then we have statistically relevant information about the experimental procedure used to reach these results. Specifically, the procedure followed— including now the procedure for choosing the cuts — has different error characteristics than the procedure just described, on which CDF based their significance estimate of 0.041. If experimenters know that they have tuned their cuts on the signal, then this would be the wrong reference class for calculating that probability. The reference class used in calculating a significance level is a hypothetical population of repetitions of a certain experiment. However, if it were known that the SLT cuts had been chosen specifically in order to increase the value of the test statistic, and yet the statistical significance calculation were performed without taking this information into account, then the reference class chosen would not be a homogeneous reference class (see (Salmon 1984)).[3]

It is not homogeneous because it can be further partitioned according to a statistically relevant factor that has been ignored. The statistically relevant factor that one would need to include, if the cuts had been tuned on the signal, is the method of selecting cuts.

In calculating a significance level, one first supposes that the null hypothesis is true. One then asks, suppose I were to perform an infinite sequence of repetitions of this experiment, how often would I get such a result as this? A great deal turns, however, on how *this experiment* is specified. Here are two possible ways to specify an experiment such as that performed by CDF:

[2] CDF members worried about tuning on the signal in other aspects of the experiment as well. I discuss another manifestation of the worry in (Staley 1996).

[3] The requisite notion of homogeneity here remains to be explicated fully. Salmon's concept of homogeneity, which he formulates for use in his theory of statistical explanation, is too stringent for use in cases of experimental inference. A first step, and one that is sufficient for the present discussion, is to note that a reference class A used in calculating the statistical significance of an outcome E is *not* homogenous if there is a statistically relevant factor B, under the control of the experimenter, such that p(E/A) ≠ p(E/A&B), where B was present in that instance of the experiment that resulted in E.

Experiment 1: collecting this much data, with a detector configured in this way (etc.) *using these cuts*

Experiment 2: collecting this much data, with a detector configured in this way (etc.) *using these cuts, chosen according to this method of choosing cuts*

If the cuts were tuned on the signal, then the question with *this experiment* specified according to (2) is the appropriate one, for the method of selecting cuts is statistically relevant.

The problem with experiment 2 is not that it is an inherently bad experiment. Rather, the obstacle posed by experiment 2 is that attempting to calculate a significance level for an experiment specified in that way can be practically impossible.

For Experiment 1, a reliable model of the experiment yielding a probability distribution is available -- that is the model used by CDF in their significance calculation. Such probabilistic models of the experiment are a prerequisite for significance calculations. No such reliable model is available for Experiment 2, however, when cuts have been chosen by tuning on the signal. Such a model would have to take into account information about the intensity of the experimenter's motivation to increase the value of the test statistic, the magnitude of the desired enhancement, and so on. If such models were available, tuning on the signal would not pose a problem. If you chose your cuts to maximize the value of the test statistic, you would simply need to remember to use the probability distribution for a type-2 experiment rather than a type-1 experiment.

The "no peeking" rule gets its force from the difficulty of generating such a distribution. It is not the act of peeking itself that is troublesome, it is our inability to reliably represent the effect it has on the probability of various experimental outcomes.[4] We simply cannot generate a reliable probability distribution for the experiment in which the experimenter's zeal enters into the determination of the test statistic and the probability of getting an apparently positive outcome. Observing the no peeking rule helps to secure the conditions necessary for producing reliable probabilistic models of the experimental test, which are in turn necessary for generating significance calculations. When the rule is violated, significance calculations become unavailable, and so, it would seem, do severity assessments.

But one might be able to salvage an experiment in which peeking or tuning on the signal has occurred.

V. The Method of Counterfactual Significance Calculations

Error-statistical assessments of experimental results involve three elements: the model of the hypothesis, the model of the experiment, and the model of the data (Suppes 1962; Mayo 1996, esp. ch. 5). The model of the hypothesis provides us with a probability distribution. The model of the experiment contains all of the statistically relevant information about the experimental test itself. The model of the data yields a test statistic. One important experimental strategy, emphasized by Mayo, involves holding the experimental model and the data model constant while varying the model of the hypothesis, in order to see what can be learned about various hypotheses from the testing results at hand.

