UCLA Fall Undergraduate Enrollment Trends (1999-2019)

A Hypothesis Test

Leonid Shpaner

2020-08-13

UCLA undergraduate fall enrollments have been averaging 27,340 for the last 20 years (1999-2019), with a 1.26% growth rate year over year. A noticeable milestone was reached in 2017 when the population of new students crossed the 31k mark. Notwithstanding, from 2018-2019 we have seen a reduction in 0.11% from 31,577 to 31,543 new student enrollments. Given the current state of COVID-19 physical distancing restrictions, coupled with the slight drop off in enrollments between 2018-2019, will we be seeing a notable decrease in new student enrollments from 2019-2020?

To shed more light on potential outcomes of this scenario, we will conduct the following hypothesis test:

H_0 (initial hypothesis): the number of students enrolled in 2020 will be at least 30,000.

$H_0: \mu \geq 30{,}000$

H_a (alternative hypothesis): the number of students enrolled in 2020 will fall below 30,000.

$H_a: \mu \leq 30{,}000$

The last 20 years' worth of data are observed with actual fall enrollment numbers (source: https://www.universityofcalifornia.edu/infocenter/fall-enrollment-glance). The average number of enrolled students is measured as 27340.48 from the 21 observations. Generally speaking, it is important to note that for a true hypothesis test to be more robust, an adequate sample size of at least n = 30 must be gathered or presented; this is based upon the central limit theorem, where the spread of sample means approaches a normal distribution as the sample size, n increases. That being said, we use the parametric t-test in this experiment (which is useful for smaller sample sizes) as we draw a conclusion on our findings. In order to test these numbers, a z-value measuring the difference between the average (mean) of the last 20 years' observations and our null (initially hypothesized value) must be calculated. This mathematical value is expressed in the following:

$$TS = z = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} \Longrightarrow SE_x = \frac{s}{\sqrt{n}} \Longrightarrow TS = (27340.48 - 30000) \cdot \frac{\sqrt{21}}{2443.37} = -4.988.$$

Prior to conducting the test, comments must be made about the potential downside risk of the false positive and false negative errors. A type I (false positive) error would indicate that we have rejected the null hypothesis when it was indeed true. To mitigate this risk, we will assume an alpha (critical p-value) of 0.05. This is chosen based on a wider, yet common 95% confidence level where P = 1 – 0.95 = 0.05 (the probability that the null hypothesis is true). Using this large confidence level will ultimately allow a wider range of data, with enough "wiggle room" for a larger margin of error. In determining any potential type II

(false negative) errors, we may run the risk of not rejecting our null (initial hypothesis) by not factoring in a large enough sample size.

Finally, the hypothesis test is conducted as follows using R.

First, we load the library (readr) so we can load the csv file "UCLAEnrollment" that contains our data set into a data frame called enrollment.

```
library(readr)
enrollment = read.csv("UCLAEnrollment.csv")
```

To see the data set, we proceed to view(enrollment) which loads this into R environment. A quick summary gives us a glance at the minimum and maximum, median, mean, as well as the 1st and 3rd quartiles of each variable in the data set.

```
View(enrollment)
summary(enrollment)
```

```
##       Year         Enrolled        Delta          
##  Min.   :1999   Min.   :24668   Min.   :-0.029900  
##  1st Qu.:2004   1st Qu.:25328   1st Qu.:-0.001225  
##  Median :2009   Median :26536   Median : 0.016200  
##  Mean   :2009   Mean   :27340   Mean   : 0.012555  
##  3rd Qu.:2014   3rd Qu.:29585   3rd Qu.: 0.026475  
##  Max.   :2019   Max.   :31577   Max.   : 0.043500  
##                                 NA's   :1
```

To help us better understand the characteristics of the enrollment figures, let us view the summary statistics of this variable.

It becomes apparent that the average number of enrolled students year over year is roughly 27,340.

```
mean(as.numeric(enrollment$Enrolled, na.rm = TRUE))
```

```
## [1] 27340.48
```

with a standard deviation of 2,443.369.

```
sd(as.numeric(enrollment$Enrolled, na.rm = TRUE))
```

```
## [1] 2443.369
```

The minimum number of enrolled students in the Fall from 1999-2019 was 24,668.

```
min(as.numeric(enrollment$Enrolled, na.rm = TRUE))
```

```
## [1] 24668
```

