Measurement of Top Quark Properties Using Public Data from CMS

Young-Woo Choe[1,2]

[1]Dept. of Physics, Yonsei Univ.
[2]SCC, Yonsei Univ.

E-mail: `ywchoe92@gmail.com`

Abstract. We study basic data analysis methods in High Energy Physics experiments through the measurement of the cross section $t\bar{t}$ at $\sqrt{s} = 7$ TeV and the top quark mass. We use the public data and MC simulation samples released by CMS Collaboration. A brief review of top quark physics and CMS experiment is presented, and the analysis on the semi-leptonic decay of $t\bar{t}$ with muons is performed. We report the cross section of $\sigma[pp \to t\bar{t}] = 194.6 \pm 25$ pb, and $m_t = 165.5 \pm 4.2, 163.9 \pm 3.8$ GeV for each hadronically and leptonically decaying top.

Acknowledgements

This study could not have been possible without the effort of the International Particle Physics Outreach Group(IPPOG) at CERN that disclosed their data for an educational purpose. And we would like to thank Prof. Y. J. Kwon for reviewing and commenting on this article.

1. Introduction

Our goal is to measure the cross section and mass of top quark by analyzing the semi-leptonic decay channel of top quark pair production. We use the public data and MC samples released by CMS Collaboration at CERN[1].

Natural Unit In this article, we use the natural unit as usual in High Energy Physics.

$$\hbar = c = 1$$

In this unit, length and time has the same dimension (usually in GeV), and energy and mass has the inverse dimension of length GeV^{-1}.

First, we review the commonly used quantities and variables in HEP. And the coordinate system for HEP anaylsis is introduced.

1.1. Cross Section, Luminosity and Efficiency

The total number of detected events is related to the total cross section σ of process and the integrated luminosity L, where integrated luminosity is the time-integrated instant luminosity $\mathcal{L}$.

$$N \;=\; \sigma \cdot L, \qquad where\ L = \int dt\,\mathcal{L}$$

But, in real cases, we need to take into account a few more factors. The first factor is the efficiency ϵ of event selection cut. We do not count all the signal events if we apply our selection cuts in order to reject background events. And the trigger efficiency also contributes to N. In addition, we need to subtract the estimate number of background events from N. The expected number of background events is estimated using simulation samples. With these considerations, we can measure the total cross section as the following:

$$\sigma \;=\; \frac{N^{obs}_{data} - N^{est}_{BG}}{\epsilon \cdot L}$$

1.2. Detector Coordinate System

In HEP experiments, z-axis of the coordinate system is defined as parallel to the beam line with the origin at the interaction point. And the x-axis is on the horizontal plane, and y-axis on the vertical. The polar angle θ starts on the z-axis and ϕ on x-axis, as the usual spherical coordinates.

There are some other commonly used variables:

- Transverse momentum
$$p_T = \sqrt{p_x^2 + p_y^2}$$
The transverse component of the momentum. This is preferred because arbitrary boost on the z-axis does not influence on the transverse momentum, which is called boost invariance.
- Pseudorapidity
$$\eta = -\ln\left(\tan\frac{\theta}{2}\right)$$
Pseudorapidity is used as an alternatvie to the polar angle θ. For massless particles or in an extreme relativistic condition, it is equal to rapidity, which is boost invariant.
- Transverse energy
$$E_T = E \cdot \sin\theta = \frac{E}{\sinh\eta}$$
It is also boost invariant.

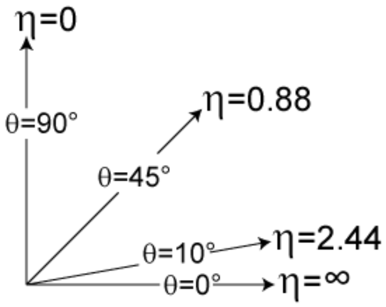

Figure 1. Some values of η depending on θ. Most of the central region of detector is covered by the region $|\eta| \leq 2.4$ [7]

2. Signal and Backgrounds

We will analyze the semi-leptonic decay channel of top pair production.

$$t\bar{t} \to W^+W^-b\bar{b} \to b\bar{b}q\bar{q}'l\nu$$

This process is illustrated in Figure 2. The semi-leptonic channel decay modes are relatively easy to analyze because the branching fractions add up to more than 40 percent, and there is a distinctive signature in the final state with a single isolated lepton of high p_T.

