CDF/ANAL/TOP/GROUP/2370

Estimating the Statistical Significance of a Top Quark Search Counting Experiment

Tony Liss
University of Ilinois

Guillaume Unal
University of Pennsylvania

Pekka K. Sinervo
University of Toronto

December 1, 1993

Abstract

We present some of the issues relating to the statistical significance of the sample of top quark candidates from Run 1A. We present a specific method for calculating the probability that our observation could be due to a background fluctuation.

1 Classical Definition of an Upper Limit

In an earlier CDF note, the concept of a Classical (or "anti-Bayesian") confidence level and upper limit was described [1]. An experiment is performed that measures a number of events, N_o (equivalently, this could be expressed as an observed cross section by dividing N_o by the integrated luminosity of the experiment). We can interpret this measurement as either an estimate of the true event rate for a given process or to make an inference about how large or small the true rate, N^*, of the expected signal is.

There are two schools of thought on this issue: the Bayesian and the Classical (or anti-Bayesian) points-of-view. In most cases, these result in similar upper limits although they actually make different statements. We will in keeping with past practice stick to the Classical approach to setting an upper limit. We assuume that we have no a priori knowledge of the the true rate of events. However, if we interpret the observed number of events as having some contribution from our expected signal (and perhaps an expected background), we can ask the question "How large can the true signal rate be such that if I repeat the experiment I will have a given probability of observing the same signal, or a signal that has even fewer events." So for example, if we observe 1 event, we would determine the expected event rate N_e such that the probability of observing 0 or 1 events is equal to 5%. The value of N_e is called the 95% confidence level upper limit for N^*.

We are often interested in setting an upper limit on not the observed rate of events, but the cross section for the signal, which is related to n_o by

$$\sigma = \frac{n_o}{\epsilon \int \mathcal{L}\, dt}, \tag{1}$$

where ϵ is the acceptance for the signal. The procedure for handling the case when we have uncertainties on ϵ is straightforward. We want to determine the confidence level interval for the situation where if we repeat the experiment, now allowing the observed number to vary and the acceptance to vary in a way consistent with our uncertainty in it, we would observe a number of events $\leq N_o$. The likelihood function for this is defined in [1], and is given by

$$L = \tag{2}$$

2 The Significance of an Observation

We are often interested in making a statement about the "significance" of an observation. This usually means that we want to determine how likely our observation would be under a specific hypothesis. We are particularly interested in what we would call the "null hypothesis" where we assume that the observed events only arise from background sources in the absence of signal. A popular way of stating this is: Assuming that there is only a background with an expected mean rate μ, what is the probability that we would observe a result that is as significant as the one we have observed or even more significant? In a counting experiment, this means how likely are we to observe at least as many events as were observed.

If the probabiity of observing n background events is given by $P_b(n, \mu)$, then the probabiity of observing at least n_o events is given by

$$\sum n = n_o{}^{\infty} P_b(n, \mu) = 1 - \sum_{n=0}^{n_o - 1} P_b(n, \mu). \tag{3}$$

For example, the probabiity of observing 15 events with an expected background of 6 is 2×10^{-4}.

This statement is often criticised on the basis that it doesn't take into account the likelihood of the "alternative" hypothesis, which in our currrent context is the presence of a signal above the expected background. In the presence context where the signal (*i.e.*, top production) depends on the unknown top quark mass, the likelihood of our observation being consistent with top quark production is a function of mass. For very large top quark masses, where the expected signal is small compared to the expected background, the probability that we would observe a signal as large as the current event sample is essentially the same as the probability that it is due to purely background. Thus, we would have to conclude that a heavy top signal is as consistent with our observation as is the null hypothesis.

In the top search, where the expected signal rate is unknown, the fairest conclusion we could draw is one that is a function of top mass. In that case, we would want to determine the significance of the nul hypothesis relative to the likelihood that we would have observed as significant a signal due to a combination of background and top coming in at the rate predicted by the standard model.