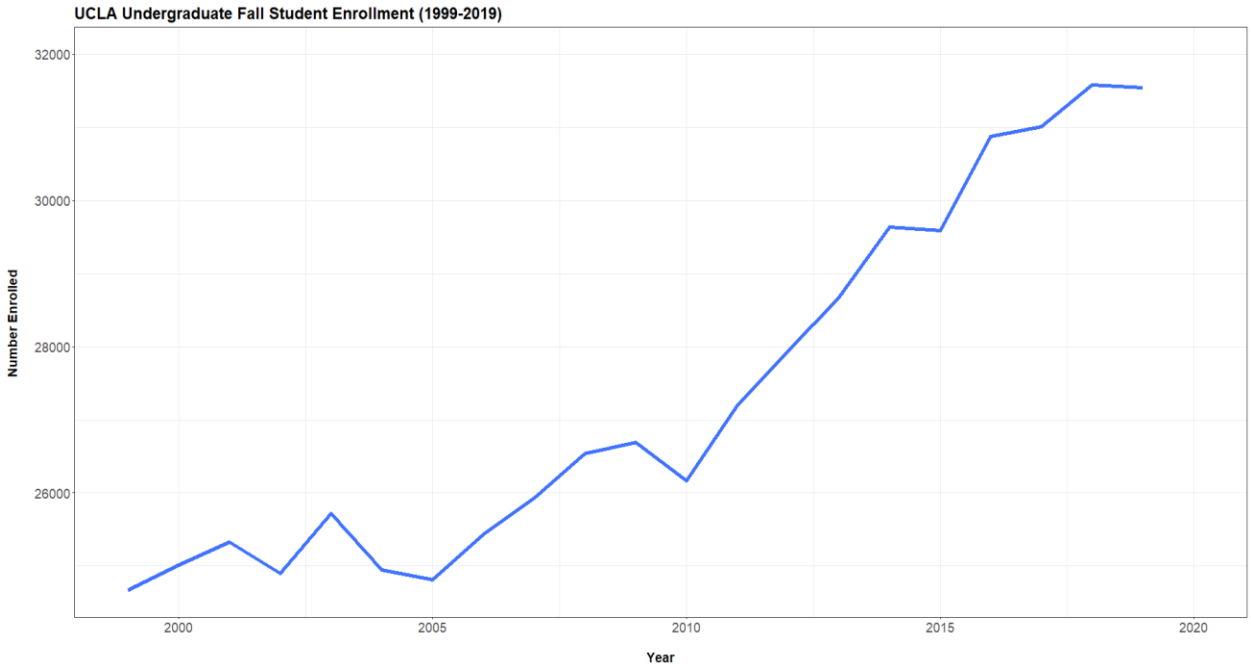
with a maximum number in the same category expressed as 31,577.

```
max(as.numeric(enrollment$Enrolled, na.rm = TRUE))
```

```
## [1] 31577
```

To see this trend represented using a line plot, the ggplot library is loaded and the code executed as follows:

```
library(ggplot2)
knitr::opts_chunk$set(fig.width = 25, fig.height = 8)
options(repr.plot.width = 1, repr.plot.height = 8)
ggplot(data = enrollment[1:2], aes(x = Year, y = Enrolled, group = 3))+
  expand_limits(x = 2020, y = 32000)+
  geom_line(color = "#2a63ff", size = 2, alpha = 0.9, linetype = 1)+
  theme_bw()+
  theme(axis.text = element_text(size = 14),
        axis.title = element_text(size = 14,face = "bold"))+
  labs(title = "Enrolled",
       x = "\nYear",
       y = "Number Enrolled \n")+
  ggtitle("UCLA Undergraduate Fall Student Enrollment (1999-2019)")+
  theme(plot.title = element_text(size = 18, face = "bold"))
```



From the line plot alone, we can gather that enrollment numbers have been trending upward year over year for the last 20 years. However, to examine if we will we be seeing a notable decrease in new student enrollments from 2019-2020, a t test with a confidence level of 95% is conducted below. Here, we are surmising that enrollment numbers will be greater than or equal to 30,000. Alternatively, we are postulating that the numbers will fall below 30,000. Let us now see the results by plugging in the statistics we have calculated manually earlier:

```
xbar = 27340.48
mu0 = 30000.00
sigma = 2443.37
n = 21
z = (xbar-mu0)/(sigma/sqrt(n))
z
```

```
## [1] -4.987968
```

The z-value of -4.988 still holds true from our prior manual test, and when calculated against alpha, we retrieve the critical value as follows

```
alpha = 0.05
z.alpha = qnorm(1-alpha)
# The critical value (-z.alpha)is shown below
-z.alpha

## [1] -1.644854
```

The test statistic -4.988 is lower than the critical value of -1.645. Therefore, at a 0.05 significance level, we must reject our initial claim that student enrollment will reach or exceed 30,000 in the fall of 2020.

As another measure, we can apply the pt function to calculate the lower tail p-value of the test statistic.

```
pval = pt(z, df = n-1)
pval

## [1] 3.5329e-05
```

As shown by the p-value of 3.5329e-05 above being less than the 0.05 significance level, we must reject the null that $\small x \geq 30,000.$

A t-test is also performed on the data, thereby further strengthening the testing grounds, and confirming what we have previously concluded.

```
# t test of x against a null hypothesis

t.test(enrollment$Enrolled, mu = 30000,alternative = "less", conf.level = 
0.95)

##
##  One Sample t-test
##
## data:  enrollment$Enrolled
## t = -4.988, df = 20, p-value = 3.533e-05
## alternative hypothesis: true mean is less than 30000
## 95 percent confidence interval:
##      -Inf 28260.07
## sample estimates:
## mean of x
##  27340.48
```

As evident from the t-test above, the p-value of 3.533e-05 is statistically significant and falls substantially below the alpha level of 0.05. As such, we must reject the idea or notion that Fall Student Enrollment in 2020 will equal or surpass that of 30,000 students (the null hypothesis).