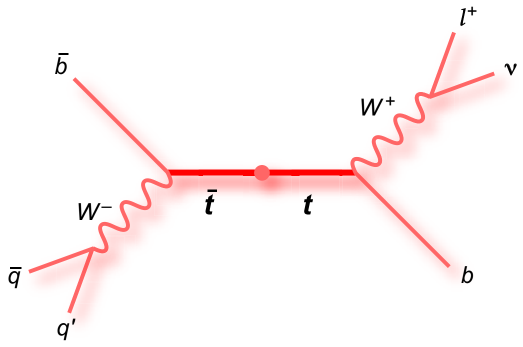

Figure 2. The Feynman diagram for the semi-leptonic decay of $t\bar{t}$ [5]

The hadronic channel($t\bar{t} \to b\bar{b}q\bar{q}'q''\bar{q}'''$) is the statistically most important, but has no lepton signature which makes it difficult to distinguish it from other hadronic backgrounds. The dileptonic channel($t\bar{t} \to b\bar{b}l\nu l'\nu'$) has too small branching fraction and more than one neutrinos would bring additional complication in analysis. Therefore, we will set the semi-leptonic channel to our signal, and treat other $t\bar{t}$ channels as backgrounds. In particular, we will restrict the range of our anaylsis to the semi-leptonic channel with muons because CMS pubic samples are only suitable for such channel. We will explain more about this soon.

There are many other background processes that have similar signatures with $t\bar{t}$ production. Among these, the dominant processes are W+jets production, QCD multijet production, and Drell-Yan process. We will apply several selection cuts to distinguish our signal from these backgrounds.

Data and Monte Carlo Simulation Samples

Since the collision rate in LHC is extremly high, which is about 20MHz, we cannot store all the data. Therefore, trigger system is used to select only the meaningful events. In our study, the data is taken through the muon trigger system that is triggered by single isolated muons with $p_T > 24$ GeV. In general, reconstruction quality of trigger system is inferior than offline reconstruction. So, it is possible that trigger is fired because it found a muon, but this object do not pass the quality cuts during offline reconstruction[9]. Therefore, we should also care about the trigger efficiency.

In order to examine properties of each process before we apply our selection cuts to the real data, simluation samples generated by Monte Carlo simulations are exploited. In this article, we use the simulation samples from 5 different processes whose details are shown in Table 1.

Name	N_{Events}	σ (pb)	L (pb^{-1})	Triggered Samples Only
data	508,561		50	yes
$t\bar{t}$	383,167	165	100	no
W+jets	76,196	31,300	100	yes
QCD multijets	137	10^8	1000	yes
Drell-Yan	109,656	2,475	100	yes

Table 1. Data and simulated Monte Carlo Samples

The last column of the table indicates whether the sample contains also events which did not pass the trigger selection. Naturally, real data contains triggered events only. For simulation, only the $t\bar{t}$ sample also contains non-triggered events, which means it contains not only detector level variables, but also event generator level variables.

3. Event Selection

3.1. Muon Selection

The semi-leptonic channel with muons is characterized by an isolated single muon with high p_T. Generally, whether a particle is isolated or not should be determined by the particle track variables, but we have a precalculated parameter in our dataset that can be used as a criteria to determine the isolation of particle. And we will also restrict the range of $|\eta| \leq 2.4$, because we want to deal with top quarks decaying into the central region. This range covers almost all the anglular distribution of the events except for small parts that decay into the front region of the detector.

The events with the following properties are selected through this cut:

- Contains exactly one isolated muon with $p_T \geq 30$ GeV
- $|\eta| \leq 2.4$

The efficiency of this selection is presented in Table 2. Note that the signal efficiency is relatively low in exchange for the strong reduction in QCD background.

Name	Before Cut	After Cut	Efficiency
$t\bar{t}$	9518.8	661.142	0.070
W+jets	273507	197774	0.723
QCD	78626.2	3955.08	0.050
Drell-Yan	26991.5	6204.7	0.230

Table 2. Efficiency of Muon Selection

3.2. Jet Selection

The topology of the semi-leptonic event shows us that there should be at least 4 jets in the final states. Since the mass of top quark is heavy, we require these jets to have high p_T.

Our jet selection is based on the following cut:

- at least 4 jets which have $p_T \geq 30$ GeV
- $|\eta| \leq 2.4$

And the efficiency for each simulation sample is shown in Table 3.

Name	Before Cut	After Cut	Efficiency
$t\bar{t}$	9518.8	7528.95	0.791
W+jets	273507	1957.38	0.007
QCD	78626.2	414.078	0.005
Drell-Yan	26991.5	826.102	0.031

Table 3. Efficiency of Jet Selection

3.3. B-tagged Jet Selection

After these two preselections, we apply one more restriction on our cut. Among selected 4 jets, two of them should be b-tagged jets. However, it turned out that if we require at least two b-tagged jets, the selection efficiency drops significantly. Therefore, we loosely take our cut to require $N_{bjet} > 1$.

