Using Applications, Admissions Data to Forecast Enrollment

Lawrence J. Redlinger*, Sharon Etheredge & John Wiorkowski*****
The University of Texas at Dallas
Presented at the annual meeting
Association for Institutional Research
Long Beach, California May, 18-22,2013

* Professor and Executive Director Strategic Planning and Analysis, **Associate Director Strategic Planning and Analysis, *** Professor and Vice Provost

Questions

1. What are the key issues in selecting a forecast model?
2. Given an institution's mission and student characteristics, what are the appropriate forecasting models for the institution?
3. What are the central issues with the use of applications, admissions data to forecast new enrollment?
4. What are the effects of continuing students, stop-outs and graduates on enrollment forecasts?
5. How can enrollment forecasts be utilized to set enrollment targets?

UT DALLAS
The University of Texas at Dallas

Forecasting Questions*

1. **The Time Horizon**: forecasts are generated for some point in time that vary from immediate to long-term.
2. **Data Patterns**: trend, cyclic, seasonal or a combination of these.
3. **Associated Costs** of: model development, procedures, complexity, data gathering, operational.
4. **Degree of Accuracy** desired or necessary.
5. **Availability of data** that is sufficient, accurate and timely
6. **Ease of Operation and Understanding** (administrators should be able to understand how the forecasts are made).

UT DALLAS
The University of Texas at Dallas

* see Bowerman, O'Connell and Koehler, 2005.

"Predictions are best made in *a stable system* where *trends are well established* and *rates of change* for all variables *are known*. It is even better if that stable system is nested in *a stable environment*."

FAT CHANCE

UT DALLAS
The University of Texas at Dallas

"The farther away the projected time is from the present the more likely the projection will err by some degree. This is especially so when the environment is turbulent, the prediction involves many variables, and/or the system is undergoing continuous change."

UT DALLAS
The University of Texas at Dallas

DIMENSIONS OF CHANGE

The Source of Change: Internal to External

The Duration of Change : Short to Long

Magnitude of Change : Small to Large

Frequency of Change : Single or Multiple

The Intensity of Change

The degree of Connectivity of Changes

The Threshold Where the Change makes a difference

Impact of Change(e.g., immediate or delayed)

And of course the
System's Response to Change

UT DALLAS
The University of Texas at Dallas

$$E_t = I_t + (C - O)_t$$

Where E = enrollment

I = Input Streams

(for example, First Time In College, Transfers, Masters Students and Doctoral Students)

Where C = Continuing Students

Where O = Output Streams

(Graduates, Drop-outs, Stop-outs, Transfers-out)

And *t = time*

UT DALLAS
The University of Texas at Dallas

$$E_t = I_t + (C - O)_t$$

$$E_{t+1} = I_{t+1} + (C - O)_{t+1}$$

So enrollment at time E_{t+1} will be greater than enrollment at time E if the sum of new students and continuing students is positive.

Increasing new students and decreasing the proportion who leave (retention) are both drivers of enrollment gains.

Retained students make up the highest percentage of enrollment.

UT DALLAS
The University of Texas at Dallas

Two *"Simple"* Steps

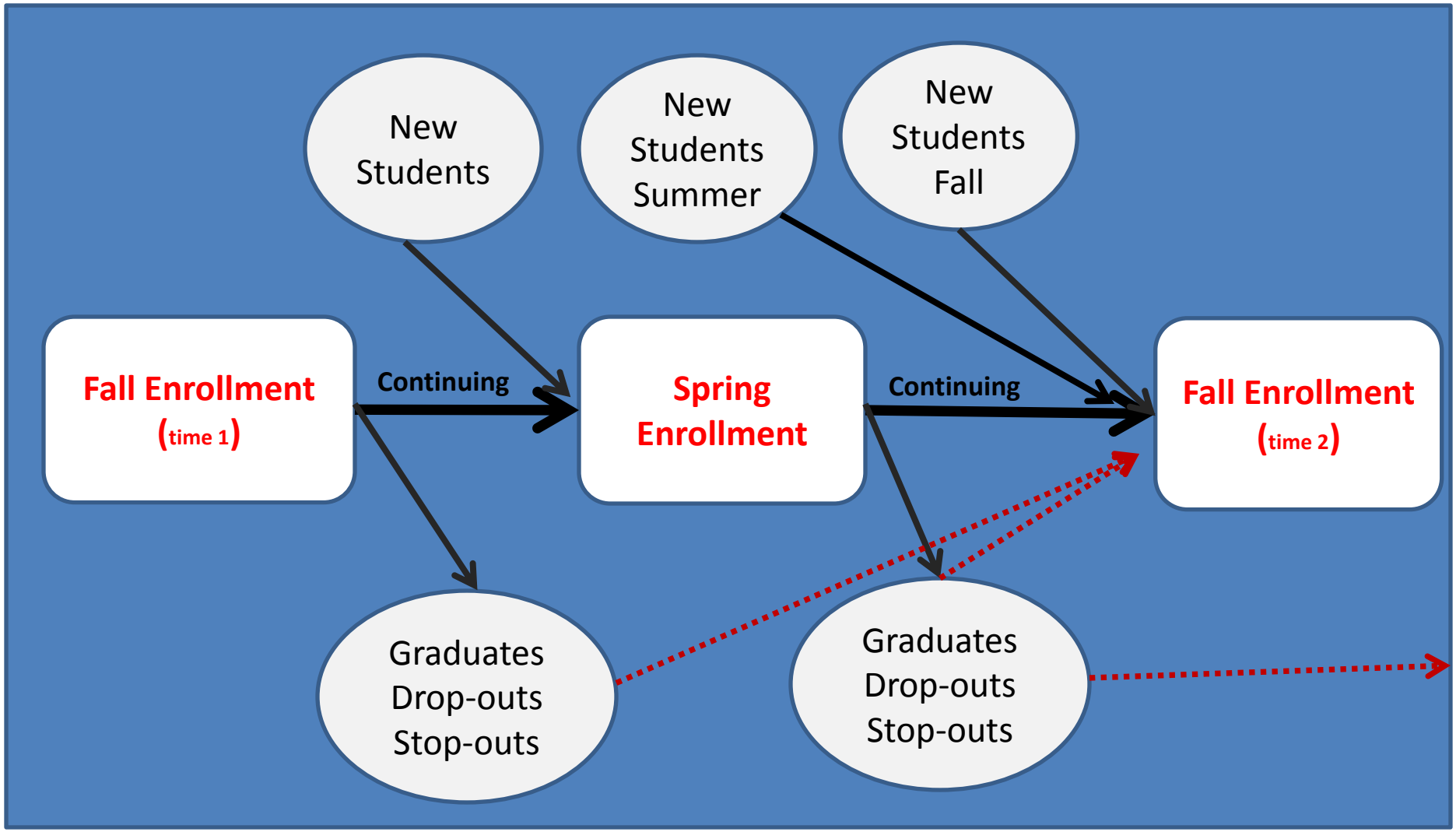
1. Accurately estimate the Number of Continuing Students (C-O)

2. Accurately estimate the Number of New Students (I)

With enough lead time to allow organizational adaptations should they be needed.

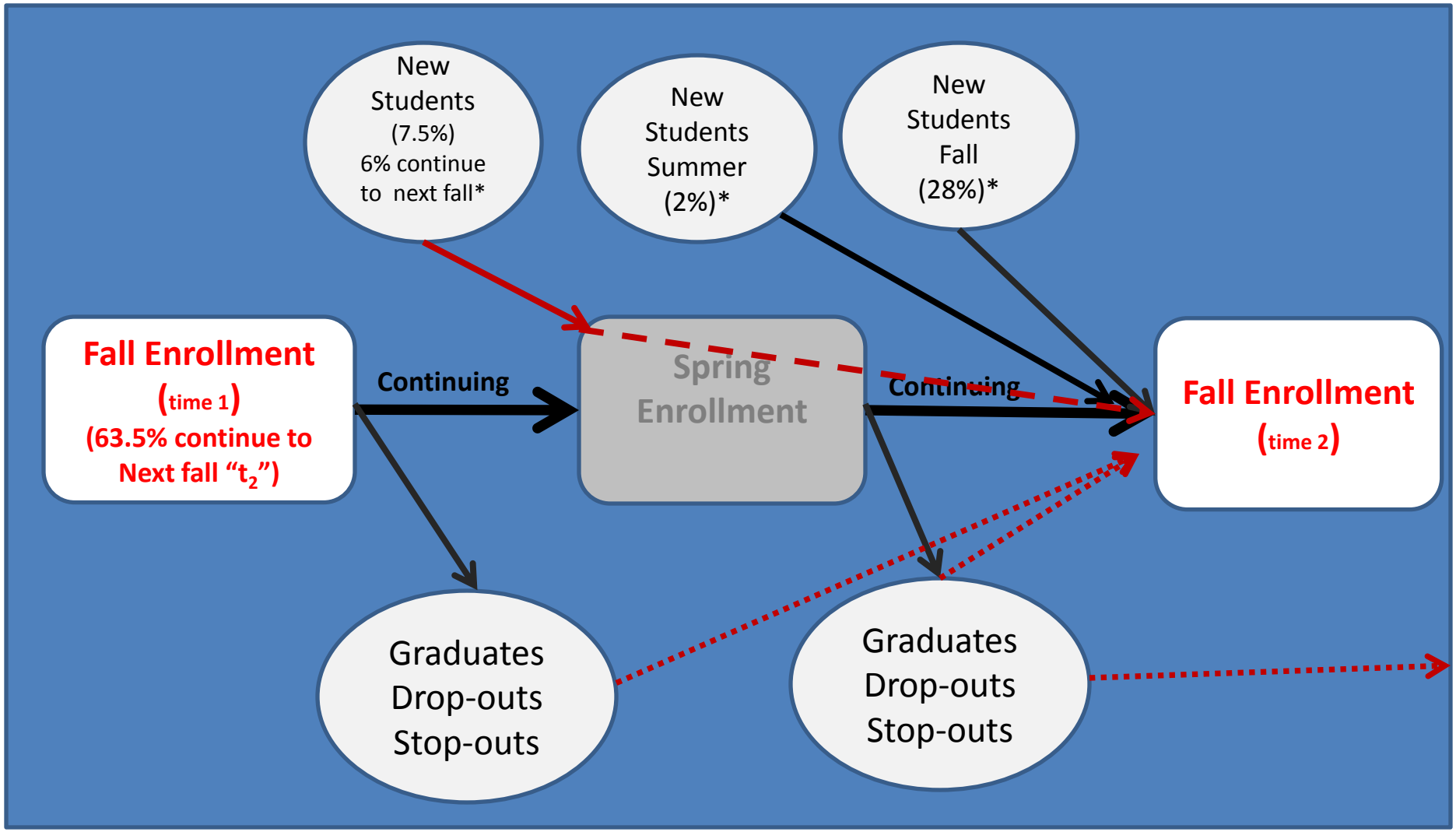
UT DALLAS
The University of Texas at Dallas

Enrollment: Inflow and Outflow

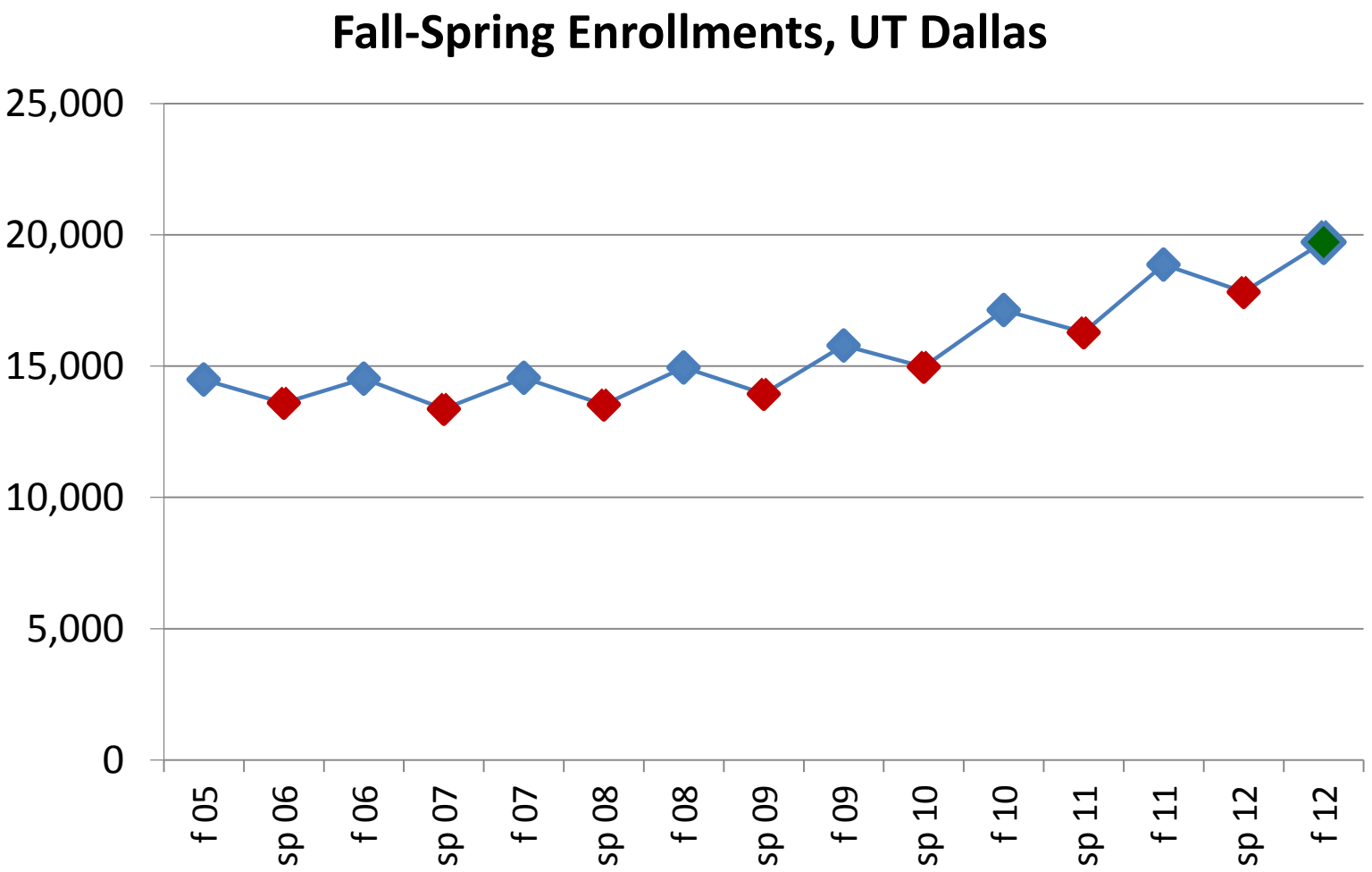


Retained students make up the highest percentage of enrollment.

