Investigation of College Tution Prices

Raphael Cherney, Franklin W. Olin College of Engineering

I. Introduction

The problem is not new and has been discussed many times before: college is getting more expensive. Tuition prices increase every year by more than inflation, and nothing appears to be slowing it down. The educational system in the United States may be the envy of the rest of the world, but it may soon become simply unaffordable.[1] This paper broadly investigates the following question*: Are college tuition prices really spiraling out of control?* Along those lines, it will look into the current cost of a college education and several of the factors that play into it. Hopefully by the end of this investigation you will have a better idea of how the tuition expenses have evolved over the past several decades and where we are today.

II. Sources

As one would expect, much data has been collected on the subject of college expenses. This paper is largely based on data from the Integrated Postsecondary Education Data System (IPEDS). IPEDS is a system of interrelated surveys conducted annually by the National Center for Education Statistics (NCES). The government requires all colleges, universities, and technical and vocational institutions that participate in federal student financial aid programs to complete the IPEDS surveys. This includes 7316 different institutions for the 2009 dataset. The data was downloaded from the IPEDS Data Center[2] and analyzed using custom Python statistical analysis tools.

III. Current Costs

The question seems almost trivial: What does college cost today? However, the answer is not as simple as it may first appear. If we simply look at the 4188 schools with tuition listed in the IPEDS dataset, we find that published tuition and fees range from zero to $45,800 with an mean of $12,016. However, these numbers alone are incomplete and even a bit misleading. Not all schools are created equal: land grant universities have a much lower sticker price than private Ivy League schools. As we might expect with such a varied pool of values, the standard deviation is quite high: $9,585 - approaching the value of the mean. This is not necessarily unusual for a distribution that is dependent on factors ranging from geography to type of school; there is a lot of variability. In fact, it warns us to keep our eyes open for differentiating factors.

[1] *College May Become Unaffordable for Most in U.S.*. New York Times. December 3, 2008.
<http://www.nytimes.com/2008/12/03/education/03college.html>

[2] <http://nces.ed.gov/ipeds/datacenter/DataFiles.aspx>

It is quite difficult to understand what is going on by simply calculating simple summary statistics like the mean tuition. It is much more valuable to observe the whole distribution and look for interesting effects. To do this, we can create a probability mass function (PMF) of tuition as shown in Figure 1 (grouped into $1000 bins). This plot shows the probability that a random college will have a given tuition.

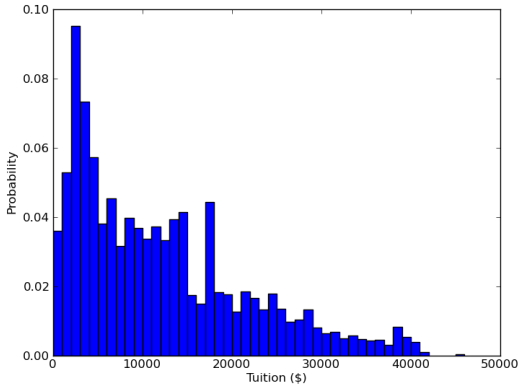

Figure 1. PMF of 2009-10 tuition for all respondent colleges.

While the results are still difficult to gather all that much meaning from, there are some interesting things to note: to begin, for all of the talk of exorbitant tuition, it appears that a majority of colleges (50.7%) actually have tuition and fees under $10,000. This would initially seem like a good sign, but again can be misleading. We must dig deeper into the data. In this case, we see that this PMF includes data from all different types of colleges ranging from private for-profit institutions with under-2-year programs (21.8% of respondents) to public institutions with 4-year and above degrees (9.5% of respondents). There is obviously a disparity between these different groups in both cost and mission.

IV. Enrollment Differences

As suggested in the previous section, there are many differences between the schools listed in the IPEDS dataset. Consider, for example, the enrollment of each institution – not all of these schools are the same size. Figure 2 shows the PMF and CDF of undergraduate enrollment values for all of the colleges in the IPEDS dataset. Unsurprisingly, there are many more small schools than large institutions. In fact, the distribution looks very much like an exponential decay. We can quickly test to see how well model matches the distribution by plotting the complementary CDF on a log-y scale. If the data fits an exponential distribution, we would expect a straight line with a slope of λ (where λ is the parameter of

an exponential distribution that determines the shape). Figure 3 shows this test applied to the undergraduate enrollment data. We see that the data fits an exponential distribution very well; the result is very linear, especially after the small enrollment values (< 2,000 students). When we do a linear least squares fit to the data and look at the slope, we find that the $\lambda \approx 1.35e-4$ for this distribution. An exponential relationship seems reasonable in this case and suggests that there are many more small educational institutions that we normally might overlook. Many of these schools may cater to specialized markets, but they are a part of the educational fabric of the country.