[4] Mayo makes a similar point regarding the predesignation of test specifications in general in (Mayo 1996, ch. 9).

In the strategy I wish to discuss (we might call it "exploring significance space") the model of the hypothesis and the data are held constant, while the model of the experiment is varied.[5] The question that such a strategy allows one to address is this: How sensitive is my assessment of the severity with which this hypothesis passed a test to changes in the description of that test? This question becomes important when an experimenter is uncertain as to which of several distinct descriptions of the test is most accurate. The greater that uncertainty is, the more important this question becomes.

Sometimes, although one may be uncertain about which experiment one did, i.e., which reference class to use when calculating significance, one can nevertheless evaluate the experimental outcome counterfactually. On this approach, the experimenter evaluates a single set of data in the light of a number of different experiments that *might* have been done. The actual significance level of the result may remain forever unknown, but one can gain insight into just how sensitive the *apparent* significance level is to those aspects of experimental procedure about which one is uncertain.

In the method of counterfactual significance calculation one hypothetically reconstructs the experiment without any link between the choice of cuts and the data at hand. Absent such a link, other cuts might have been chosen, (within reasonable bounds -- certain choices would not be physically reasonable given the aims of the experimenter). In this reconstruction, the experimenter can address the following question: assume we

[5] The model of the data in its entirety cannot be held exactly constant under the proposed variation in the experimental model. When the experimental model changes, this changes the definition of the test statistic, which is part of the model of the data. However, the data which determine the value of the test statistic can be held constant. In the present case, holding the data model roughly constant while varying the experimental model amounts to supposing that all measurements made on collision products remain the same, but with different cuts applied to those measurements.

had not tuned our cuts on the signal (which in the case of CDF's SLT analysis may or may not have been done); what cuts might we have chosen? What, then, would we now be saying about the significance level of our results?[6]

The goal of this procedure is to determine whether one has evidence for a given hypothesis or not, and if so, to determine how strong that evidence is. This goal is pursued by attempting — qualitatively — to evaluate the severity of the test that the hypothesis has passed. When an experimenter is uncertain about what the appropriate reference class is for her experimental results, she cannot specify an accurate significance level. However, on the error statistical approach, significance levels are not ends in themselves, but a means to evaluating severity. Through counterfactual significance calculations, the experimenter may be able to gauge the severity of her test qualitatively although a quantitative determination is impossible.[7]

To illustrate this point, consider the significance calculations presented by CDF for different parts of top search results, as well as other counterfactual calculations that they

[6] Some of these calculations were in fact carried out within CDF and shown at collaboration meetings. They did not at that time become part of CDF's official presentation of the results of their top search. At least one member of CDF, Fermilab physicist Morris Binkley, proposed that the official results include just such calculations. Skeptical of the officially quoted significance levels being presented by CDF, Binkley proposed that results be shown using a variety of different choices of cuts (Binkley 1995).

[7] A similar approach can be employed where questions arise about whether a pre-determined "stopping rule" has been followed: From the data collected, sample smaller subsets of data, and calculate apparent significance levels based on those subsets. In this way, the sensitivity of the significance level to the precise stopping-point in gathering data can be evaluated.

might have presented.[8] The SLT analysis was just one of three search strategies that CDF employed, and their full results involved combining the outcomes of all three. In **table 1**, I present CDF's calculated significance levels for each of the three parts of their top search, the dilepton (DIL), secondary vertex tagging (SVX), and soft lepton tagging (SLT) searches. This table also presents my own calculations for those same values, along with calculations based on various changes that might have been made to the SLT analysis. (My calculations are based on simple Poisson statistics, using data presented by CDF in (Abe, Albrow et al. 1994) and ignoring "systematic" uncertainties. I was not able to reproduce their full statistical analysis, which involved subtleties beyond my means. Although my results are close to CDF's in the cases where they can be compared, these numbers are meant only to be suggestive of the type of strategy involved, and can not be used to draw any reliable conclusions about CDF's actual results.)