Enrollment: Inflow and Outflow



Retained students make up the highest percentage of enrollment.



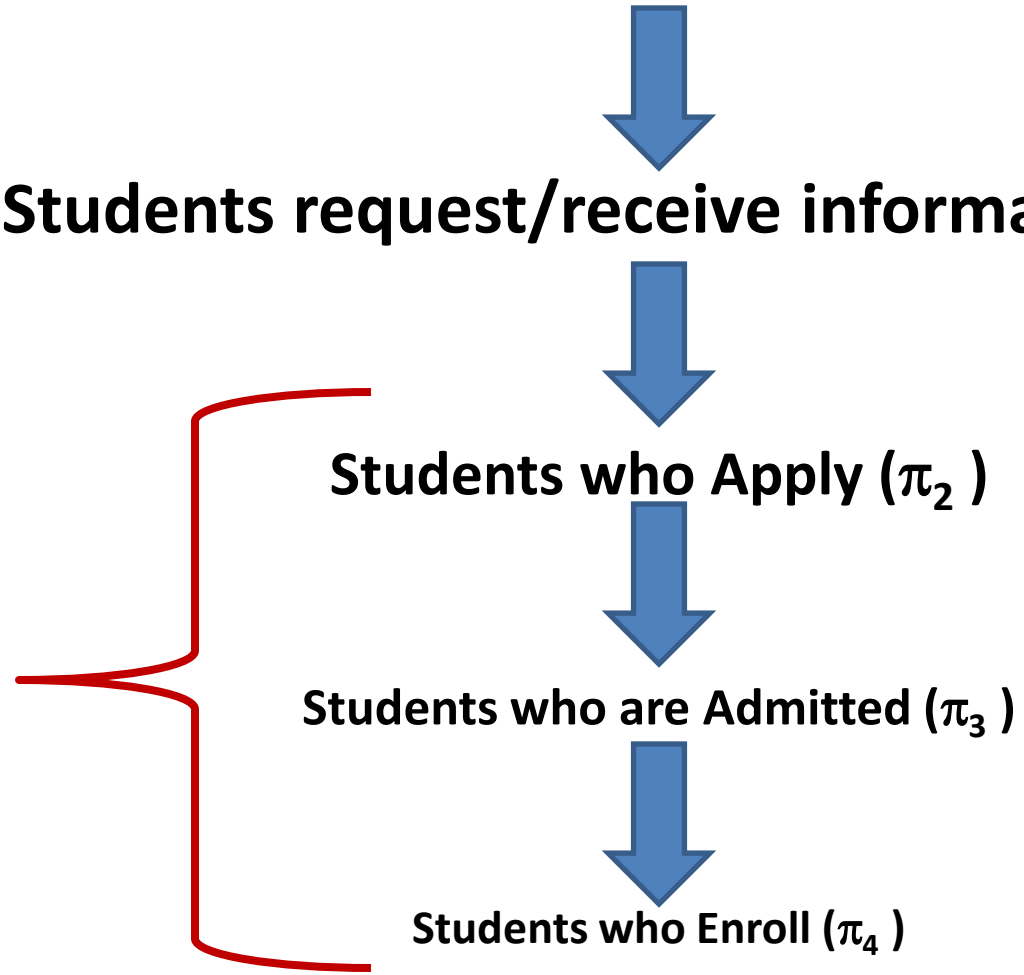
The correlation between fall-to-fall persistence and spring-to-next fall persistence is .97.

CONTINUING STUDENTS

1. Establish if persistence is a stable element or if there have been changes in persistence.
2. Establish the appropriate data unit(s).

Fall-to-Fall	2008-2009	2009-2010	2010-2011	2011-2012
Academic Persistence	61.50%	62.00%	62.50%	63.50%

Potential Population of New Students (θ)



$$\lambda = \theta \pi_1 \pi_2 \pi_3 \pi_4$$

See Wiorkowski, John and Lawrence J. Redlinger, "The Statistical Anatomy of Academic Enrollment Data."

$$\lambda = \theta \pi_1 \pi_2 \pi_3 \pi_4$$

This equation (1) conforms to the facts that if we:

- **Increase initial exposure (i. e. increasing the value of) and/or**
- **Increase the probability of requesting information, and/or**
- **Increase the probability of applying for admission, and/or**
- **Increase the probability of actually being admitted and/or**
- **Increase the probability of enrolling**

Then the average new enrollment will increase.

Over time, efforts in these segments are not likely to be constant. Varying efforts are made to increase exposure, increase or sustain student response, student application, and the enrollment of admitted students (the process of admission is sometimes manipulated to make the probability of admission increase or decrease).

See Wiorkowski, John and Lawrence J. Redlinger, "The Statistical Anatomy of Academic Enrollment Data."

Establish Input Streams

Example: Fall 2012

Fall 2012 Applications, Admits and Enrolled counts as of September 6, 2012

	Applied	Admitted *	Enrolled	% Applied to Admitted	% Admitted to Enrolled	% Applied to Enrolled
Freshmen -- FTIC**	7,081	3,941	1,519	55.7%	38.5%	21.5%
Freshmen Transfers & Transients	660	283	149	42.9%	52.7%	22.6%
Sophomore Transfers & Transients	1,520	1,106	769	72.8%	69.5%	50.6%
Junior Transfers & Transients	1,682	1,312	927	78.0%	70.7%	55.1%
Senior Transfers & Transients	335	209	143	62.4%	68.4%	42.7%
Graduate Non-Degree Seeking & 2nd Bacc.	689	445	284	64.6%	63.8%	41.2%
Graduate Degree Seeking	11,706	6,449	2,578	55.1%	40.0%	22.0%
TOTAL	**23,673**	**13,745**	**6,369**	**58.1%**	**46.3%**	**26.9%**

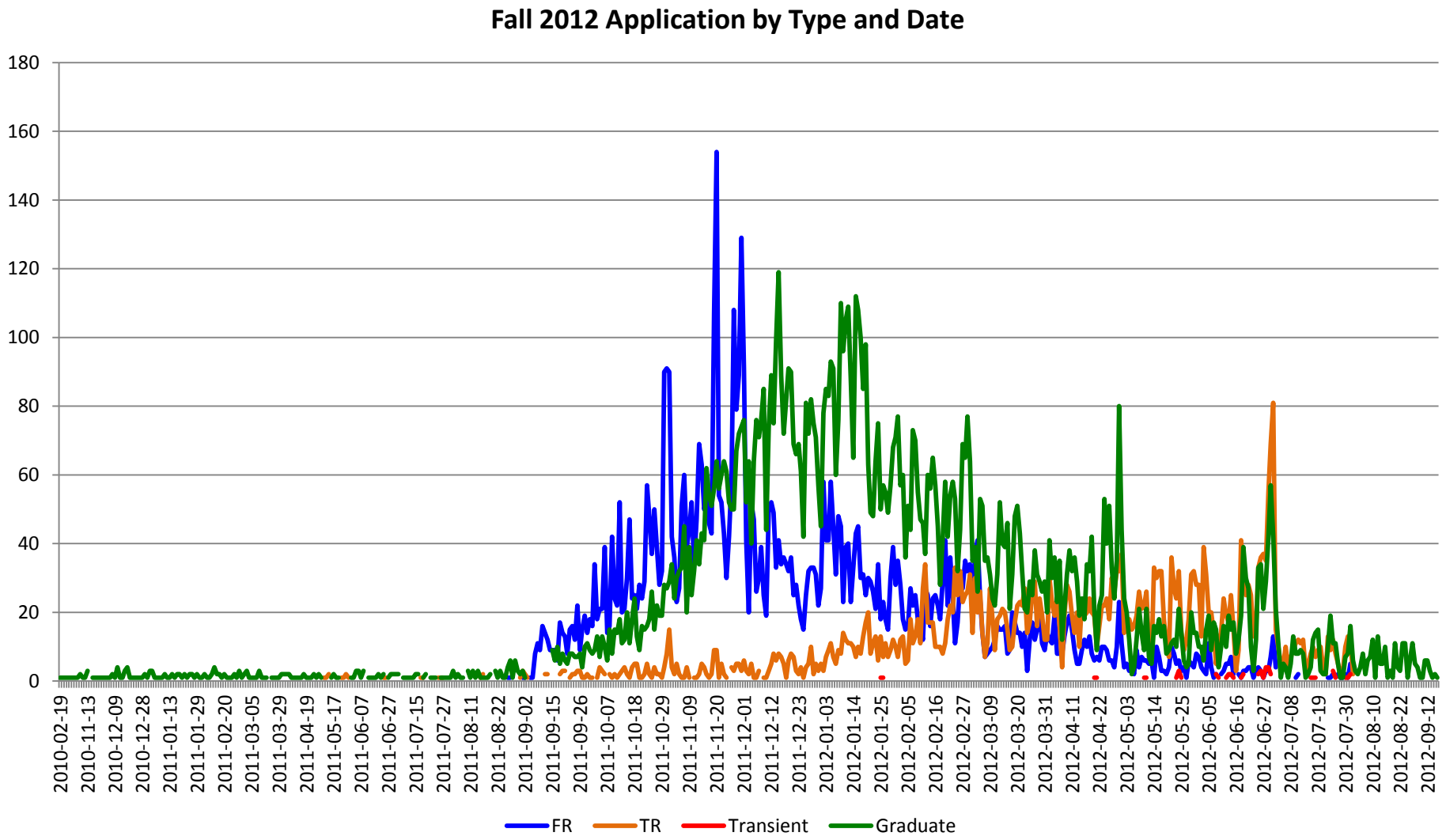
* Includes admitted applications that were later cancelled.

** Includes FTIC "sophomores" and "juniors"

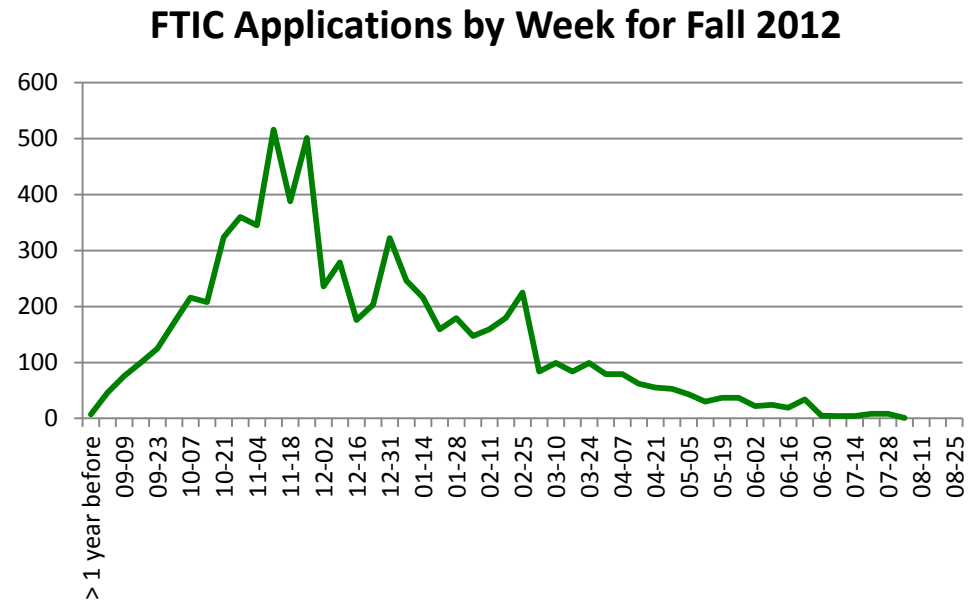
UT DALLAS
The University of Texas at Dallas

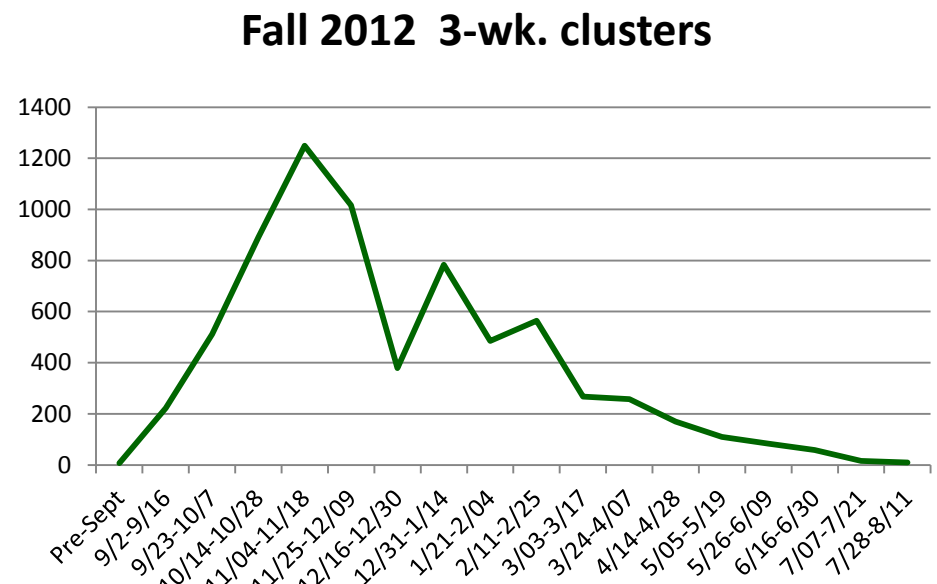
APPLICATIONS: PERIODICITY

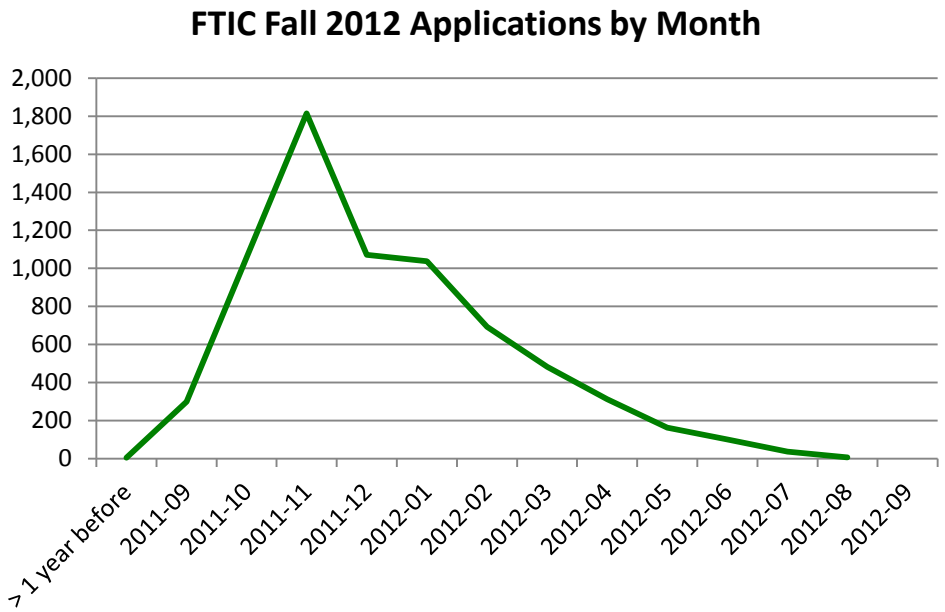
Daily level data are very "noisy"