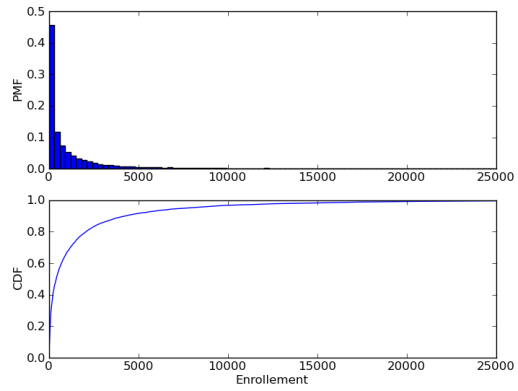

Figure 2. PMF and CDF for enrollments for all colleges in the IPEDS dataset.

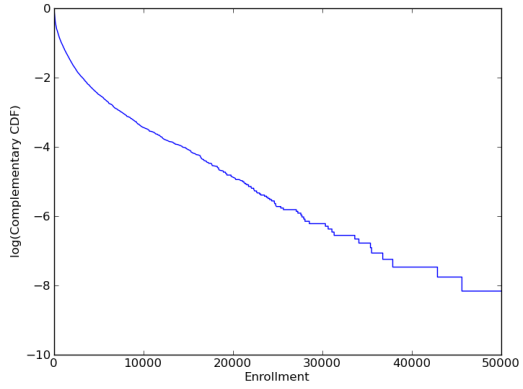

Figure 3. Complementary CDF of enrollment on a log-y scale to show exponential relationship.

V. Different Institution Types

In addition to measuring a variety of different college sizes, the IPEDS dataset includes entries from many different types of educational institutions. These include private, 2-year, for-profit technical schools to public, 4-year universities. For the sake of this study, we are primarily interested in public and private 4-year institutions. These two groups represent some of the most reputable colleges in the world, and they are what most people think of when discussing the increasing costs of higher education.

How exactly does tuition vary between private and public institions? We can get a measure of the difference by looking at the conditional PMFs (showing the probability of a tuition given that the institution is private or public). Figure 4 shows the conditional PMF for 4-year private institutions and 4-year public institutions. From these plots, it is clear to see the difference between the types of schools. Public institutions are much more likely to have lower tuition and fees than corresponding private colleges. Table I shows the summary statistics calculated given the same conditions.

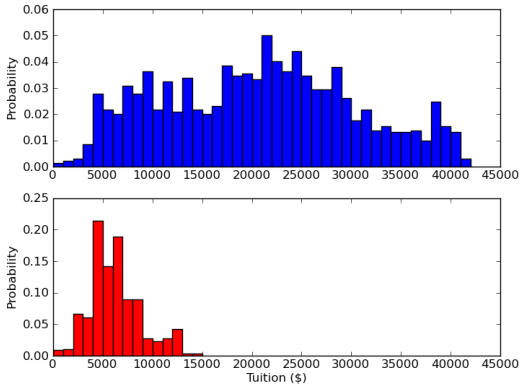

Figure 4. Conditional PMFs for 4-year private institutions (above) and 4-year public institution (below).

Table I
SUMMARY STATISTICS FOR 4-YEAR PRIVATE AND PUBLIC INSTITUTIONS.

Institution Type	*Mean Tuition and Fees*	*Standard Deviation*	*Coefficient of Variation*
Private 4-year	$20,691	$9,740	0.47
Public 4-year	$6,253	$2,600	0.42
Combined 4-year	$15,888	$10,575	0.67

From this investigation, we find that the mean cost for tuition and fees at a private 4-year institution is over three times as the cost of a corresponding public education (at least on paper). Similarly, the standard deviation, the square root of the variance, is several times larger for private colleges than public ones. For this reason, it is better to compare variability between the groups by using the coefficient of variation ($\frac{\sigma}{\mu}$), which standardizes the variations for comparison purposes.