3.4. Summary

In summary, our selection cut consists of the following conditions:

- One isolated muon with $p_T \geq 30$ GeV, with $|\eta| \leq 2.4$
- At least 4 jets which have $p_T \geq 30$ GeV, with $|\eta| \leq 2.4$
- At least one b-tagged jets

The application of this cut to MC samples leads to the result shown in Table 4.

Parameter	Formula	Value
$N_{sig,gen}$	A	9518.8
$N_{sig,selected}$	B	86.656
$N_{sig,sel+trig}$	C	76.833
ϵ^{acc}	B/A	0.009
ϵ^{trig}	C/B	0.887
$N_{BG,sel+trig}$	D	8.294
purity (π)	$C/(C+D)$	0.903

Table 4. Event selection result for MC samples.

4. Measurement

In this section, we apply our selection cut to the data, measure the total cross section of $t\bar{t}$ production, and reconstruct the invariant mass distribution of top quark. Table 5 shows the selection result for the data.

Parameter	Value
N^{obs}_{data}	87
$N^{obs}_{data} - N^{est}_{BG}$	78.524

Table 5. Event selection result for the data.

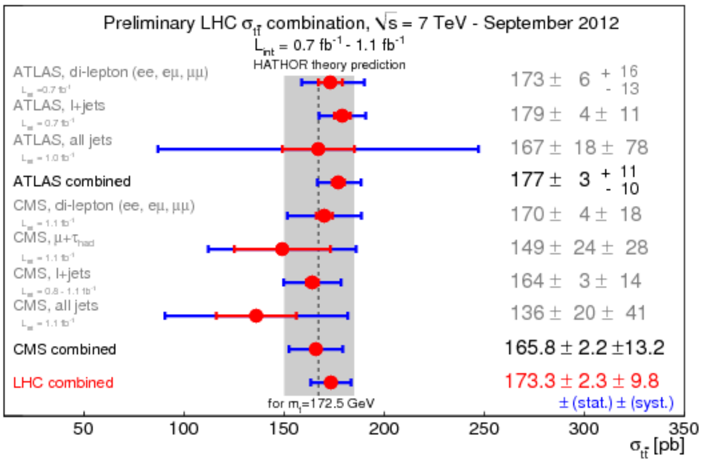

Figure 3. The official measurements of $t\bar{t}$ cross section at $\sqrt{s} = 7$ TeV from CMS and ATLAS Collaborations[8].

4.1. Cross Section

Using the cross section formula discussed in (1.1), we measure the cross section of $t\bar{t}$ production in p-p collision at $\sqrt{s} = 7$ GeV.

$$\sigma_{t\bar{t}} = \frac{N_{data}^{obs} - N_{BG}^{est}}{\epsilon^{acc} \cdot \epsilon^{trig} \cdot L} = \frac{N_{data}^{obs} \cdot \pi}{\epsilon^{acc} \cdot \epsilon^{trig} \cdot L} = 194.6 \pm 25\, pb$$

Error Analysis We estimate the statistical error of cross section as (See Appendix A for the calculation of each term)

$$\frac{\delta\sigma_{t\bar{t}}}{\sigma_{t\bar{t}}} = \sqrt{\left(\frac{\delta\epsilon^{acc}}{\epsilon^{acc}}\right)^2 + \left(\frac{\delta\epsilon^{trig}}{\epsilon^{trig}}\right)^2 + \left(\frac{\delta\pi}{\pi}\right)^2} = \sqrt{0.011 + 0.001 + 0.004} = 0.13$$

4.2. Top Mass Reconstruction

In the semi-leptonic channel with muons, one top quark decays into 3 jets including one b-tagged jet, and the other into one muon, one neutrino, and one b-tagged jet. We exploit these kinematic constraints to reconstruct top quark. We can also impose additional constraint from W boson mass.