SMOOTHING THE DATA

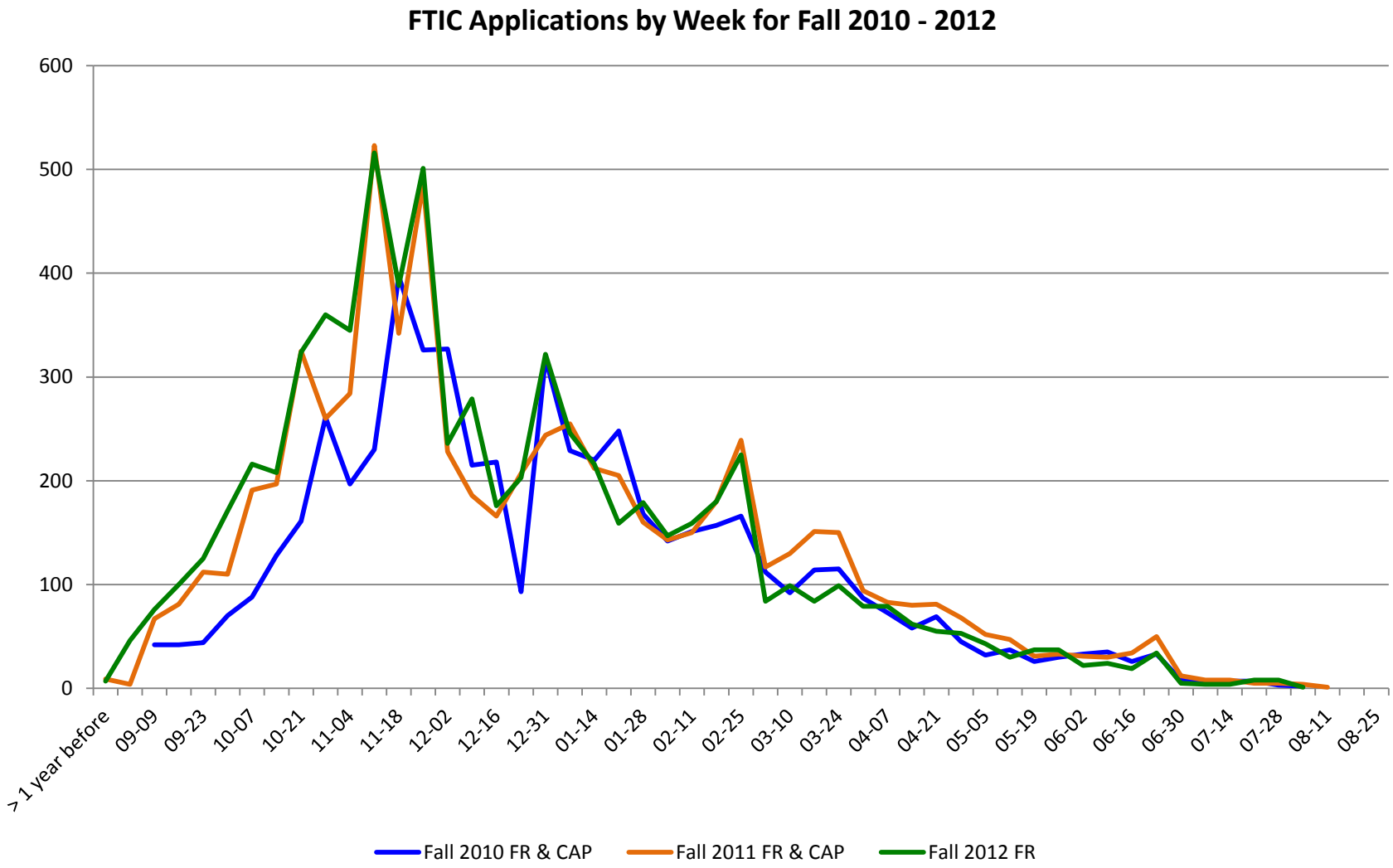






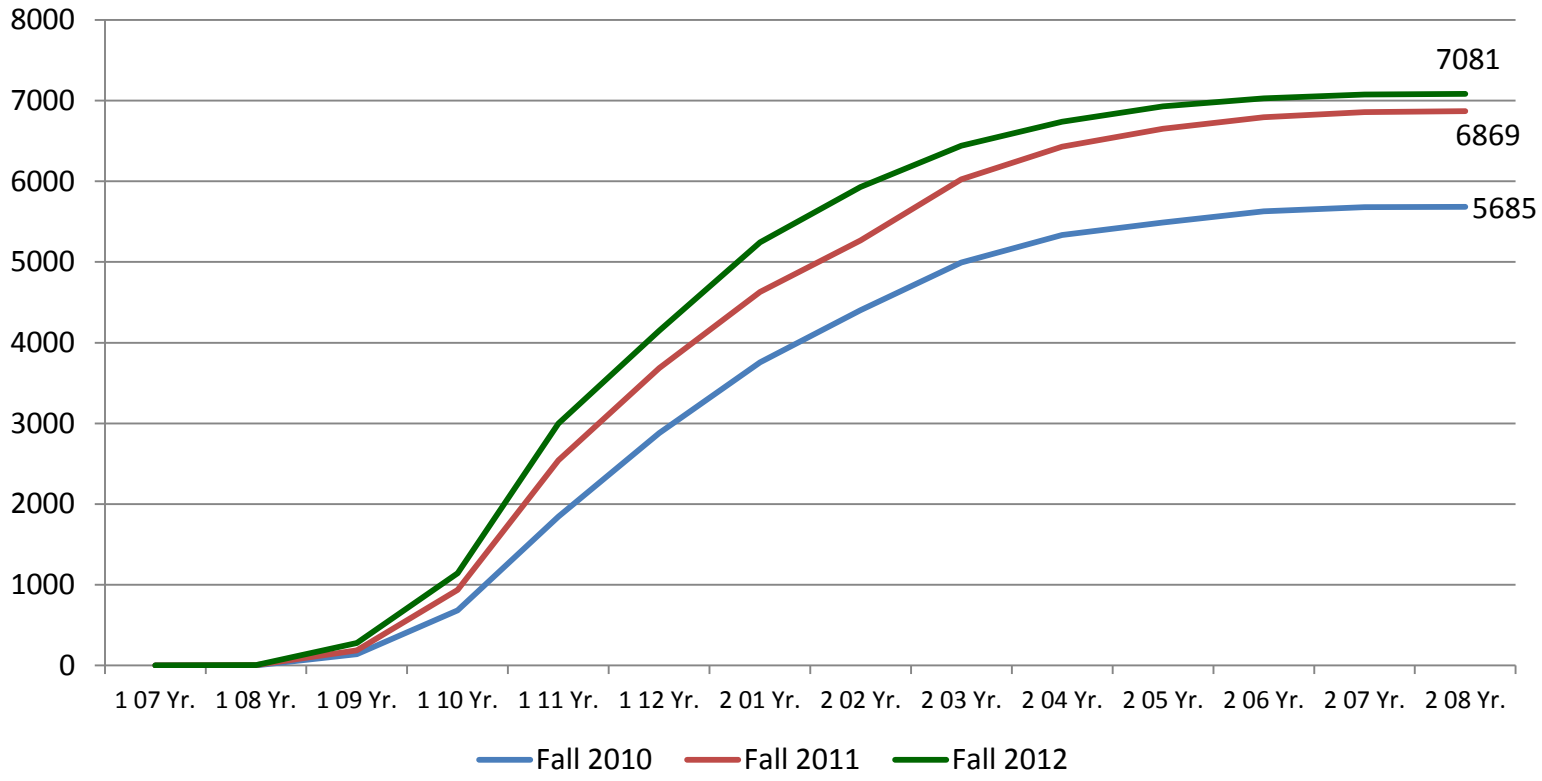
PERIODICITY

Three Year Comparison for FTIC Applications



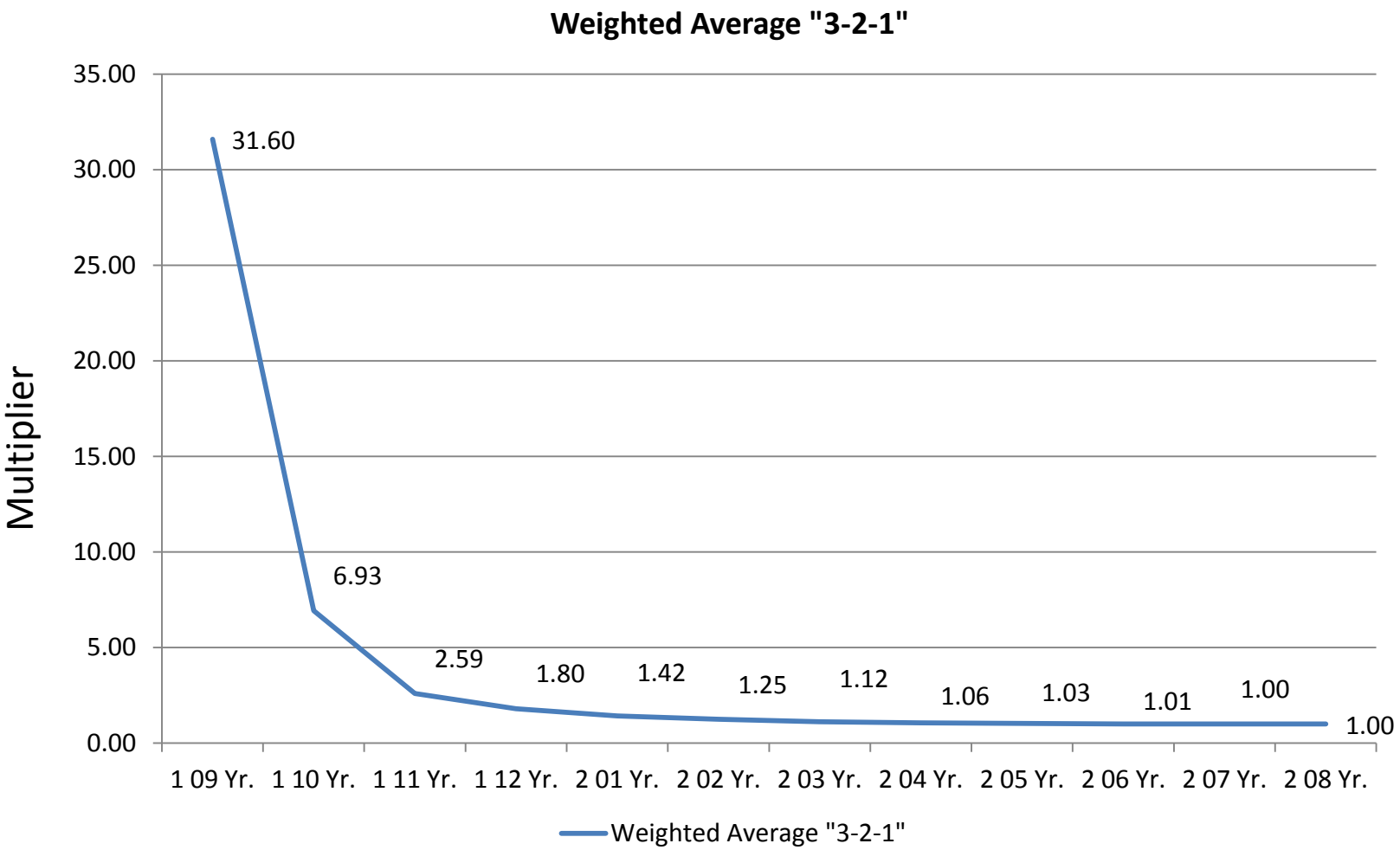
PERIODICITY

Three Year Comparison for FTIC Applications Cumulative



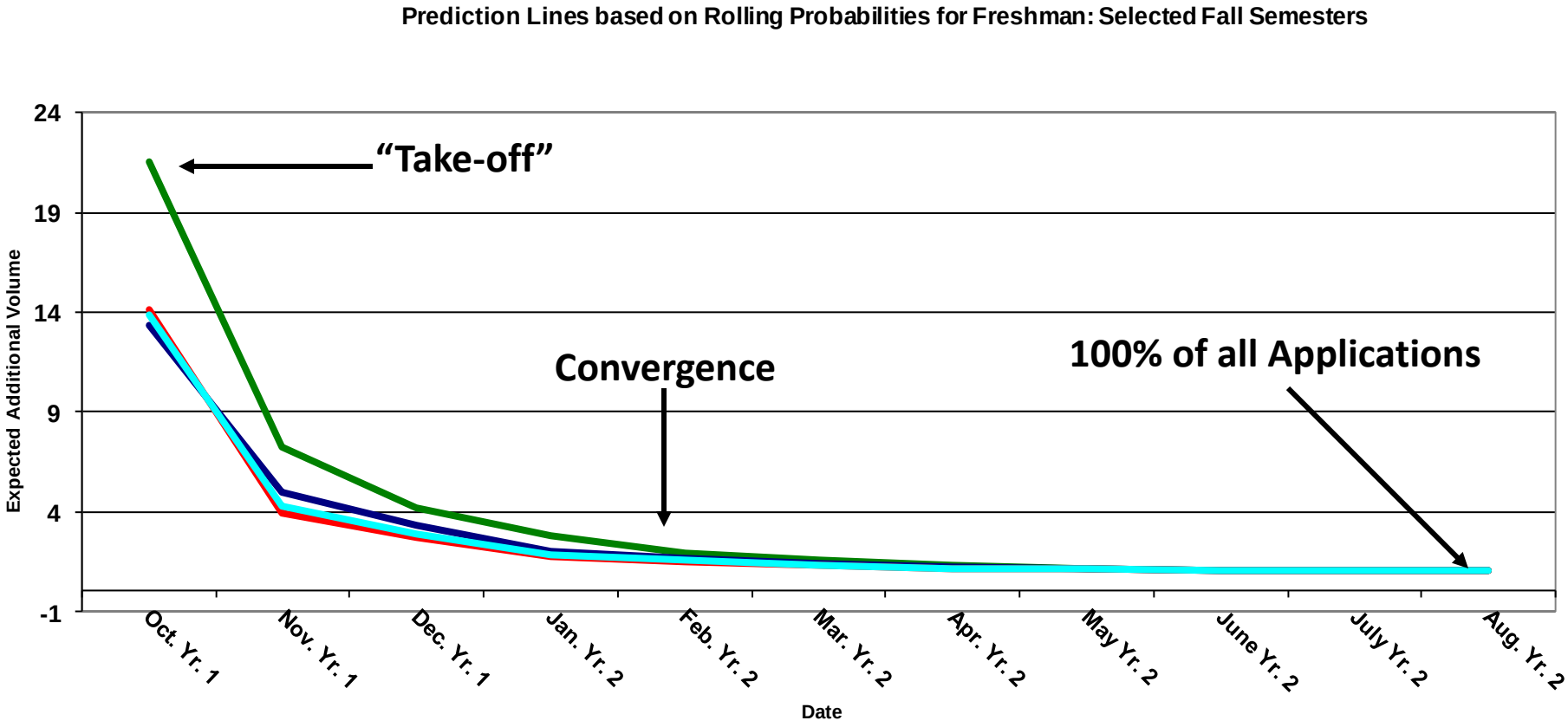
Fall 2012 Forecast Line with State Multipliers Based on a Weighted Average Model:

$$\frac{F09 + 2(F10)+3(F11)}{6}$$

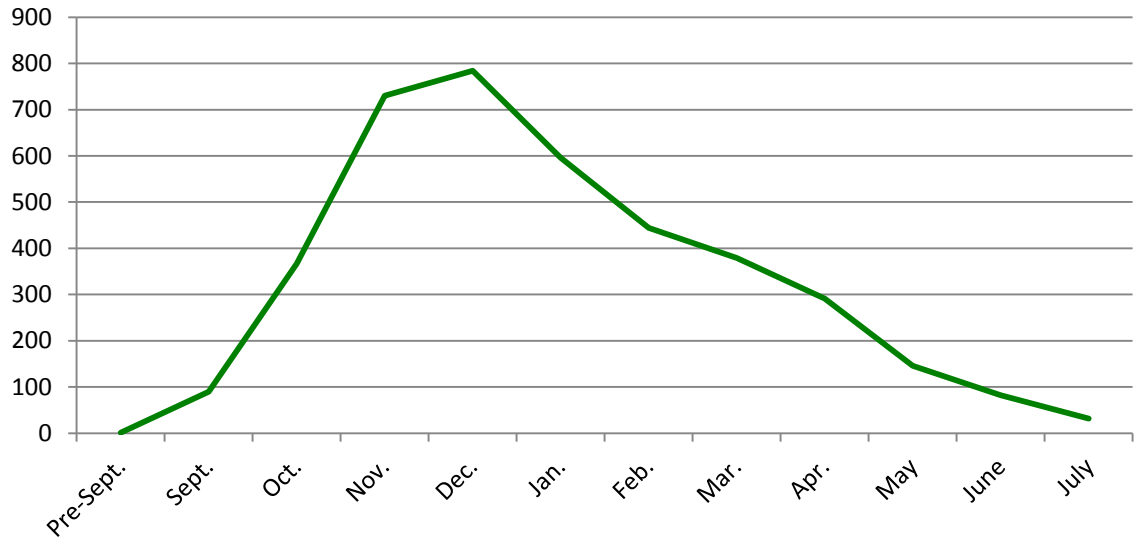


$$\Sigma A^f = (a^{t1,2,\ldots n})(m^{t1,2,\ldots n})$$

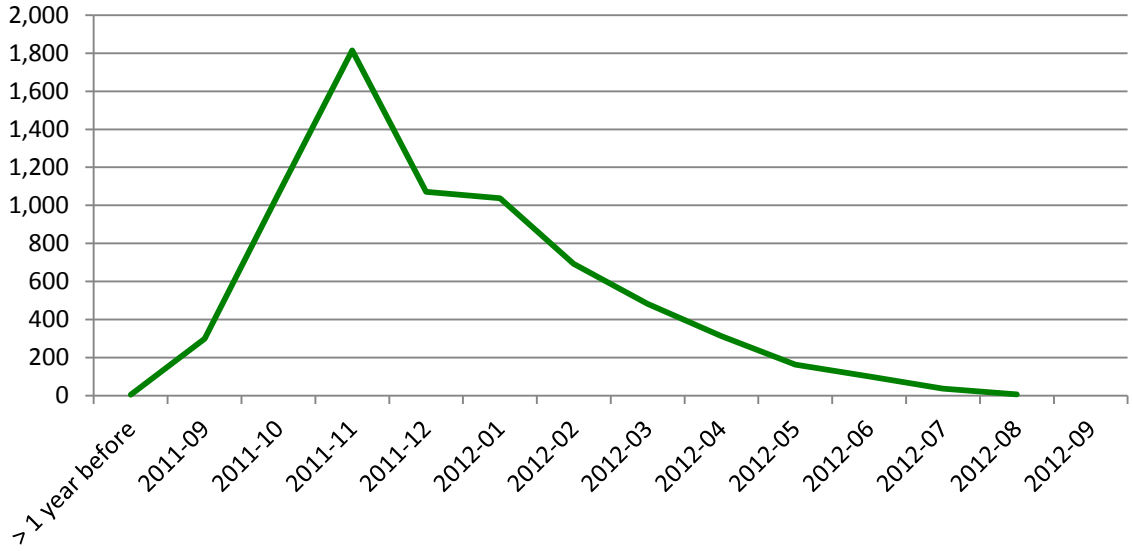
Where A^f = total applications; a = applications at time 1…n, and m = state multiplier at time 1…n


Prediction Lines based on Rolling Probabilities for Freshman: Selected Fall Semesters
24
19
14
9
4
-1
Expected Additional Volume
"Take-off"
Convergence
100% of all Applications
Oct. Yr. 1
Nov. Yr. 1
Dec. Yr. 1
Jan. Yr. 2
Feb. Yr. 2
Mar. Yr. 2
Apr. Yr. 2
May Yr. 2
June Yr. 2
July Yr. 2
Aug. Yr. 2
Date

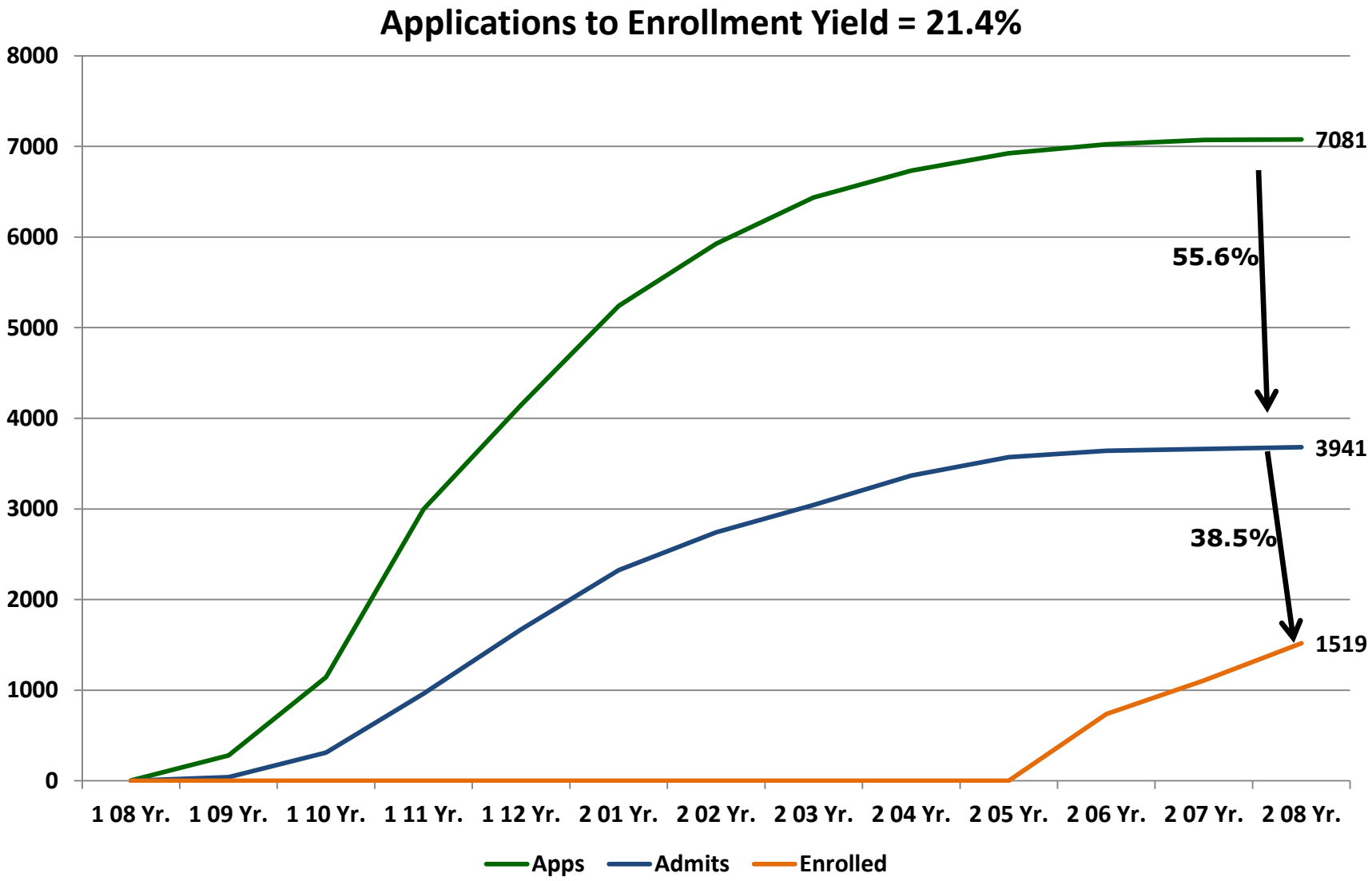
FTIC 2012 Admissions



FTIC Fall 2012 Applications by Month



Fall FTIC* Applications, Admissions and Enrolled



*does not include summer applications, admissions and enrolled for FTIC

Additional Comments

- Scan the internal environment
  - Housing deposits
  - Orientation Registrations
  - I-20 Applications
- Scan external environment
  - Universities you have identified (e.g. by SAT/ACT scores) as alternatives to your university. Have they changed their goals, policies, financial aid packages?
  - Feeder high school demographics

Selected References

- Bowerman, Bruce L., Richard T. O'Connell and Anne B. Koehler, Forecasting, Time Series, and Regression: *An Applied Approach*. 4th ed. 2005. Brooks/Cole a division of Thomson Learning, Inc. ISBN: 0-534-40977-6.
- Brinkman, Paul T. and Chuck McIntyre, "Methods and Techniques of Enrollment Forecasting," in New Directions for Institutional Research, no. 93, Spring, 1997, Jossey-Bass.
- Guo, Shuqin, "Three Enrollment Forecasting Models: Issues in Enrollment Projection for Community Colleges," presented at the 40th RP Conference, May 1-3, 2002, Asilomar, Pacific Grove, Cal.
- Redlinger, Lawrence J. and Sharon Etheredge, "Using Student Classification Specific Applications and Admissions Data to Forecast Enrollment, "presented at AIR 2004, Boston, Mass.
- Redlinger, Lawrence J. and Stanley L. Gordon, "A Comparison of Time Horizon Models to Forecast Enrollment," presented at AIR 2004, Boston, Mass.
- Reiss, Elayne, "Best Practices in Enrollment Modeling: Navigating Methodology and Processes, presented at 2012 FACRAO Conference – St. Augustine, FL, June 5, 2012 (http://uaps.ucf.edu/doc/FACRAO_2012_Enrollment.pdf
- Rylee, Carol and Dale Trusheim, "Enrollment Projections and the Budget Process: A Technique for Smart Planning," presented at SCUP-39 Annual Conference, Toronto, Canada, July 20, 2004. (http://www.udel.edu/IR/reports/presentations.html)
- Wiorkowski, John and Lawrence J. Redlinger, "The Statistical Anatomy of Academic Enrollment Data."

UT DALLAS
The University of Texas at Dallas

Thank you!

This presentation will be available online at:

http://www.utdallas.edu/ospa/research/Conference%20Presentations/AIR/AIR.html

The Statistical Anatomy of Academic Enrollment Data

by

John J. Wiorkowski
The University of Texas at Dallas

and

Lawrence J. Redlinger
The University of Texas at Dallas

Model Development

The size of the enrollment in a particular academic category (for example a discipline like Biology, or a rank such as Undergraduate, or a degree such as Ph.D) is made up of two components. The first is the number of students who have been admitted in the category at a particular time, and at subsequent time points the number of those previously admitted who are still at the institution. Consider the first component, the number of students who are admitted at a particular time *t* (e.g. a semester or quarter), which we shall denote by x_t. This is a random quantity which is not known with exactitude until the time period begins. A plausible statistical model for this random quantity is the Poisson Distribution. The Poisson Distribution has a long history and is often used for random quantities which are counts. Consider the process of obtaining new freshmen. The first step is to make prospective students aware of an academic institution. This is done through media advertising, use of mailing lists, personal visits to feeder institutions, etc. The total number of prospective students reached is unknown and but can be modeled by the Poisson Distribution with mean θ (for the Poisson distribution only one parameter is necessary since the standard deviation is $\sqrt{\theta}$). If we assume that each of the reached individuals requests further information with probability π_1, then conditional on the total number of reached individuals, the number requesting further in formation would follow the Binomial Distribution. This process of a total "pool" being generated by the Poisson Distribution followed by binomial distribution, conditional on the actual total, responding with a fixed probability is known as a compound Poisson process. It is shown in the Technical Appendix (Result 1) that a consequence of this is that the number of students requesting information will follow a Poisson distribution with parameter $\theta\pi_1$. Of the number who request further information, assume that the probability of applying for admission is π_2. Again assuming that the Binomial Distribution applies, the number applying for admission would follow the Poisson Distribution with parameter $\theta\pi_1\pi_2$. Letting π_3 represent the probability of an applicant being accepted, it follow that the number accepted follows a Poisson Distribution with

parameter $\theta\pi_1\pi_2\pi_3$. Finally, if π_4 is the probability that and accepted student actually enrolls, the number of new freshmen will follow the Poisson Distribution with parameter λ, where

$$\lambda = \theta\pi_1\pi_2\pi_3\pi_4 \qquad \textbf{(1).}$$

Equation (1) conforms to the well known fact that by increasing initial exposure (i. e. increasing the value of θ) or increasing any of the probabilities of requesting information, applying for admission, actually being admitted and enrolling (the π_i terms) the average new enrollment will increase. Since constant efforts are made to increase exposure, student response, student application, and the enrollment of admitted students (the process of admission is sometimes manipulated to make the probability of admission increase or decrease), it is likely that λ is not constant from time period to time period. Accordingly we will assume that the number of new students in a category, x_t follows the Poisson Distribution with mean λ_t.