3 Combining Information in Multiple Channels

We now turn to the case where we search for a signal in several different channels. There is usually a straightforward method of determining the probabiity (or likelihood) of making an observation consisting of n_o^i events in each of the i channels. Assuming that each observation

Figure 1: The likelihood of observing n_1 events in channel 1 and n_2 events in channel 2 for $\mu_1 = 1.5$ and $\mu_2 = 3.0$.

is statistically independent, the likelihood is simply the product of the individual likelihoods for each channel. The real problem arises when we wish to combine the information from the channels and determine a confidence level interval. This is because there is not a unique method for combining the information together to assess whether one observation is more significant than the other.

A simple example illustrates the problem. Suppose we have two channels, and we observe n_1^o and n_2^o events in each channel. Further, let's assume that we epect a background rate μ_i in each channel. The probability of observing n_1 and n_2 events assuming the null hypothesis is

$$P(n_1; \mu_1) P(n_2; \mu_2) \tag{4}$$

where the function $P(n; \mu)$ is the Poisson probability distribution. This is shown plotted in Fig. 1 as a function of n_1 and n_2.

In order to determine the likelihood of the null hypothesis, we have to construct a criterion that labels a given outcome more consistent with the null hypothesis or less consistent. One natural choice is to classify the outcomes based on their relative probabilities–consider all outcomes with a lower probability than the observation as being less consistent with the null hypothesis than the outcome we did observe. We would then sum up the probabilities of all outcomes that are less likey than $P(n_1^o; \mu_1) P(n_2^o; \mu_2)$ and interpret it as the probability that the expected backgrounds could fluctuate to produce a signal at least as significant as the one we observed.

There are several problems inherent with combining several data together that this example does not ilustrate. For example, if we had one very unlikely outcome in one channel (*e.g.*, observing 10 when we expect only 1), then this would make our observation very significant regardless of what we observed in the other channel. In some sense, it tends to give less reliable information because it can be skewed by one very improbable measurement even when we have other information that is inconsistent with the "outlying" observation. It also is more difficult to generalise to cases where an analytic expression for the likelihood is not available and Monte Carlo techniques have to be used.

For these reasons, we concluded (as did others before us) that a reasonably robust, moderately conservative technique to combine data from several channels was to simply sum up the total number of observed events and the total expected background and base confidence interval estimation on the total rate. There are of course other techniques that one could in principle use, but this technique is generally accepted.

4 Estimating the Significance of our Top Result

We now want to specifically address the question of how to estimate the significance of our recent observation of dilepton events and W events with b-quark candidates tagged using either the SVX (SVX tags) or the soft lepton algorithm (SLT tags). We will for the purpose of this note assume that we observe 6 SVX tagged events, 7 SLT tagged events and 2 dilepton events. In addition, We will consider the fact that we observe 3 events that are tagged by both SVX and SLT algorithms and one dilepton event that also has an SVX tag. We will for the purpose of illustration, also assume that the estimated background rates are 1.70 ± 0.41, 2.94 ± 0.51 and 0.61 ± 0.13, for the SVX, SLT and DIL channels, respectively. These are, in fact, similar to the preliminary values for these quantities, but we point out that these probably differ from the accepted values now.

4.1 The Simple Approach Ignoring Systematic Uncertainties

We first calcuate the significance of observing the signals in each channel by summing up the Poisson probabilities of observing at least as many events as have been observed in each channel using the formula in Eq. 3. We ignore the systematic uncertainties on the background rates to simpify the discussion. The resulting confidence levels are 0.8%, 3.1%, and 12.5% for the SVX, SLT and DIL channels.

We now wish to combine these data together to calculate the combined significance of the result. One way to do it is to sum up the number of tags (15) and the number of background events expected (5.25 ± 0.67), and calculate the significance of observing at least 15 tags given the expected backgrounds. Ignoring the uncertainty in the background estimates, we obtain a significance of 0.037%.

This calculation can be criticised [2] on the basis that it ignores the potential correlations in the backgrounds contributing to the various tagged samples. There are several ways of addressing this problem. Perhaps the most conservative method is to simply count events instead of b-quark tags. The probability of observing 12 events or more given an expected background of 5.25 is 0.78%, which implies the observation is marginally significant.