MC Truth Before working on the data, we use the generator level information of $t\bar{t}$ MC samples to estimate the actual mass of top quark. Figure 4 shows the peak in 172-173 GeV. Therefore, we expect the data to show the peaks in the simliar region. Our reconstruction strategy goes like the following:

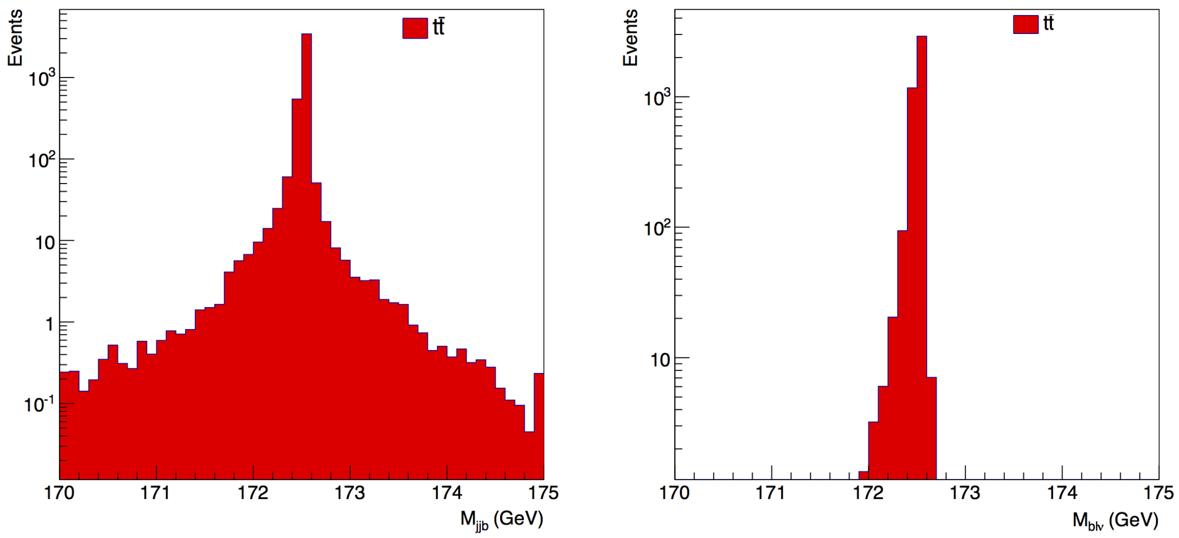

Figure 4. The reconstruction of top mass using the generator level MC truth information. The left one is hadronically decaying top, and the right leptonically.

Hadronic top Find two jets which are not b-jet and have the invariant mass within $|M_{jj} - M_W| < 15$ GeV. We regard these jets as the products of W boson decay. Then, combine these two jets with one of the b-tagged jets. Since there are more than one b-tagged jets, we choose the closest one in terms of the angular distance. The resulting combination is the reconstruction of hadronically decaying top quark.

Leptonic top Due to the conservation laws, sum of transverse momentum of all particles in an event should be 0. But, if there is neutrino in an event, it escapes the detector leaving no trail. Therefore, we indirectly measure its momentum in terms of missing energy. Still, we do not have information on z component of missing energy because only x and y components are measured.

In the semi-leptonic channel, we can put W mass constraint on a neutrino and a lepton. It enables us to determine the whole fourvector of a neutrino.

$$M_W^2 = \{E_\mu + E_\nu\}^2 - \{(p_{\mu\,x} + p_{\nu\,x})^2 + (p_{\mu\,y} + p_{\nu\,y})^2 + (p_{\mu\,z} + p_{\nu\,z})^2\}$$
$$E_\nu^2 = p_{\nu\,x}^2 + p_{\nu\,y}^2 + p_{\nu\,z}^2$$

This constraint is quadratic in $p_{\nu\,z}$, which results in two possible solutions. We take both solutions into our reconstruction. Similar to the hadronic top quark reconstruction, we combine one muon, one neutrino, and one b-tagged jet into leptonically decaying top quark.