The second component of enrollment is the number of students who entered in previous time periods and are still at the institution. Denote this number by y_t. If we denote the total number of students enrolled at time *t* by n_t, then at time *t + 1*, we have the result that

$$n_{t+1} = x_{t+1} + y_{t+1} \qquad \textbf{(2).}$$

We shall assume that a student leaves the university during time period *t* with probability p_t. It should be noted that those leaving include students graduating, students dropping out, students moving to other categories such as undergraduate to graduate or Master's level to Doctoral Level, and students who "stop out", that is temporarily do not enroll for a semester but intend to re-enroll at some later time period. Under this assumption, it is plausible to assume that that given the number of students enrolled in the previous time period, n_t, the number who will be present at time period $t + 1$ follows the Binomial distribution with probability $1 - p_t$.

The actual data series of enrollments $n_1, n_2, n_3, \ldots$ forms the observed time series of total enrollment. From institutional records it is possible to determine the values of x_t and y_t for any time period *t*. But to understand the statistical behavior of the enrollment time series, we must determine the basic statistical properties of the time series. By this we mean the expected value (or mean) and variance of n_t as well as the correlation between the enrollment values at adjacent time points *t* and *t+k*.

If we denote the theoretical mean of n_t by μ_t, it is shown in the Technical Appendix (Result 2) that the following relationship exists between the mean of the enrollment at times *t* and *t + 1*,

$$\mu_{t+1} = \lambda_{t+1} + \mu_t(1 - p_t) \qquad \textbf{(3).}$$

Equation (3) shows the dependence of the expected size of the enrollment on the two main drivers of enrollment, the expected number of new students (λ_{t+1}) and the proportion of students who leave the institution (p_t). Increasing the number of new enrollees will increase the enrollment as will decreasing the proportion that leave. In fact equation (3) can be rewritten in the form

$$\frac{\mu_{t+1} - \mu_t}{\mu_t} = \frac{\lambda_{t+1}}{\mu_t} - p_t$$

which implies that the relative change in expected enrollment is the difference between the relative size of the new enrollees (relative to the expected total enrollment at time *t)* less the proportion that leave. Simply put if the proportion of new students is greater than the proportion that leave, expected enrollment will increase, and vice versa.

The variance structure of the series is complicated. If we denote the variance of n_t by σ_t^2, then it is shown in the Technical Appendix (Result 3) that

$$\sigma_{t+1}^2 = \lambda_{t+1} + \mu_t p_t(1 - p_t) + \sigma_t^2(1 - p_t)^2 \qquad \textbf{(4).}$$

(Note that if the n_t follow the Poisson distribution, then $\sigma_t^2 = \mu_t$ and equation (4) simplifies to

$$\sigma_{t+1}^2 = \lambda_{t+1} + \sigma_t^2(1 - p_t)$$

which is the same recursion relationship as expressed by equation (3).)

Either form of equation (4) indicates that the series has changing variance from time period to time period. This invalidates the use of most standard statistical procedures which assume constant variance from time period to time period. Further complicating analysis is the fact that the observed enrollments are auto correlated, i.e. not independent of each other from time period to time period. In fact it is shown in the Technical Appendix (Results 4, 5 and 6) that the correlation between the number enrolled at time *t* and the number enrolled at time *t +k*, is given by

$$Correlation(n_t, n_{t+k}) = \frac{\sigma_t (1-p_t)(1-p_{t+1})\ldots(1-p_{t+k-1})}{\sigma_{t+k}} \quad \textbf{(5).}$$

Finally using equations (4) and (5), it follows that

$$Var(n_{t+k} - n_t) = \sigma_{t+k}^2 + \sigma_t^2 - 2\sigma_t^2 (1-p_t)(1-p_{t+1})\ldots(1-p_{t+k-1}) \quad \textbf{(6),}$$

which determines the variability of changes in the levels of enrollment from time period *t* to time period *t + k*.

Equilibrium

Enrollment will stabilize, i.e. cease to grow or contract, if the two key parameters, λ_t *and* p_t cease to change. (It is possible that stabilization could occur by λ_t *and* $(1-p_t)$ moving in different directions, for example increasing retention, $(1-p_t)$, to compensate for a decrease in enrollment λ_t, but this situation is unlikely to be sustainable for protracted periods of time). Accordingly, assume that at some point in time t_0 we have the situation that $p_t = p$ *and* $\lambda_t = \lambda$ and that this continues for all $t \geq t_0$. Then by substituting into equations (3) to (6) above, subsequent to t_0 the following relationships will hold

$$\begin{aligned}
&E(n_{t+k}) = \mu_{t+k} = \frac{\lambda}{p} \quad \text{for all } k \geq 0 \\
&Var(n_{t+k}) = \sigma_{t+k}^2 = \frac{\lambda}{p} \quad \text{for all } k \geq 0 \\
&Correlation(n_{t+k}, n_t) = (1-p)^k \quad \text{for all } k \geq 0 \\
&Var(n_{t+k} - n_t) = \frac{2\lambda}{p}[1-(1-p)^k] \quad \text{for all } k \geq 0
\end{aligned} \qquad \textbf{(7).}$$

Figure 1 below shows a simulation of a constant parameter process with $\lambda = 3{,}000$ and $p = .2$, so that that the mean of the series would be *15,000.*

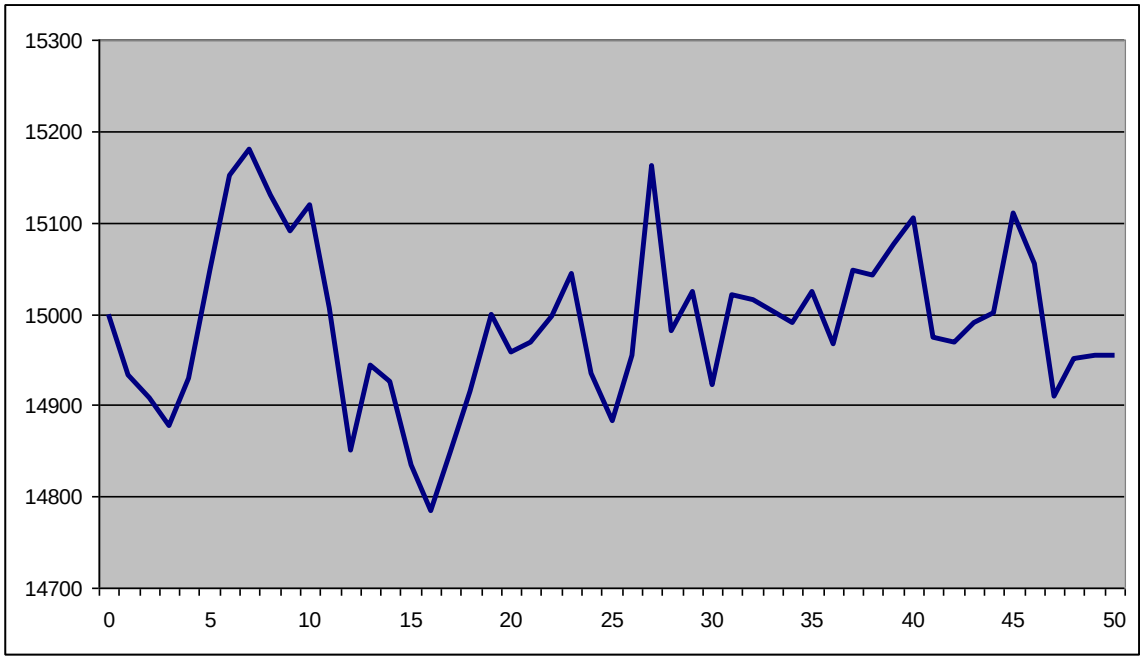


Figure 1. Simulation of an Equilibrium Series with $\lambda = 3{,}000$ and $p = .2$

Equations (7) imply that the enrollment process is in a no change situation with the variability due to the inherent uncertainties in the enrollment process as reflected by the stabilized variance and correlation structure of the the process. Since all the components of n_t follow the Poisson distribution and the sum of Poisson distributions also follows the Poisson distribution, it is to be expected that the mean and variance of n_t would be the same. What is important is the correlation structure of the series. Contrast the simulation in Figure 1, with that of Figure 2, the latter showing a time series of <u>uncorrelated</u> Poisson variables with mean *15,000.*

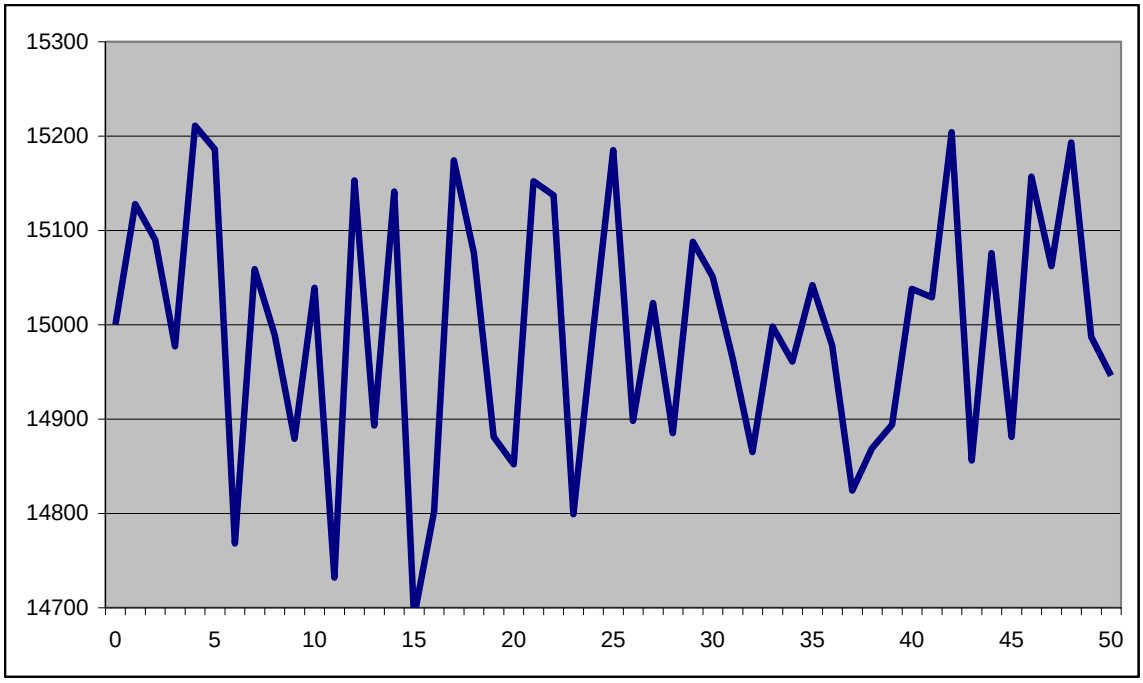


Figure 2. Simulation of Uncorrelated Poisson Variables with mean *15,000*

Although both series have the same mean and variance, the difference in correlation structure is very apparent with the uncorrelated series of Figure 2. showing greater period to period volatility than the data in Figure 1. which changes more smoothly from period to period. When the series of Figure 1 moves away from the mean of *15,000* it returns to the mean more smoothly than the more erratic period to period behavior of Figure 2.

The strong correlation structure of enrollment series means that the effects of changes in the parameters will not have rapid effects on the enrollment, rather, such change will gradually emerge over time. For example, suppose at time $t = 11$, the university was able to permanently increase the average admissions from *3,000* to *4,000*. By the equations above, with $p = .2$, this would mean that the total enrollment would increase to *20,000*. However, the increase would be gradual and not show up fully for almost 15 time periods. This is shown in Figure 3. , below, which simulates such a change. At first the rise is quite swift, but by $t=16$, the enrollment has only risen to *19,000* and then tapers, geometrically, slowly approaching *20,000*. It is shown in the Technical Appendix (Result 8), that if the average new enrollment changes permanently at time $t + 1$ from λ to $\lambda + \Delta_\lambda$, then at time $t + k$, the new mean will be

$$\mu_{t+k} = \frac{\lambda + \Delta_\lambda}{p} - \frac{\Delta_\lambda}{p}(1-p)^k \qquad \textbf{(8).}$$

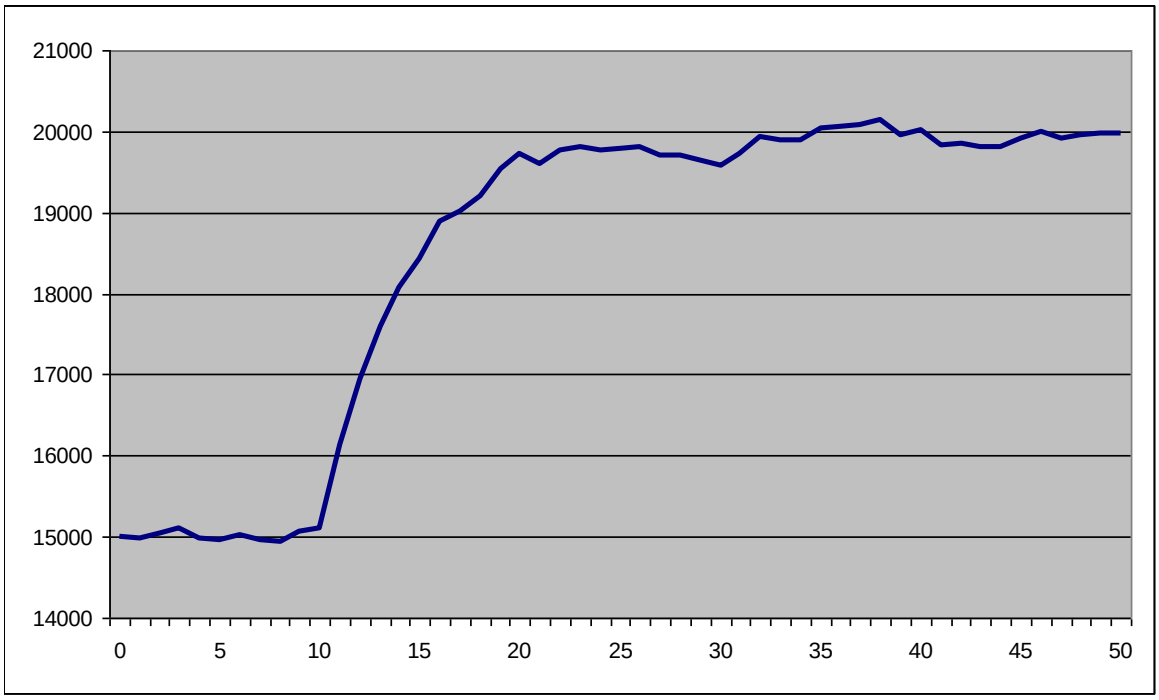


Figure 3. Change in Enrollment Average from *3,000* to *4,000* at *t=11, p =.2*

Further it can be shown (Result 7 in the Technical Appendix) that every permanent percent increase in average new enrollment will eventually result in the same percent increase in average total enrollment. Thus a *10*% increase in average new enrollment from *3,000* to *3,300*, will eventually result in a *10*% increase in average total enrollment from *15,000* to *16,500*. In headcount terms, every permanent increase of *1* new enrollee is eventually on average worth *1 / p* new total enrollees.