VI. Statistical Significance

These two groups have a significant difference in their mean tuition and fees, but could this simply be due to chance? Intuitively, it seems very unlikely; however, we can more rigorously determine the statistical significance of this observation by testing the difference in means. The null hypothesis in this case is that the distributions for the two groups are the same, and therefore the observed difference in means is completely due to chance. To check whether this hypothesis is true, we resample the pooled tuition values with the same

number of samples as observed public and private institutions (645 and 1,294 respectively). We then compute the mean of these randomly generated groups and take the difference between the two. We can then measure how often this random difference in means is greater than the observed difference in means. What we find is that it is exceedingly unlikely that the observed effect is do to chance; even running the test 1,000,000 times didn't yield a single difference that was greater that the observed difference in means (p-value = 0.0). This suggests that the probability of this phenomena occurring by chance is less that 0.0001% and strongly suggests that the effect is statistically significant. In other words, we are right to separate these tuitions into distinct groups of public and private 4-year colleges - the two are not exactly the same and have different trends and effects.

VII. Normal Distributions

Given these two different types of educational institutions, we notice that the conditional PMF plots look a lot like Gaussian distributions. We can actually test for a normal distribution by creating a normal probability plot. This is done by plotting sorted values from the dataset versus sorted random values from a standard normal distribution. A straight line in a normal probability plot indicates that the data matches a normal distribution well. Figure 5 shows the normal probability plots for tuition at both private and public 4-year institutions. In both cases, the results show a strong linear correlation, suggesting that a normal distribution is a reasonable model for the data. If we take a linear least squares fit of the data in the normal probability plot, we can get estimates for the mean (given by y-intercept) and standard deviation (given by slope). In this case, we find that the estimated mean tuition is $20,367 with a standard deviation of $10,036 for private 4-year institutions while public 4-year institutions have an estimated mean cost of $6,207 with a standard deviation of $2,596. These values are extremely close to the calculated statistics shown in Table I (under 3% error). Note that both graphs have a slight S-shaped curve to them indicating shorter than normal tails (less variance than expected in a perfect Gaussian distribution).

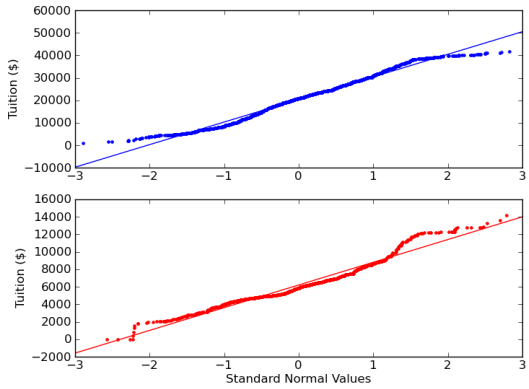

Figure 5. Normal probability plots for tuition at private 4-year institutions (above) and public 4-year institutions (below) along with linear least squares fits.

It actually makes a lot of sense that tuition and fees at 4-year colleges would fall into normal distributions (within their respective categories). Tuition is affected by many different independent factors ranging from professors' salaries to equipment maintenance. Given that each of these independently varying costs are combined to get the total cost (and resulting tuition and fees), the central limit theorem suggests that the final distribution should be normal. It is a near-perfect example of this principle. Public institutions are slightly different than private colleges in that they have additional government support and the economies of scale (hence the separation); however exact same effect can be observed in both, distinct groups.

VIII. Identifying Individual Institutions

Given this understanding of the distribution of college tuitions, it seems logical to ask: where do particular schools fall on this list? To calculate the percentile ranks for individual institutions, we can use a continuous distribution function, or CDF. Figure 6 shows the CDF of tuitions for public and private 4-year institutions. The distribution looks surprisingly linear, showing how variable the costs of tuitions really are – there is not a single tuition value that most of the colleges are especially near. From this CDF, we can find the percentile rank for any given school: we simply look up the probability for a given tuition. Table II shows the cost and percentile rank for several well known private and public institutions. It is not too surprising that many of the high-profile private institutions have high percentile ranks for tuition and fee expenses. It is important to keep in mind, however, that this chart just shows the published tuition and fees and does not take into account financial aid packages. For example, every student admitted to Olin College receives a half-tuition merit scholarship.