Search Used (SLT momentum cut)	no. of candidate events	expected background	CDF's significance calculation	KWS's significance calculation
1. DIL	2	0.56	0.12	0.11
2. SVX	6	2.3	0.032	0.030
3. DIL+SVX	8	2.86	—	0.0091
4. SLT(2)	7	3.1	0.041	0.039
5. SLT(4)	4	1.7	—	0.093
6. SLT(6)	3	1.1	—	0.10
7. DIL+SVX+SLT(2)	12	5.7	0.016	0.014
8. DIL+SVX+SLT(4)	9	4.3	—	0.032
9. DIL+SVX+SLT(6)	9	3.7	—	0.014

[8] The calculations presented here are based on counts of the number of candidate events observed. The significance calculations officially presented by CDF are based on a count of "tags" rather than events (Abe, Albrow et al. 1994). Significance levels for the result in terms of number of tags are much more difficult to calculate. Furthermore, because the counting of tags allows for a single event to be counted in more than one search channel, the conclusions suggested by the calculations presented here do not necessarily carry over to the case in which tags rather than events are calculated. Hence the full story of how searching significance space can shed light on the results in CDF's Evidence paper exceeds what can be presented in one short paper.

Table 1

The apparent significance of the SLT search by itself is strongly dependent on the placement of the soft lepton momentum cut (lines 4–6). Taken by itself, then, this suggests that if there are doubts as to whether the SLT cuts were tuned on the signal, then the calculated significance based on the 2 GeV/*c* cut may indeed be a poor indicator of the actual severity of the test.

Combining all three searches yields a different picture, however. The SVX search and the SLT search picked out some of the same events. Hence, the number of candidate events that were selected by at least one search algorithm does not simply equal the sum of the numbers selected by each algorithm (line 7 is not equal to the sum of lines 1, 2, and 4). None of the events selected by the SLT search that fell into the momentum region between 2 and 4 GeV/*c* were tagged by the SVX algorithm, but three of the events in the higher-momentum region were. Hence, if the result is reported in terms of number of events chosen by at least one search algorithm, the SLT contributes 4 events to the total (beyond those in the SVX+DIL sample) provided that the cut is kept at 2 GeV/*c*. However, although three of those events are lost by moving the cut to 4 GeV/*c*, there is also a significant decrease in the expected background. Hence a further increase in the cut to 6 GeV/*c*, which does not remove any events from the candidate sample, cuts out still more background, and the apparent significance is restored to what it was with the cut kept at 2 GeV/*c*.

While the apparent significance of the result based on all three search channels appears to be somewhat sensitive to the placement of the SLT momentum cut, it is not strongly sensitive, provided the result is reported in terms of the number of events selected by at least one algorithm. (In fact, this is not how CDF actually reported their results — see note 8).

If they could be taken seriously (they cannot) the above calculations would suggest that the results of the SLT present at best very weak evidence for the top quark. The reason for this is that the apparent significance of the SLT results depends strongly on where the SLT momentum cut is placed. It could easily have been chosen to take on a value that would have resulted in a much higher value for its apparent significance. For the evidence claim based on the results of all three searches, the picture is not so bleak (in light of this over-simplified analysis). Although the placement of the soft lepton momentum cut makes some difference in the apparent significance of the combined result, it is not so great as to undermine radically whatever evidence claim might be made on the basis of these results.