Similarly, Figure 4. shows a simulation of what happens when the retention rate changes from *.8* to *.85* (i.e. *p* changes from *.2* to *.15*) retaining a constant average new enrollment of *3,000*. As in Figure 3, there is at first a steep rise but then a slower geometric approach to the new equilibrium mean of *20,000*. It is shown in the Technical Appendix (Result 9) , that if the attrition decreases permanently at time *t + 1* from *p* to $p - \Delta_p$ (or alternatively the retention increases from *(1 – p)* to $(1 - p + \Delta_p)$), then at time *t + k*, the new mean will be given by the equation

$$\mu_{t+k} = \frac{\lambda}{p - \Delta_p}\left[1 - (1 - p + \Delta_p)^k \frac{\Delta_p}{p}\right] \qquad \textbf{(9).}$$

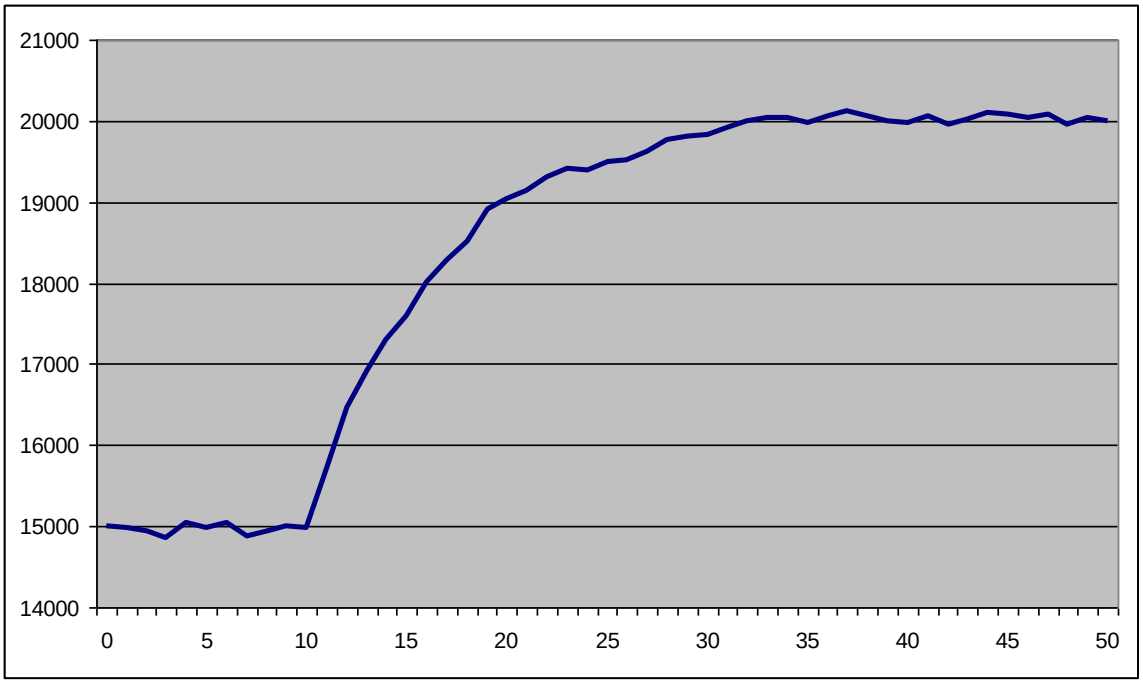


Figure 4. Change in Enrollment *p* from *.2* to *.15,* at *t=11,* mean = *3,000*

In this case *p* decreased from *20%* to *15%,* which is a relative decrease of *25% (.05/.20).* It is shown in the Technical Appendix (Result 10), that this will result in a relative increase in average total enrollment of *.25/(1-.25) = .3333.* Therefore average total enrollment will grow by *33.33%* from *15,000* to *20,000.* The effect of a small change in *p* can be quite profound. For example, imagine an enrollment situations where on average *3,700* new students apply each year and *p=.37* so that there is a very high attrition rate. By our previous formulas, this gives an equilibrium average total enrollment of *10,000.* If the attrition rate is reduced by just *.02* to *p=.35,* the average total enrollment will rise to *10,571* or by almost *6%.*

Using the Model for Statistical Analysis

The model above allows one to answer some basic questions about the enrollment series. First and paramount, is the simple question as to whether the change in total enrollment from time *t* to *t + 1* indicates that total enrollment has changed or whether it is likely just statistical fluctuation. More specifically suppose we want to test the statistical hypothesis:

$$H_0 : \mu_{t+1} = \mu_t = \mu$$
$$H_A : \mu_{t+1} \neq \mu_t$$

It is shown in the Technical Appendix (Result 11) that if H_0 is true and *p* is known, then the statistic

$$z = \frac{n_{t+1} - n_t}{\sqrt{(n_{t+1} + n_t)\,p}} \qquad \textbf{(10),}$$

is approximately normally distributed with mean *0* and standard deviation *1*. Accordingly to test H_0, at significance level $\alpha = .05$, one need only compute *z* and compare it to the limits ± 1.96. If *z* is within these limits, then one would accept H_0 and declare that no statistically significant change has occurred. Otherwise, one would reject H_0 and declare that a significant change has occurred.

Unfortunately *p*, may not always be available. However, the statistic

$$z_C = \frac{n_{t+1} - n_t}{\sqrt{n_{t+1} + n_t}} \qquad \textbf{(11),}$$

which does not require knowledge of *p*, can be used and is statistically conservative in the sense that if you reject H_0 using (11), you will also reject H_0 using (10), but not vice versa.

To illustrate the above, data on enrollment from The University of Texas at Dallas collected for Fall Semesters from 1998 to 2008 will be used. Figure 5, shows the data, both listed and graphically, for total undergraduate enrollment and the associated z_C statistics for each Fall semester. (z_C was used since *p* was not available). It is clear that the conservative z_C statistics are mirroring the pattern evident in Figure 5. Specifically, there is a strong rise in enrollment from Fall 1999 until Fall 2005 followed by no growth from Fall 2005 through Fall 2008.

Total UG Enrollment

Fall	n_t	z_C	Significant Change
1998	5,501		
1999	5,879	3.543	Yes
2000	6,439	5.046	Yes
2001	7,416	8.300	Yes
2002	7,969	4.458	Yes
2003	8,516	4.260	Yes
2004	8,904	2.940	Yes
2005	9,397	3.644	Yes
2006	9,193	-1.496	No
2007	9,096	-0.717	No
2008	9,202	0.784	No

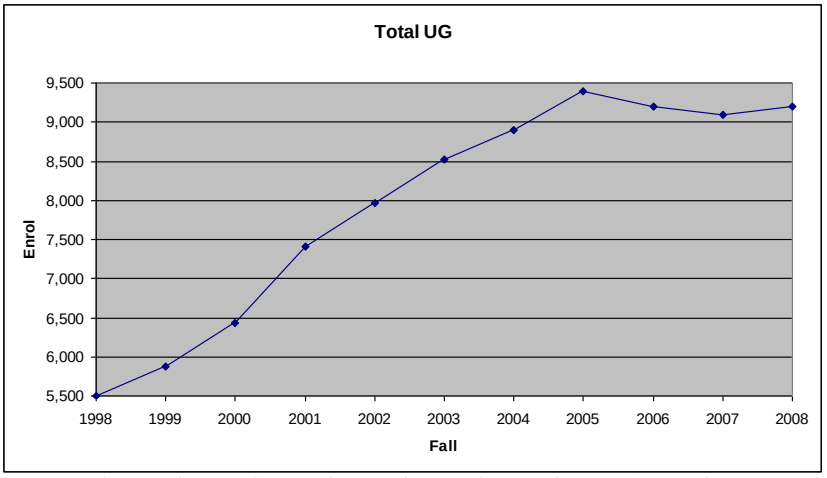


Figure 5. UG Enrollment UTD Fall 1998 – Fall 2008

Figure 6. below shows the Enrollment data for Master's students at UTD from Fall 1998 through Fall 2008. Again the z_C statistics mirror the pattern of the data. Specifically there is strong growth from Fall 1999 through Fall 2001. The series does not change significantly in the Fall of 2002 but drops, significantly, in the Fall of 2003 after which there has been no real change year to year change from Fall 2003 through Fall 2008. However, the period from Fall 2004 through Fall 2008 seems to indicate a possible recovery. In order to test this, we can use the last equation given in (7) which gives the $Var(n_{t+k} - n_t)$. If we assume the series is in equilibrium, then the statistic

$$z_C = \frac{n_{2008} - n_{2004}}{\sqrt{n_{2008} + n_{2004}}} = \frac{4705 - 4311}{\sqrt{4705 + 4311}} = 4.15,$$

is also conservative and when compared to the normal distribution with $\alpha = .05$, indicates that the recovery may be real since the value is statistically significant. This indicates that although the year by hear changes are not significant, the cumulative effect over the four year period may indicate a slow recovery.

Master's Enrollment

Fall	n_t	z_C	Significant Change
1998	3,519		
1999	3,714	2.293	Yes
2000	4,014	3.413	Yes
2001	4,760	7.964	Yes
2002	4,731	-0.298	No
2003	4,446	-2.975	Yes
2004	4,311	-1.443	No
2005	4,444	1.421	No
2006	4,410	-0.361	No
2007	4,547	1.448	No
2008	4,705	1.643	No

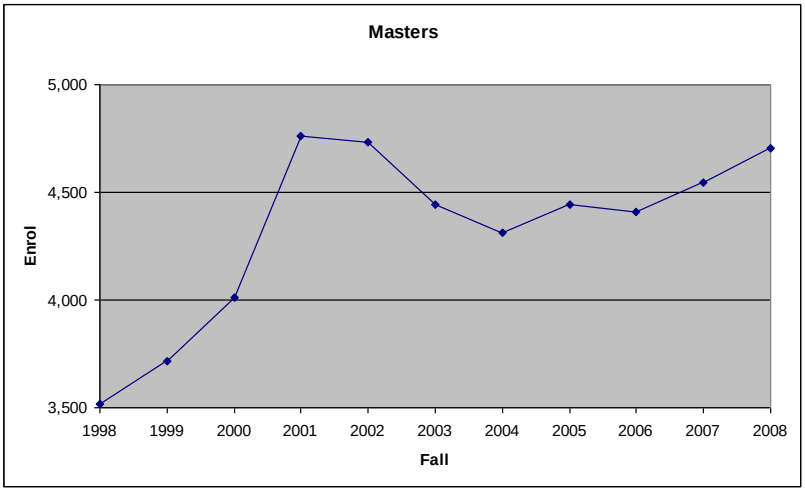


Figure 6. Master's Enrollment UTD Fall 1998-Fall 2008

Since we did not know the value of *p*, we have been forced to use the conservative *z* statistics. In some cases, it is possible to use data to estimate *p*. Equation (3) above, when *p* is constant from time period to time period can be written as

$$\mu_{t+1} = \lambda_{t+1} + \mu_t(1-p),$$

which suggests a regression like relationship between μ_{t+1} and μ_t with slope *(1 – p),* but with constantly changing intercept λ_{t+1}. This suggests that a similar relationship might hold for the relationship between n_{t+1} and n_t. If one could come up with an estimate of λ_{t+1}, say $\hat{\lambda}_{t+1}$, then it would not be unreasonable to expect that

$$\begin{aligned} & n_{t+1} = \hat{\lambda}_{t+1} + n_t(1-p) + \varepsilon_{t+1} \\ \Rightarrow \quad & n_{t+1} - \hat{\lambda}_{t+1} = n_t(1-p) + \varepsilon_{t+1}. \end{aligned}$$

Here ε_{t+1} would not have the properties usually assumed in standard regression analysis (uncorrelated with constant variance). However, it is well know that even with auto-correlated, heteroscedastic errors, least squares analysis will still give unbiased parameter estimates.