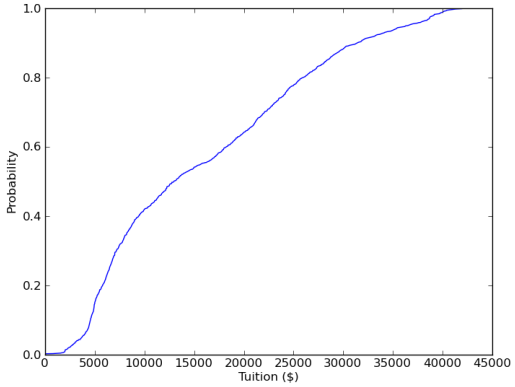

Figure 6. CDF of tuition for public and private 4-year institutions.

IX. Recruiting Success

The difference in costs between various colleges brings up some interesting questions about where the money and "brightest" minds go. Obviously institutions want to have the best programs and build their reputations; however, this does not happen overnight. It takes investments in talented

Table II
COST AND PERCENTILE RANK OF SELECT COLLEGES AND UNIVERSITIES.

College	*Tuition and Fees*	*Percentile*
Harvard University	$37,012	95.6%
Princeton University	$35,340	94.2%
Yale University	$36,500	95.1%
Columbia University	$41,316	99.8%
Olin College	$36,795	95.4%
UC - Los Angeles*	$8,266	35.3%
University of Florida*	$4,373	7.8%
UNC - Chapel Hill*	$5,625	19.0%

*in-state tuition and fees

professors and students. Recruiting is a big part of this; the best schools in the nation recieve far more applications than they can accept each year. The question is: do selective institutions get the better students because of this?

Obviously we cannot get a completely straight answer to this question, but we can use the dataset that we have to look for correlations. In particular, we examine the correlation between 75th percentile SAT scores[3] and the selectivity of the college as given by the percent of applicants accepted. Figure 7 shows a scatter plot with these two variables. it is quite clear from this plot that there is some level of correlation between the two variables. We calculated a Pearson's correlation of -0.314 using the equation $\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$. The two variables are negatively correlated as we expect from the plot (an increasing percent of accepted applications is correlated with a lower SAT score). Therefore, to some extent, we can say that more options may indeed lead to "better" results.

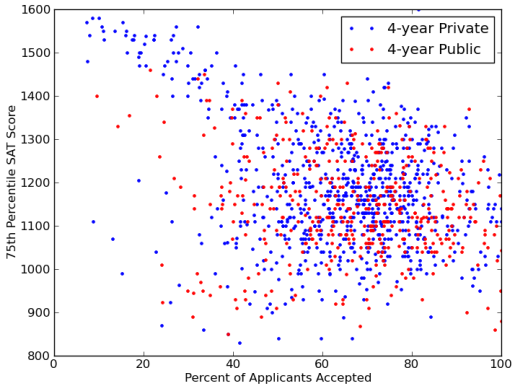

Figure 7. Scatter plot showing the correlation between institutional selectivity and the standardized test scores.

X. BAYESIAN ESTIMATION

Given the large disparity between public and private institutions, it seems plausible that we could identify what kind of school a person attends based solely on the tuition that they are being charged (whether or not they have financial aid). Obviously, such a method would not be definitive, be we could at least update the probability based on such information. As an example, I have a friend in college who I know is paying around $8,500 in tuition and fees. I would like to know what the probability is that he is attending a private college. To calculate this, we can use Bayes's theorem to update the conditional probability of the hypothesis that my friend is going to a private institution given the evidence that he is paying $8,500 per year. We can calculate the posterior probability using the equation $P(H|E) = P(H)\frac{P(E|H)}{P(E)}$, where H is the hypothesis that my friend attends a private college and E is the advance that he pays $8,500 per year in tuition in fees. The prior probability of the hypothesis does not take into account tuition and is given by $P(H) = \frac{\#\,of\,private\,colleges}{total\,\#\,of\,colleges} = \frac{1294}{1939} = 0.667$. The probability of paying $8,500 per year given that you go to a private college can be found using the conditional PMF we created earlier. This gives us $P(E|H) = 0.027$. Finally, the overall probability that a student is paying $8,500 in tuition and fees can be found using a PMF of the pooled costs for both public and private 4-year institutions. This gives us $P(E) = 0.048$. Plugging in, we get a posterior probability of $P(H|E) = 0.667(\frac{0.027}{0.048}) = 0.383$. In other words, the probability that my friend goes to a private school went down as given by the likelihood ratio ($\frac{P(H|E)}{P(E)} = \frac{0.027}{0.048} = 0.57$).