VI. Subjective Circumstances and Objective Evidence

Charles Sanders Peirce, an early and insightful advocate of the "no peeking" rule in experimental methodology, once wrote:

> [I]n demonstrative reasoning the conclusion follows from the existence of the objective facts laid down in the premisses; while in probable reasoning these facts in themselves do not even render the conclusion probable, but account has to be taken of various subjective circumstances — of the manner in which the premisses have been obtained, of there being no countervailing considerations,

> etc.; in short, good faith and honesty are essential to good logic in probable reasoning. (Peirce 1931-1958, 2.696)

Peirce raises a problem. The error-statistical approach to inference is supposed to yield an objective concept of evidence, and yet, as Peirce puts it, "account has to be taken of various subjective circumstances" -- such as how cuts were chosen, whether those who chose them had seen the data, how it was decided to stop collecting data, etc. Some critics have charged that the apparent relevance of such matters indicates that there are undesirable "subjective" elements in the theory of Neyman-Pearson significance testing.

My analysis suggests a rather different picture, however. These "subjective circumstances" are really relevant to the question of whether the experimenter is applying his statistical tools correctly, so that they can reliably be used to answer the questions -- such as the severity question -- for which they are employed. The wrong kind of behavior on the part of experimenters introduces an element into the experiment itself that renders the results of the standard statistical calculations unreliable. Similarly, if I am using a thermometer to measure the air temperature, but absent-mindedly leave my thumb on the bulb of the thermometer, I will get an unreliable reading. But we would not say that because facts about my "subjective circumstances" affect the functioning of the thermometer, its output is merely subjective.[9] Rather, in both cases, the experimenter must simply take care that the instrument is measuring what he thinks it is measuring.

Experimenters need to have reliable probabilistic models of their experimental tests. Some recent social and rhetorical studies of science have stressed the great deal of effort

[9] A similar point is made by Deborah Mayo regarding the fact that experimenters must exercise judgment in the pre-trial specification of a test's properties. See (Mayo 1996, 405–411).

that scientific investigators invest into making themselves “invisible” in their official reports, so that their audience will see those reports as, in a sense, nature reporting directly. Whatever may be the social, political, or rhetorical motives for this practice, it also reflects an epistemic requirement of central importance. Having a reliable model of the experimental test being performed *demands* that the mysterious workings of individual or collective experimenters’ psyches be made statistically *irrelevant*. A complete picture of the scientific enterprise must recognize this importance of this requirement, but must also recognize just how much work, how much active engagement of those very same psyches, it can take to achieve that elusive goal.

References

Abe, F., M. G. Albrow, et al. (1994). “Evidence for Top Quark Production in $\overline{p}p$ Collisions at $\sqrt{s} = 1.8$ TeV.” *Physical Review D* **50**: 2966–3026.

Abe, F., D. Amidei, et al. (1992). “Limit on the Top-Quark Mass from Proton-Antiproton Collisions at $\sqrt{s} = 1.8$ TeV.” *Physical Review D* **45**: 3921–48.

Binkley, M. (1995). Oral History Interview by K. Staley. Tape Recording. October 19, 1995. Fermilab.

Mayo, D. (1996).*Error and the Growth of Experimental Knowledge*. Chicago, University of Chicago Press.

Peirce, C. S. (1931-1958).*Collected Papers of Charles Sanders Peirce*. Edited by C. Hartshorne and P. Weiss. 8 vols. Cambridge, Mass., Harvard University Press.

Salmon, W. (1984).*Scientific Explanation and the Causal Structure of the World*. Princeton, NJ, Princeton University Press.

Staley, K. W. (1996). “Novelty, Severity, and History in the Testing of Hypotheses: The Case of the Top Quark.” *Philosophy of Science* **63** (supplement, Proceedings of the 1996 Biennial Meeting of the Philosophy of Science Association): S248–55.

Suppes, P. (1962). “Models of Data.” In *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*, edited by E. Nagel, P. Suppes and A. Tarski. Stanford, Stanford University Press: 252-61.