To illustrate the above, we will use data from the University of Texas at Dallas for Lower Level Undergraduates (the first two years of college) and for PhD students. As an estimate of $\hat{\lambda}_{t+1}$ we will use the value x_t, the number of new enrollees in the category in the previous time period since x_{t+1} would be unknown at time *t.* Figure 7. shows the data and results of performing such a regression. Since the regression was done without an intercept, the usual value of R^2 is not valid. Instead a pseudo R^2 is computed using the formula

$$Pseudo\ R^2 = 1 - \frac{\sum_t (n_t - \hat{n}_t)^2}{\sum_t (n_t - \bar{n})^2} = .8944.$$

This indicates very high predictability. From the slope estimate, we get that $\hat{p} = .4148$ which indicates that about 42% of the students leave the Lower Level

Lower Level Enrollment

Year	n_t	x_{t-1}	$n_t - x_{t-1}$	n_{t-1}	Forecast	error
1999	1,764	732	1,032	1,498	1,609	155.44
2000	1,992	914	1,078	1,764	1,946	45.78
2001	2,180	1112	1,068	1,992	2,278	-97.63
2002	2,596	1060	1,536	2,180	2,336	260.36
2003	2,826	1208	1,618	2,596	2,727	98.93
2004	2,865	1265	1,600	2,826	2,919	-53.65
2005	2,880	1204	1,676	2,865	2,880	-0.47
2006	2,710	1171	1,539	2,880	2,856	-146.25
2007	2,687	1144	1,543	2,710	2,730	-42.77
2008	2,683	1205	1,478	2,687	2,777	-94.32

$$(1-p) = .5852 \Rightarrow \hat{p} = .4148$$
$$Pseudo\ R^2 = .8944$$

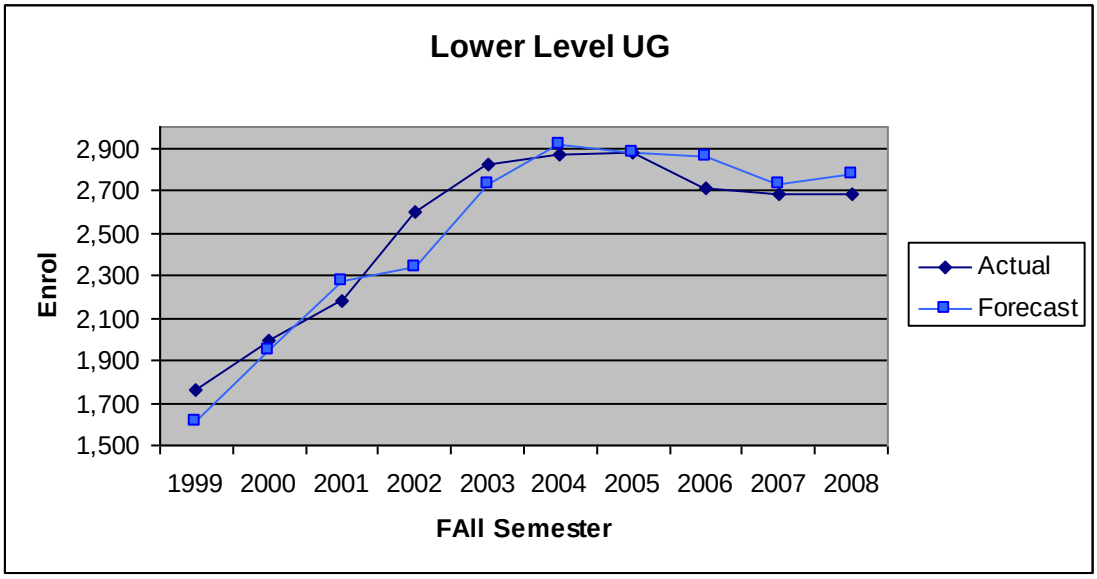


Figure 7. Data and Regression Results for UTD Dallas Lower Level Undergraduate Enrollment

undergraduate category each year. Assuming that this value of *p* is relatively stable, it follows that the number of years a student stays in the lower level category follows a geometric probability distribution with parameter *p*. The mean of the geometric distribution is given by *1 / p* which in this case is *1 / .4148 = 2.41*. This indicates that on average, a student spends about 2.41 years in the category of lower level undergraduate. This value conforms well with other estimate of the length of time a student stays in this category.

Forecasting Using the Model

When forecasting an enrollment series two distinct situations arise. The first is when the process is in equilibrium. In the case the governing equations for the system are given by Equation (7). That is the mean, variance, and correlation structure of the system is stable and not changing with time. Assuming you know the value of n_t, a simple forecast of n_{t+k} and an approximate 95% confidence interval on the forecast is, based on Equation (7),

$$n_t \pm 1.96\sqrt{2n_t[1-(1-p)^k]} \qquad \textbf{(12).}$$

Note that as you forecast further and further into the future, i.e. *k* gets larger, Equation (12) approaches a maximal forecast range of $\pm 1.96\sqrt{2n_t}$. However, for one period ahead the forecast range is much smaller and is given by $\pm 1.96\sqrt{2n_t p}$. Figure 8. below shows the one step ahead forecasted and confidence intervals for a simulated series with $\lambda = 3{,}000$, $p=.2$, so that $\mu = 15{,}000$. The dark squares in

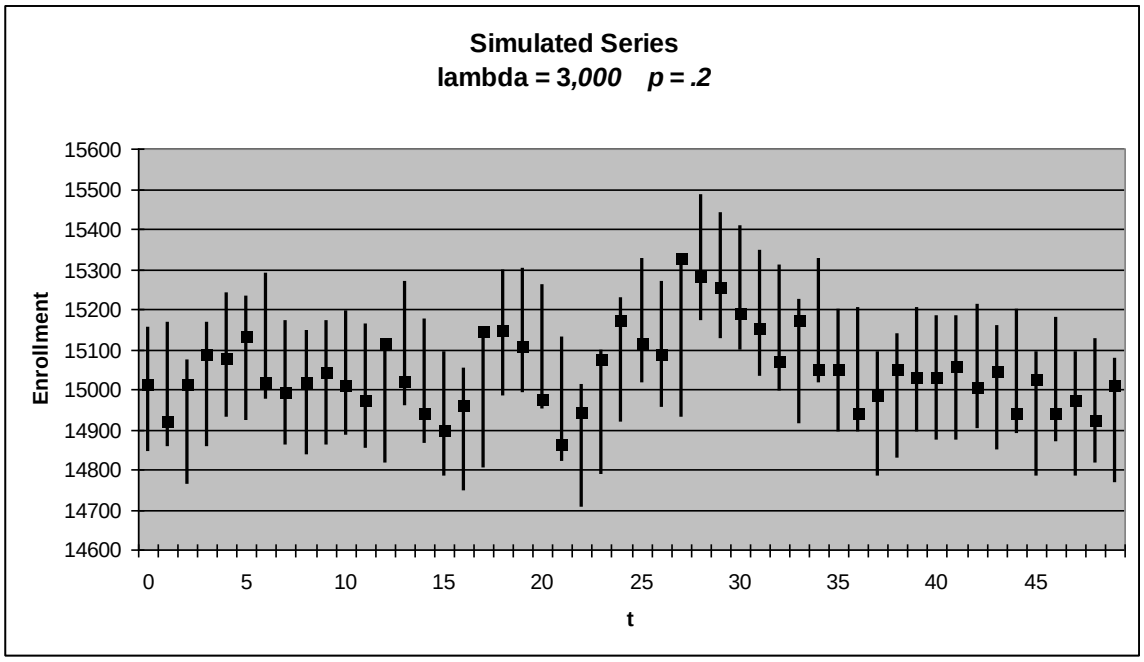


Figure 8. One Step Ahead Forecasts for a Series in Equilibrium

Figure 8. represent the actual values of the time series while the vertical lines represent the confidence interval. As can be seen, the confidence interval brackets the actual value most of the time. If one used the conservative limits of $\pm 1.96\sqrt{2n_t}$, the limits would have ranged between approximately *14,600* and *15,400* which would have been much too wide. Accordingly, knowledge of the parameter *p* is

quite important in forecasting the series. The relative error of the one step ahead forecast is given by the equation

$$\pm \frac{1.96\sqrt{2n_t p}}{n_t} = \pm 2.772\sqrt{\frac{p}{n_t}} \qquad \textbf{(13).}$$

In the simulated case above with $\mu = 15{,}000$ and $p = .2$, Equation (13) indicates that the one step ahead confidence interval is approximately $\pm$ *1 %* around the value at time *t* .

If the system is not in equilibrium, then it is still possible, under certain circumstances, to forecast the series. It is shown in the Technical Appendix (Result 13) that the following "regression like" relationship exists for the series

$$n_{t+!} = \lambda_{t+1} + n_t(1 - p_t) + \varepsilon_{t+1} \qquad \textbf{(14),}$$

where

$$Var(\varepsilon_{t+1}) = \lambda_{t+1} + \mu_t p_t(1 - p_t) \qquad \textbf{(15).}$$

The strength of this "regression like" relationship can be measured by an analogue to the usual regression coefficient of determination, specifically,

$$\rho^2 = 1 - \frac{Var(\varepsilon_t)}{Var(n_{t+1})} \ .$$

Using Equations (4) and (15), one arrives at the formula

$$\rho^2 = \frac{\sigma_t^2(1 - p_t)^2}{\lambda_{t+1} + \mu_t p_t(1 - p_t) + \sigma_t^2(1 - p_t)^2} \qquad \textbf{(16).}$$

The use of equation (14), is quite difficult since λ_{t+1} is not constant from time period to time period. If one can replace λ_{t+1} with an estimate $\hat{\lambda}_{t+1}$ and p_t is known, then equation (14) can be used for forecasting. To illustrate, we will look at Doctoral enrollment data at The University of Texas at Dallas. In this case we will use as our estimate of $\hat{\lambda}_{t+1} = x_t$ the number of new doctoral students in the previous year. It is shown in the Technical Appendix, Result 15, that with this estimate, the variance of the error $(n_{t+1} - \hat{n}_{t+1})$ where

$$\hat{n}_{t+!} = x_t + n_t(1 - p_t) \qquad \textbf{(17)}$$

is given by the equation

$$Var(n_{t+1} - \hat{n}_{t+1}) = \lambda_{t+1} + \mu_t p_t (1 - p_t) + \lambda_t \qquad \textbf{(18).}$$

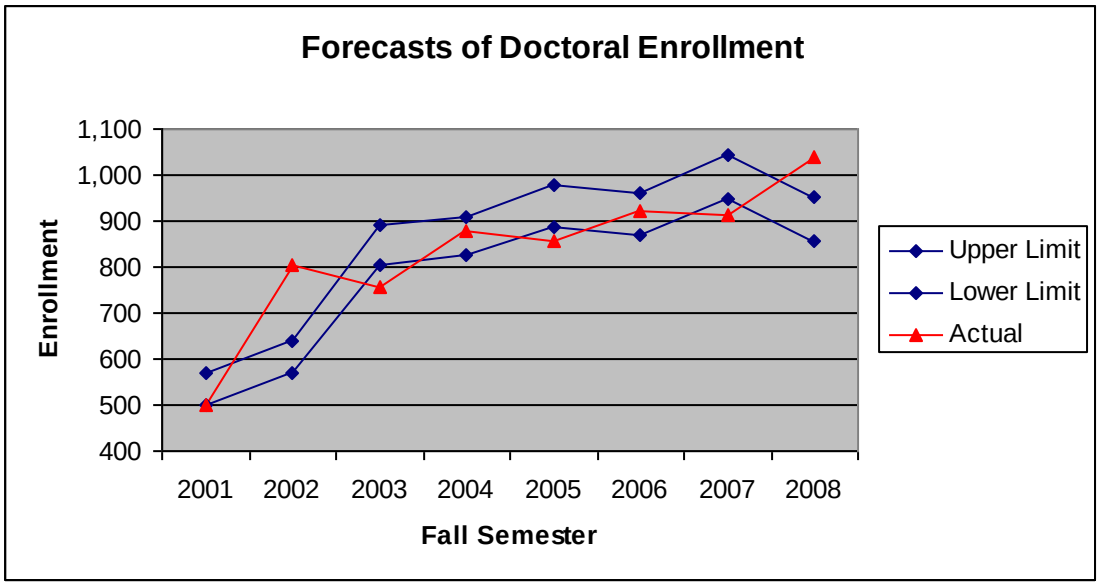
As an approximation, we will take $\hat{\lambda}_{t+1} = \hat{\lambda}_t = n_t p_t$ ***and*** $\hat{\mu}_t = n_t$**, which yields the estimate**

$$Var(n_{t+1} - \hat{n}_{t+1}) = n_t p_t (3 - p_t) \qquad \textbf{(19).}$$

From equations (17) and (19), it follows that an approximate 95% confidence interval on the enrollment at time ***t + 1,*** **is given by the limits**

$$x_t + n_t (1 - p_t) \pm 1.96 \sqrt{n_t p_t (3 - p_t)} \qquad \textbf{(20).}$$

These limits are shown in Figure 9. As can be seen, the forecasting intervals show a rather odd pattern of just enclosing the actual values or just missing them. The forecasting intervals seem to show a "high" followed by a "low" forecast. In the State of Texas, university funding is determined every two years. The legislative session is held in the summer with new funding for the next two years beginning in the following Fall semester. Since offers to new doctoral students are usually made in the late spring or early summer, this means that every two years there is uncertainty as to how much state funding will be available in the upcoming academic year. It is our conjecture, that university departments tend to be conservative in the number of doctoral offers in years when the legislature is in session since future funding is uncertain. In the subsequent year, when funding is known, the number of offers tends to be less conservative. This could induce a two year cycle, which is reflected in the number of new doctoral students enrolled. Since we used the previous year's new enrollment as our estimate of $\hat{\lambda}_{t+1}$**, this would put our forecasts out of synchronicity with the actual enrollment cycle and could be the source of the two year oscillating forecasting pattern.**



Year	x_{t-1}	n_{t-1}	Forecast	High	Low	n_t
2001	155	492	536	570	501	500
2002	216	500	603	638	568	805
2003	224	805	847	891	803	756
2004	282	756	867	910	824	877
2005	253	877	932	978	886	858
2006	253	858	917	962	872	920
2007	285	920	997	1,044	950	913
2008	197	913	904	950	857	1037

$$(1-p) = .7739 \Rightarrow \hat{p} = .2261$$

Figure 9. Forecasts of Doctoral Enrollment
The University of Texas at Dallas

Semester Credit Hours

The above analysis can be used to forecast and test hypotheses about changes in the numbers of semester credit hours generated. Let h_t denote the number of semester credit hours in an enrollment category generated by n_t students in that category. Then the credit hours and the enrollment are related by the formula

$$h_t = w_t n_t \qquad \textbf{(21),}$$

where w_t is the average number of credit hours taken by students in the enrollment category. The statistical structure of the semester credit hour series can then be inferred from the information about enrollment.