This result fits with what we would expect – there are far more public universities with tuitions in the $5,000-$10,000 range than private colleges with such low expenses. Therefore, knowing that he is paying $8,500 per year might lead us to believe that he goes to a public institution. Similarly, if he had been paying $20,000 per year in tuition, then we would essentially know that he attended a private institution, since there are almost no 4-year public programs that cost that much (at least with in-state tuition).[4]

XI. GEOGRAPHIC DISTRIBUTION

In addition to the different school types within the IPEDS dataset, institutions from different geographic regions are identified. Table III lists the different regions across the United Stats and costs associated with attending a 4-year college within that geography. Based on this chart, it would appear that tuition is, in fact, dependent upon geography. After all, places like New England have over twice the mean tuition and fees as the Southwest. This begs the question of what can cause such large variations. We know that there is a disparity between the costs of public and private institutions, so it seems logical to say that the regions with the lower costs will have a higher percentage of public institutions.

We can validate this hypothesis using a chi-square test. We set up our test by defining a set of cells that a college might fall into. In this case, we have 16 cells as given by our two groups (public and private institutions) and eight different geographic regions. Next, we compute the number of colleges in each cell. Under the null hypothesis we assume that geography doesn't matter and all regions have the same percentage of public institutions (29.3%). This means that each region should have 0.293*(number of colleges in that region) public institutions assuming the null hypothesis were true. We then compute the

[3]Obviously, students who score higher on the SAT are not necessarily "better" students – they are simply better at taking the SAT. Nevertheless, it is one of the few quantitative measures of ability that we have available.

[4]It turns out that my friend actually attends Doane College in Nebraska, a private institution.

Table III
MEAN TUITION COSTS BY GEOGRAPHIC REGION.

Region	*Mean Tuition and Fees*	*Standard Deviation*
New England	$22,793	$11,109
Mid East	$18,902	$11,413
Great Lakes	$16,963	$9,444
Plains	$15,433	$8,660
Southeast	$13,203	$8,904
Southwest	$11,580	$8,645
Rocky Mountains	$10,250	$8,807
Far West	$17,047	$11,908

difference between the observed value (O_i) and the expected value (E_i) and use it to find the chi-square statistic ($\chi^2 = \Sigma\frac{(O_i-E_i)^2}{E_i}$) which gives us a measure of the total deviation. Finally, we use a Monte Carlo simulation to compute the p-value (probability of seeing a chi-square statistic as high as the observed value under the null hypothesis). In this case, we find that $\chi^2 = 61.10$ which would occur by chance less than one time every 100,000 (p-value < 0.00001). This tells us that there is a statistically significant difference in the number of public vs. private institutions within these different regions. That is to say: geography is important! Not all regions are the same, suggesting that we were correct to expect a higher rate of public colleges in regions with lower mean tuition costs. Table IV shows the actual percentage of colleges that are public in each region. Note that the regions with lower mean tuition do, in fact, have a higher percentage of public institutions. And we were able to determine this by simply observing the differences in mean tuition!

Table IV
PERCENTAGE OF 4-YEAR COLLEGES THAT ARE PUBLIC FOR EACH GEOGRAPHIC REGION.