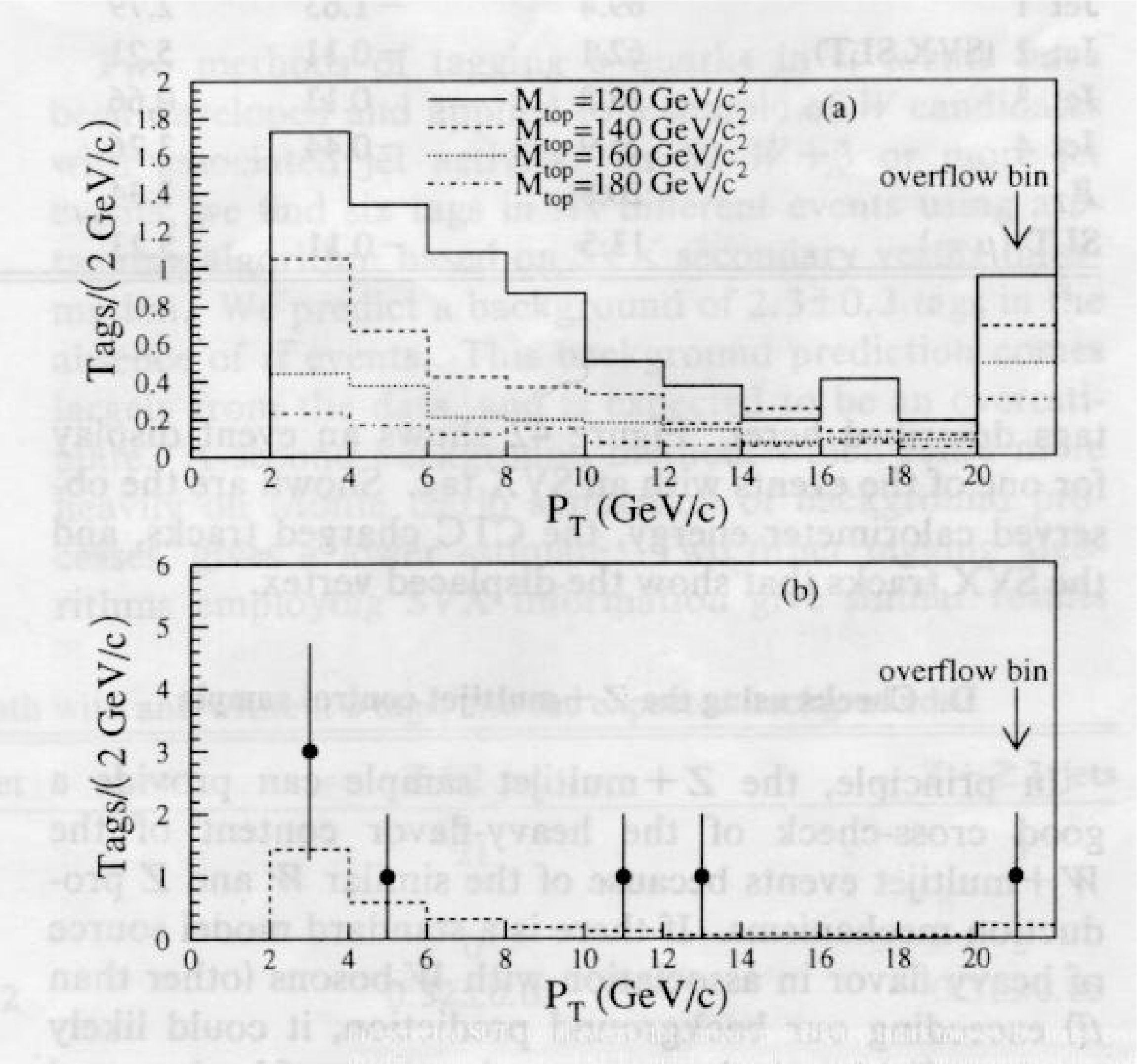

Figure 1. (a) expected SLT candidates from top quark, various possible top quark masses, by transverse momentum (P_T) of soft lepton; (b) observed (*data points*) and expected (*dotted line*) SLT candidate events, by P_T of soft lepton. (Abe 1994, 3001)