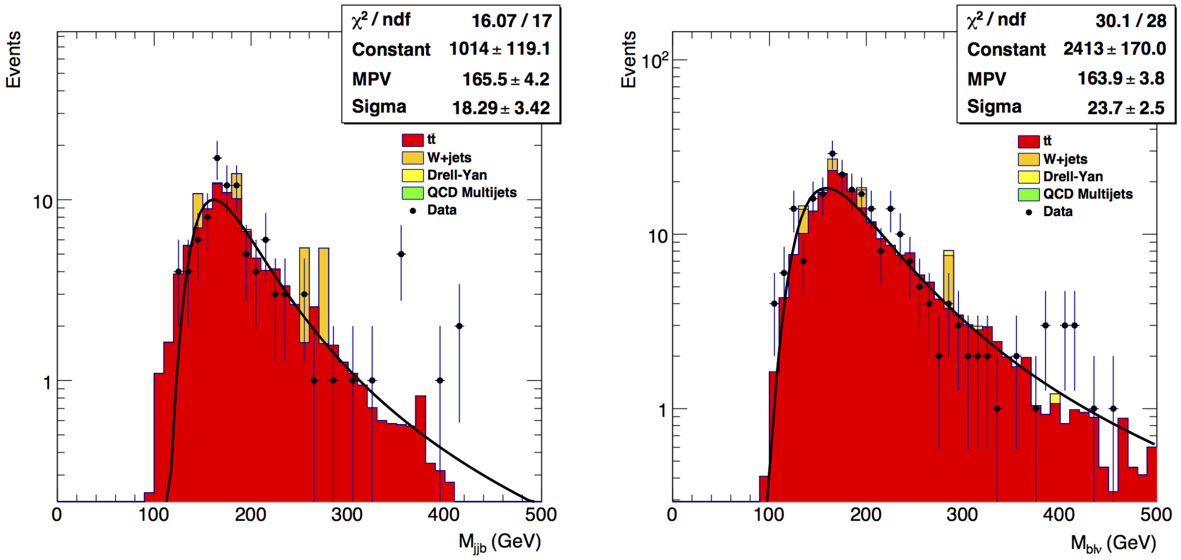

Figure 5. Reconstructed mass distribution. The left figure is the distribution of hadronically decaying top quark and the other is of leptonically decaying one.

The reconstruted mass distributions are shown in Figure 5. Our result shows much lower value than the official one. The current official measurment[2] is 173.21±0.51±0.71 GeV.

5. Conclusion

We studied basic data analysis methods in HEP by studying a pedagogical example: measurment of top quark properties. We employed a simple cut-based method to reject backgrounds. We examined our MC samples, and optimized thresholds for various cuts. As a result, we report our measurment of the total cross section of $t\bar{t}$ production at $\sqrt{s} = 7$ TeV as 194.6 ± 25 pb, and mass of top quark as 165.5 ± 4.2 GeV for hadronically decaying top, and 163.9 ± 3.8 GeV for leptonically decaying top.

In fact, there are a lot of well-known methods to improve the accuracy of measurement for this case. For example, Roseman[6] exploited some useful event selection cuts. He used the centrality C, a kind of the event shape variables, to better distinguish the $t\bar{t}$ signal from W+jets backgrounds, and kinematic fitting. Also, neural network selection, one of the machine learning methods, can be utilized as an alternative approach to cut-based event selection. However, we reserve our interest to further dig into the study of such methods for later studies, as the goal of this study is to introduce the basic data analysis methods and the overall process of research.

Appendix A : Statistical Uncertainty

Purity

$$\pi = \frac{N_{sig}}{N_{sig} + N_{bg}} = \frac{\sum_i W_i}{\sum_i W_i + \sum_j W_j}$$

$$\begin{aligned}(\delta N)^2 &= \sum_i (\delta W_i)^2 = \sum_i W_i^2 \\ (\delta \pi)^2 &= \left(\frac{\partial \pi}{\partial N_{sig}}\right)^2 (\delta N_{sig})^2 + \left(\frac{\partial \pi}{\partial N_{bg}}\right)^2 (\delta N_{bg})^2 \\ &= \left(\frac{N_{bg}}{(N_{sig} + N_{bg})^2}\right)^2 \sum_i W_i^2 + \left(\frac{-N_{sig}}{(N_{sig} + N_{bg})^2}\right)^2 \sum_j W_j^2\end{aligned}$$

Efficiency We assume the selection efficiency follows binomial model, where the estimator of standard deviation is

$$\delta \hat{\epsilon} = \sqrt{\hat{\epsilon}(1 - \hat{\epsilon})/N}$$

References

[1] Data and MC Samples used in this article can be found at CMS HEP Tutorial page. URL: http://ippog.web.cern.ch/resources/2012/cms-hep-tutorial

[2] K.A. Olive et al. (Particle Data Group), Chin. Phys. C, 38, 090001 (2014).

[3] F. Abe et al. Observation of Top Quark Production in pp? Collisions with the Collider Detector at Fermilab. Phys. Rev. Lett., 74:2626?2631, 1995.

[4] S. Abachi et al. Observation of the Top Quark. Phys. Rev. Lett., 74:2632?2637, 1995.

[5] Useful Diagrams of Top Signals and Backgrounds. URL: http://www-d0.fnal.gov/Run2Physics/top/top_public_web_pages/top_feynman_diagrams.html

[6] Rosemann, C. Measurement of top quark properties from pair production and decay with the CMS detector. (2008). doi:10.3204/DESY-THESIS-2009-006

[7] Pseudorapidity, Wikipedia. URL: http://en.wikipedia.org/wiki/Pseudorapidity

[8] CMS Collaboration [CMS Collaboration], CMS-PAS-TOP-12-003.

[9] Christian Oliver Sander and Alexander Schmidt provided a good comment on trigger system.