One method, described by George Brandenburg, considers the number of double-tagged events a special category that he treats as an independent statistic. He then forms a likelihood for observing the number of single-tagged events in each category and the number of double-tagged events, using an estimate for the number of double-tagged events. His likelihood is

$$P(n_1; \mu_1) P(n_2; \mu_2) P(n_3; \mu_3) P(n_4; \mu_4) \tag{5}$$

where n_4 and μ_4 are the observed and expected number of double tagged events. This likelihood suffers from the problems mentioned above (and in [1]).

5 Counting Tags Instead of Events

In order to realistically estimate the significance of our observation, it has been pointed out by several people that the correct procedure is to correctly include the correlations between the background events tagged by the different tagging algorithms.[3]

This can be done by analyzing the source of the expected background contributions and either estimating or measuring the probability of having the same event tagged by more than one algorithm. The backgrounds are broken into the physics backgrounds (dominated by $Wb\bar{b}$, $Wc\bar{c}$, and Wc), fake backgrounds due to mistagging, and miscellaneous other backgrounds. The correlations in the tag rates of the physics backgrounds are particulary important as each tagging algorithm is reasonably efficient on this class of background. The correlations in the fake backgrounds are perhaps less important because the probability that two algorithms mistag the same event is given by the product of their rejection powers.

The significance of our top observation is calculated using a Monte Carlo method where a large number of "background" experiments are performed. In each experiment, a number of physics background events are generated using a Poisson distribution and the efficiencies of each of the tagging algorithms are applied to determine how many of these events are tagged and double-tagged. The number of background events due to mistagging are then drawn at random, based on the mistag rate per jet as measured independently and applied to the data sample. Finally, the remaining miscellaneous background contributions are included by sampling a Poisson distribution. Correlations between the mistag rates for the SVX and SLT algorithms are estimated and included in the MC calculation. The various efficiencies that are used have uncertainties and these are incorporated into the Monte Caro by smearing the parameters used to determine the means of the Poisson distributions and efficiencies by Gaussian distributions.

It is important to note that the rate of single and double tags from the physics backgrounds is dominated by the fluctuations in the number of produced W + heavy flavour events. The MC calculations therefore first determine how many of these events were produced, and the efficiency of each b-tagging algorithm is then applied to these produced events. The efficiencies of the tagging algorithms are assumed to be independent on these "physics" background events, as suggested by the top MC simulations.

Each of us have written Monte Carlos that perform this calculation with slighty different assumptions about the correlations and different input parameters. These MC's are constrained by being required to produce the average rate of tags from the different physics backgrounds. There is therefore some degree of "tuning" that has been done in order to ensure that the MC calculations produce consistent results, but the results of the MC's in terms of the estimated significance of the observations in the three top candidate channels is similar.

5.1 The Physics Backgrounds

The physics backgrounds prior to tagging have been estimated using the observed SVX tagging rates and subtracting what is believed to be the non-heavy flavour background by using the negative sign DL distribution. The SVX and SLT tagging efficiencies for the different channels has been estimated by Guillaume Unal, Avi Yagill and Claudio Campagnari using MC and detector simulation. These estimates are listed in Table 1 and Table 2.

It is important to note that athough the physics backgrounds are relatively small (~ 0.85 and ~ 0.90 events for SVX and SLT algorithms, respectively), they have a significant effect on the likelihood of observing events that are double-tagged.

	I	II	III
N(Wbb)	3.62 ± 0.43	3.90 ± 1.17	2.63 ± 1.25
N($Wc\bar{c}$)	5.92 ± 0.72	5.75 ± 1.72	4.16 ± 1.24
N(Wc)	4.16 ± 1.24	4.3 ± 1.3	4.16 ± 1.24

Table 1: The observed event rates (before tagging) for the physics backgrounds. The first two columns are based on different methods for estimating the physics rate. The third column is an earlier variant that is used to demonstrate the sensitivity of the calculation.

	SVX	SLT
Wbb	0.200 ± 0.040	0.150 ± 0.050
$Wc\bar{c}$	0.040 ± 0.016	0.052 ± 0.020
Wc	0.035 ± 0.012	0.020 ± 0.004

Table 2: The efficiencies of tagging the different physical backgrounds for the two different b-tagging algorithms. These efficiencies have been determined from MC studies (?).