Region	*% Public*
New England	22.2%
Mid East	25.3%
Great Lakes	24.6%
Plains	25.1%
Southeast	34.6%
Southwest	46.5%
Rocky Mountains	53.6%
Far West	26.5%

XII. TIME VARYING TRENDS

All of the information presented to this point has dealt with data just from the 2009-10 academic year. However, much of the discussion around rising tuition prices is based around the rate at which it is increasing. Given our investigation into public and private 4-year colleges, it would be interesting to see whether these two groups have been changing at different rates. If we simply look at the mean difference from the 2008-09 academic year to the 2009-10 academic year, we find a mean 4.1% increase in tuition and fees for 4-year private colleges and a 6.3% increase for private 4-year institutions. However, these summary statistics hide some of the information. Figure 8 shows the percent increase in tuition between the past two academic years for both private and public institutions. There are a couple of interesting apparent effects happening here. To begin, note the peaks representing a zero percent change which skew the mean calculation a little bit. It makes sense that we would see some sort of peak here, given that at least a few colleges are trying to limit tuition inflation (or simply have strict guidelines on tuition changes). This is especially true for public institutions (note the higher peak in the graph). More interestingly, though, is that there is a second "peak" in the PMF for public colleges around 20%. Further investigation reveals that almost all of the schools in this minor peak are California state schools. It would appear that the state decided to raise tuition at all of its institutions by around $1,000, and because the state is so large, the effect of this single decision is apparent on this larger comparison. However, what is probably most concerning is the overall rate of increase. When the inflation rate in the United States is only 1-2% for the year[5] and the cost of tuition is rising by 4-6%, we realize just much of an issue this is becoming.

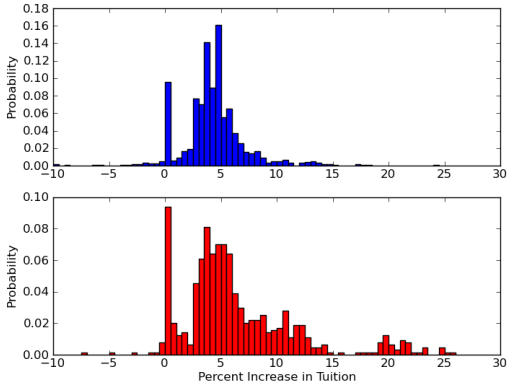

Figure 8. PMFs for the percent increase in tuition and fees between the 2008-09 and 2009-2010 academic years for private (above) and public (below) 4-year institutions.

XIII. LONG TERM TRENDS

The logical next question is how tuition has evolved over a longer period of time. For this investigation we use data from both IPEDS and the College Board's Annual Survey of Colleges which gives us tuition data from 1980-2010. In order to account for inflation, all dollar amounts have been converted into constant 2009 dollars. Figure 9 shows how the price of college has changed in the past 30 years. Note that the increase in expenses is nearly linear over this time period. In order to determine the approximate rate of increase, we found a linear least squares fit for both private and public institutions. From this we found that tuition at 4-year private colleges is going up by approximately $679 dollars per year (2.58%) and 4-year public colleges are increasing by an average of $263 (3.75%) per academic year. Note that these increases are in addition to inflation.

XIV. FUTURE TRENDS

Given all that we have found, we can continue to expect similar tuition inflation to occur for the foreseeable future.

[5]<http://www.bls.gov/cpi/>

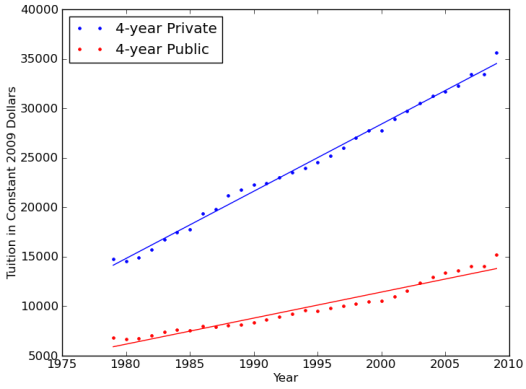

Figure 9. This plot shows the published tuition in constant 2009 dollars for 4-year private not-for-profit and 4 year public from 1979-2009.

Based on the linear model described above, we can expect to be paying at least $26,952 on average for private colleges and $7,283 for a public 4-year option. These estimates are quite conservative as they are based on minimizing the mean squared error for 30 years. It is especially problematic when we take into account the fact that last this past year saw the largest percent increase in tuition rates. Perhaps a linear model may no longer be adequte? If the existing model is wrong it should be updated. With any luck things will begin to turn around before an undergraduate education does become truely unaffordable. In the meantime, we will have to watch tuition costs continue to rise higher.