Investigation of College Tuition 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. This paper broadly investigates the following question: Are college tuition prices really spiraling out of control? Along those lines, it will also 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 U.S. Department’s National Center for Education Statistics (NCES). IPEDS gathers information from every college, university, and technical and vocational institution that participates in federal student financial aid programs. This includes 7316 different institutions for the 2009 dataset. The data was downloaded from the IPEDS Data Center 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.

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 $\lambda$ (where $\lambda$ 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 $\lambda \approx 1.35e - 4$ for this distribution.

V. DIFFERENT INSTITUTION TYPES

In addition to measuring a variety of different college sizes, the IPEDS dataset includes entries from many different type 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. So how exactly do these groups differ? We can get a measure of the difference by looking at the conditional PMFs. We do this by finding the probability of a given 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.

In addition to measuring a variety of different college sizes, the IPEDS dataset includes entries from many different type 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. So how exactly do these groups differ? We can get a measure of the difference by looking at the conditional PMFs. We do this by finding the probability of a given 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.

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 due to chance; even running the test 1,000,000 times didn’t yield a single difference that was greater than 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 Gauss-
sian 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). Both graphs have a slight S-shaped curve to them indicating shorter than normal tails (less variance than expected in a perfect Gaussian distribution).

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 college in that they have additional government support and the economies of scale (hence the initial separation); however exact same effect can be observed in both 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 at the CDF (probability) for a given tuition. Table II shows the cost and percentile rank for several well known private and public institutions. Note that many of these high-profile private institutions have very 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.

IX. 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 we can hypothesize 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
Table III
MEAN TUITION COSTS BY GEOGRAPHIC REGION.

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

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

<table>
<thead>
<tr>
<th>Region</th>
<th>% Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>New England</td>
<td>22.2%</td>
</tr>
<tr>
<td>Mid East</td>
<td>25.3%</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>24.6%</td>
</tr>
<tr>
<td>Plains</td>
<td>25.1%</td>
</tr>
<tr>
<td>Southeast</td>
<td>34.6%</td>
</tr>
<tr>
<td>Southwest</td>
<td>46.5%</td>
</tr>
<tr>
<td>Rocky Mountains</td>
<td>53.6%</td>
</tr>
<tr>
<td>Far West</td>
<td>26.5%</td>
</tr>
</tbody>
</table>

(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 difference between the observed value \(O_i\) and the expected value \(E_i\) and use it to find the chi-square statistic \(\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}\) 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!

X. 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 not-for-profit and 4 year public from 1979-2009.

XI. LONG TERM TRENDS

price in constant dollars - look at slope (% change per year) relative inflation (% change for tuition vs CPI) calculate r squared

XII. ESTIMATE FUTURE TRENDS

look at estimation chapter INCLUDE CONFIDENCE INTERVAL

XIII. CORRELATE VARIABLES

some have argued that this is due to ____. Try to correlate these two sat score and tuition (or SELECTIVITY)
XIV. ADVANCED QUESTIONS

can you identify when someone went to school based on
the tuition they are paying?

OR

develop whether they went to a private or public university
(with a certain level of confidence)

XV. WHY DOES THIS HAPPEN?

XVI. CONCLUSION

We see several interesting effects, but a lot sill missing.
For example, this study does not take into account the increas-
ing level of need-based financial aid.

XVII. NOTES

chi squared test of whether difference in means is signifi-
cant. null hypothesis - all regions are equal

given this distribution, can we hypothesize where a school

is - for example, what is the probability that Harvard is in new

england ONLY given the tuition information

enlighten us of trends in

That is not to say that there aren’t methods to help finance

your education – in fact, an increasing number of students are

receiving grants

How much does college actually cost today? This paper will
investigate several

It is undeniable.

Even with inflation, every time you seem to look up the
costs for tuition and fees at the local university, it has increased
by several percent. It is something that anyone who has seen
the numbers

Fortunately, a multitude of information has been collected
on the topic. In particular, this study will be using data from
the Integrated Postsecondary Education Data System (IPEDS).

Pose an interesting statistical question and find a dataset
that is well-suited to address it. Generate effective summary
statistics that describe several variables from your dataset (re-
codes??). Find an apparent effect that is interesting, surprising
and/or relevant to your question.

Are college prices spiraling out of control? Is the rate
increasing?