5.2 Fake Backgrounds

The other significance source of tagged events in the absence of top production are mistagged events, or what are usually considered "fakes." The rate of fake tags has been measured using the negative decay length distribution for the SVX and the fake rate measurements using jets for the SLT tag. As a representative case, we have chosen values of 0.73 ± 0.10 events and 1.75 ± 0.21 events for the SVX and SLT algorithms, respectively.

One might expect that potential correlations between the probabilities for mistagging fakes using the two algorithms could have a significant effect on the overall rate of single and double tagged events from these sources. There are some indications that the probability of mistagging is in fact correlated. In order to model these correlations, our MC calculation starts with the number of $W+\geq 3$ jet events (52) before tagging and randomly mistags each event with the jet mistagging rate estimated from the fake studies. We have made various assumptions concerning the correlation in the mistagging rates, ranging from assuming no correation to assuming that the probability of double-tagging a fake event was given by twice the product of the separate mistag probabiities.

We have also modelled the generation of fake tags in two different ways. We have first treated the fake backgrounds as being totally independent and have also included in the fake rates the additional miscellaneous physics backgrounds (0.13 ± 0.05 events and 0.33 ± 0.12 events for SVX and SLT algorithms, respectively). We have also varied the MC calculation to allow the events from the physics backgrounds to be mistagged at the same rate as the untagged $W+\geq 3$ jet sample. Allthough this is a conservative assumption, it illustrates the sensitivity of the calculation to the assumptions we have made.

5.3 Systematic Uncertainties

We included the effects of systematic uncertainties by smearing the nominal values according to Gaussian distributions given by the stated uncertainties ever time we generate a new event. Due to the fact that we are dealing with small statistics, the effect of this smearing is rather modest.

	I	II	III
SVX Only ($\geq$ 6 tags)	1.5		1.17
SVX Only (no smearing)			1.00
SLT Only ($\geq$ 7 tags)	3.57		3.59
SVX-SLT	0.29		0.38
SVX-SLT-DIL	0.12	0.11	0.17
SVX-ST-DIL (reduce fake rates by 10%)			0.09
SVX-SLT-DIL (no correlations)	0.08		

Table 3: The probabilities (in %) of observing at least as many tagged events as in the Run 1A analysis under various conditions and for different choices of correlations.

6 Significance Results

The results of the significance calculation are listed in Table 3. We show in this table the probability that the background model would result in a fluctuation of observed tags as large as or greater than our observation.

These MC calculations also provide an estimate of the number of double-tagged events we could expect from these background sources. The average number of double-tagged events expected for all events is 0.17 (this is from MC I). For those events that fluctuate to give 13 or more SVX-SLT tags, the average number is 1.17 double-tagged events, which is comparable to what we would expect if top quark production was the source of all the tags. The fact that the background and signal give similar numbers of double-tagged events implies that we cannot use this feature of the event sample to improve the significance of our observation.

We have also investigated how a potential correlation between the SVX and SLT tagging efficiency would effect the significance of our observation and the expected number of double-tagged events. We assumed that the largest possible value for the combined SVX-SLT tag efficiency is $3\epsilon_{SVX}\epsilon_{SLT}$, which in some sense is extremely conservative, as the double-tagging efficiency would then be within a factor of two or three of the single-tagging efficiencies (for example, there is no support in the data for such a large correlation). The overall significance changes from 0.11% to 0.19%, which is comparable to the changes due to other uncertainties in our background model. We would predict an average of ~ 2.3 double-tagged events in any sample with 13 or more tags, as we would expect given the larger double-tag efficiency.

7 Conclusion

We have assessed the likelihood of observing at least as many tagged events from known background sources as we have detected in the Run 1A dataset. We estimate this significance at $\sim 0.1\%$, although this can be as large as 0.2% when we vary our model of background correlations. We beieve that a conservative conclusion is that the probability of our observation arising from a background fluctuation is $< 0.5\%$.

References

[1] "The Evaluation of Upper Limits for Top Quark Production Using Combined Measurements," R. J. Hollebeek, P. K. Sinervo and H. H. Williams, CDF Note 1109, January 1990.

[2] George Brandenburg, private communication.

[3] This approach was suggested by Brig Williams, Guillaume Unal and others.