Let $\upsilon_t = E(h_t) = w_t \mu_t$, then to determine if there has been a change in the number of semester credit yours generated form time period *t* to *t+1,* is equivalent to testing the hypothesis

$$H_0 : \upsilon_{t+1} = \upsilon_t = \upsilon$$
$$H_A : \upsilon_{t+1} \neq \upsilon_t$$

However, if this hypothesis is rejected, it may be because enrollment has changed, or the number of hours taken per student has changed or both. The natural statistic to examine is

$$h_{t+1} - h_t = w_{t+1} n_{t+1} - w_t n_t$$

which by the Poisson nature of the n_t series would be approximately normally distributed with mean and variance given by

$$\begin{aligned} E(h_{t+1} - h_t) &= w_{t+1}\mu_{t+1} - w_t \mu_t \\ Var(h_{t+1} - h_t) &= w_{t+1}^2 \sigma_{t+1}^2 + w_t^2 \sigma_t^2 - 2w_{t+1} w_t \sigma_t^2 (1 - p_t) \end{aligned} \qquad \textbf{(22).}$$

If we assume that the average credit hours per student is the same in periods *t* and *t + 1,* and further that the enrollment is in equilibrium, Equations (22) becomes

$$\begin{aligned} E(h_{t+1} - h_t) &= 0 \\ Var(h_{t+1} - h_t) &= 2w^2 \mu p = 2w\upsilon p \end{aligned} \qquad \textbf{(23).}$$

Using an argument similar to that which led up to Equations (10) and (11), we can use the test statistic *Z* to test H_0 where

$$Z = \frac{h_{t+1} - h_t}{\sqrt{\left(\frac{w_{t+1} + w_t}{2}\right)(h_{t+1} + h_t)p}} \qquad \textbf{(24),}$$

where *Z* can be compared to a normal distribution with mean *0* and variance *1*. If H_0 is rejected and the test of steady enrollment is accepted, then this would imply that the credit hours per student had changed significantly. Further if H_0 is accepted and the test for steady enrollment is rejected, then this would also imply that the credit hours per student had changed significantly. If *p* is not available, then the conservative statistics

$$Z_C = \frac{h_{t+1} - h_t}{\sqrt{\left(\frac{w_{t+1} + w_t}{2}\right)(h_{t+1} + h_t)}} \qquad \textbf{(25),}$$

can be used. Figure 10. shows the total semester credit hour generation at all student levels for The University of Texas at Dallas for the Fall semesters from 1998 through 2008. The graph and the Z_C statistic shows significant year to year growth in credit hours except for the period from Fall 2006 to Fall 2007. Also shown in the figure is the z_C which tests for significant enrollment growth. This statistic indicates that for the period Fall 2002 to Fall 2003, and also Fall 2005 to Fall 2006, the growth in enrollment was not significant even though the growth in credit hours was. This indicates that for these two periods, there were significant changes in the number of credit hours per student. Indeed from Fall 2002 to Fall 2003, the average credit hours per student went from 9.66 to 10. 25 and from Fall 2005 to Fall 2006 the average credit hours per student went from 10.15 to 10.53. It is conjectured that changes in the advising programs at UTD had a positive effect on the number of hours successfully attempted by students.

t	n_t	z_C	Enrollment Change	h_t	w_t	Z_C	Credit Hour Change
1998	9,517			86,439	9.08		
1999	10,101	4.17	Yes	94,199	9.33	6.02	Yes
2000	10,945	5.82	Yes	104,398	9.54	7.45	Yes
2001	12,676	11.26	Yes	120,378	9.50	10.93	Yes
2002	13,505	5.12	Yes	130,439	9.66	6.49	Yes
2003	13,718	1.29	No	140,559	10.25	6.16	Yes
2004	14,092	2.24	Yes	144,576	10.26	2.35	Yes
2005	14,699	3.58	Yes	149,208	10.15	2.68	Yes
2006	14,523	-1.03	No	152,902	10.53	2.09	Yes
2007	14,556	0.19	No	153,115	10.52	0.12	No
2008	14,944	2.26	Yes	156,768	10.49	2.02	Yes

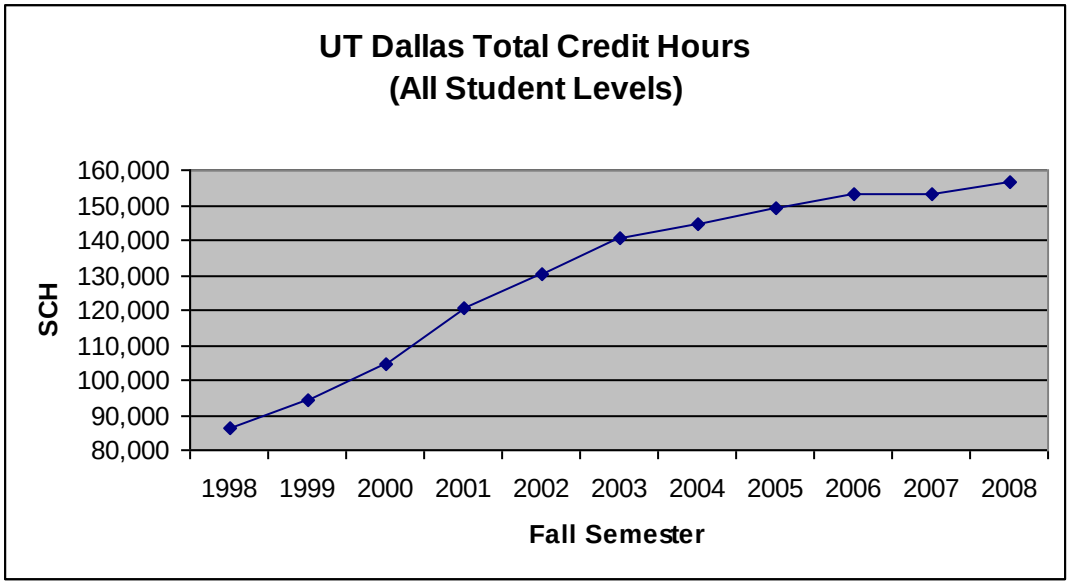


Figure 10. UT Dallas Total Credit Hour Growth

From equation (21), further statistical properties of h_t can be inferred. For example from Equations (21), (14) and (15), it follows that

$$h_{t+1} = (\lambda_{t+1} w_{t+1}) + \frac{w_{t+1}}{w_t}(1 - p_t) h_t + e_{t+1} \qquad (26),$$
$$= a_{t+1} + b_t h_t + e_{t+1}$$

where

$$Var(e_{t+1}) = w_{t+1}^2 (\lambda_{t+1} + \mu_t p_t (1 - p_t)) \qquad (27).$$

Therefore the semester credit hour series shows an auto-regressive structure similar to that of the enrollment series and if suitable information is available, regression like forecasts may be generated.

Variability in Retention

It is clear that retention, and thus the parameter *p,* may vary from school to school or within discipline in a school. For example, the undergraduate retention in the discipline of Computer Science may be lower than that of the discipline of business since the former students may have a poorer idea of the field basing their original major choice more on past "use" of a computer rather than on developing and writing programs. Our methodology tacitly assumes that the parameter *p* is constant at a student level, say lower level undergraduate, and does not take into account variability between disciplines. If one has *K* disciplines with m_i students in discipline *i* and the retention rate is $1-p_i$ within discipline *i,* then our previous analysis has treated the whole group as following the binomial distribution with parameters

$$m_+ = \sum_{i=1}^{K} m_i \text{ and } \bar{p} = \sum_{i=1}^{K} p_i / m_+ .$$

The pooled Binomial distribution, we would use, would have mean $m_+ \bar{p}$ and variance $m_+ \bar{p}(1-\bar{p})$. The correct distribution would be mixture of binomials (which is not necessarily binomial unless all the p_i are the same) with mean $\sum_{i=1}^{K} m_i p_i$ and variance $\sum_{i=1}^{K} m_i p_i (1-p_i)$. The mean values of the two distributions are the same and it is shown in the Technical Appendix(Result 16) that using the pooled binomial distribution overestimates the variance by $\sum_{i=1}^{K} m_i (p_i - \bar{p})^2$. For a deviation of *.05,* for example for $p_i's$ ranging from *.30* to *.25* , (which is quite sizable), would introduce an error of less than $.0025 m_+$, which is relatively small compared to m_+. Accordingly, we do not think that this small correction substantially alters the results above.

Conclusion

It is hoped that by illustrating the vital roles of the parameters λ_t, $1 - p_t$, and w_t, that is the average rate at which new students are enrolled, the retentions of enrolled students and the average load, on the size of the student populations, that better forecasts and enrollment management will be facilitated.

The Statistical Anatomy of Academic Enrollment Data
Wiorkowski & Redlinger
Technical Appendix

Result 1

If *n* is distributed as a Poisson random variable with parameter λ, and conditional on *n*, *x* is distributed as a Binomial random variable with parameters *n* and π , then *x* is unconditionally distributed as a Poisson random variable with parameter λπ.

Proof:

The joint distribution of *n* and *x* is given by,

$$p(n,x)=\frac{e^{-\lambda}\lambda^{n}}{n!}\frac{n!}{x!(n-x)!}\pi^{x}(1-\pi)^{n-x},$$

on the range $n = 0, 1, 2, \ldots\ldots, \infty;$ **and** $x \le n.$

The unconditional distribution of *x* is then given by

$$p(x)=\left(\frac{\pi}{1-\pi}\right)^{x}\frac{e^{-\lambda}}{x!}\sum_{n\ge x}\frac{\left[\lambda(1-\pi)\right]^{n}}{(n-x)!}.$$

By substituting $y=n-x$, and therefore $n = x + y$, one obtains

$$p(x)=\frac{(\lambda\pi)^{x}e^{-\lambda}}{x!}\sum_{y=0}^{\infty}\frac{\left[\lambda(1-\pi)\right]^{y}}{y!}=\frac{(\lambda\pi)^{x}e^{-\lambda}}{x!}e^{\lambda(1-\pi)}=\frac{(\lambda\pi)^{x}e^{-\lambda\pi}}{x!},$$

which implies that *x* is distributed as a Poisson random variable with parameter λπ.

Result 2

We assume that $n_{t+1}=x_{t+1}+y_{t+1}$ **where** x_{t+1} **follows the Poisson distribution with parameter** λ_{t+1}**, and** y_{t+1} **given** n_t **is distributed as the Binomial distribution with parameters** n_t **and** $1-p_t$**. It therefore follows that,**

$$\mu_{t+1} = E(n_{t+1}) = E_{n_t}(E(n_{t+1} \mid n_t)$$
$$= E_{n_t}[E(x_{t+1} \mid n_t) + E(y_{t+1} \mid n_t)]$$
$$= E_{n_t}[\lambda_{t+1} + n_t(1-p_t)]$$
$$= \lambda_{t+1} + \mu_t(1-p_t).$$

Here it is assumed that x_{t+1} is independent of n_t, and the symbol $E_{n_t}(...)$ denotes expectation with respect to the distribution of n_t.

Result 3

The variance of n_{t+1} can be written as

$$\sigma^2_{t+1} = Var(n_{t+1}) = E_{n_t}[Var(n_{t+1} \mid n_t)] + Var_{n_t}[E(n_{t+1} \mid n_t)]$$
$$= E_{n_t}[\lambda_{t+1} + n_t p_t(1-p_t)] + Var_{n_t}[\lambda_{t+1} + n_t(1-p_t)]$$
$$= \lambda_{t+1} + \mu_t p_t(1-p_t) + \sigma^2_t(1-p_t)^2$$

based on the means and variances of the Poisson and Binomial distributions and where the notation $Var_{n_t}(....)$ denotes the variance with respect to the distribution of n_t.

Result 4

If y_1 is Binomial with parameters *n* and p_1, and $y_2 \mid y_1$ is Binomial with parameters $n - y_1$ *and* p_2, then the random variable $n - y_1 - y_2$ follows the Binomial distribution with parameters *n* and $(1-p_1)(1-p_2)$.

Proof:

The joint distribution of y_1 and y_2 is given by

$$p(y_1, y_2) = \binom{n}{y_1} p_1^{y_1}(1-p_1)^{n-y_1} \binom{n-y_1}{y_2} p_2^{y_2}(1-p_2)^{n-y_1-y_2}$$

on the range $0 \le y_1 \le n$ ***and*** $0 \le y_2 \le n - y_1$. ***Letting*** $z = y_1 + y_2$ ***so that*** $y_2 = z - y_1$, **it follows that**

$$p(y_1, z) = \binom{n}{z} p_1^{y_1} (1-p_1)^{n-y_1} \binom{z}{y_1} p_2^{z-y_1} (1-p_2)^{n-z}$$

on the range $0 \le z \le n$ *and* $0 \le y_1 \le z$. **By regrouping terms, it follows that the distribution of *z* is given by**

$$p(z) = \binom{n}{z} (1-p_1)^n p_2^z (1-p_2)^{n-z} \sum_{y_1=0}^{z} \binom{z}{y_1} \left[\frac{p_1}{p_2(1-p_1)} \right]^{y_1} .$$

The summation above is the binomial representation of the term $(a+b)^n$ **with** $a = \frac{p_1}{p_2(1-p_1)}, b = 1,$ *and* $n = z$ **. This yields the expression**

$$p(z) = \binom{n}{z} (1-p_1)^n p_2^z (1-p_2)^{n-z} \left(\frac{p_1}{p_2(1-p_1)} + 1 \right)^z$$

which upon manipulation yields,

$$p(z) = \binom{n}{z} [1-(1-p_1)(1-p_2)]^z [(1-p_1)(1-p_2)]^{n-z} .$$

This implies that $z = y_1 + y_2$ **is Binomial with parameters** n *and* $1-(1-p_1)(1-p_2)$**, so that** $n - z = n - y_1 - y_2$ **is Binomial with parameters** n *and* $(1-p_1)(1-p_2)$ **QED.**

<u>Result 5</u>

Let $S_{t,t+k}$ **be the survivors of the** n_t **objects present at time *t* who are still in the system at time *t* + *k*, then given** n_t **,** $S_{t,t+k}$ **follows the Binomial distribution with parameters** n_t **and probability** $\prod_{i=0}^{k-1} (1-p_{t+i})$ **.**

Proof:

Apply mathematical induction to Result 4 with $z = \sum_{i=1}^{k} y_i$ **QED**

<u>Result 6</u>

Let $D_{t+1,t+k}$ be the survivors of the objects which entered the system between times $t+1$ and $t+k-1$ who are still in the system at time $t+k$, that is

$$D_{t+1,t+k} = \sum_{i=1}^{k-1} S_{t+i,t+k}$$

Then

$$n_{t+k} = x_{t+k} + D_{t+1,t+k} + S_{t,t+k} \ ,$$

where x_{t+k} are new entries to the system at time $t+k$. By the assumption of the model, all three components of n_{t+k} are independent of each other and therefore uncorrelated. It follows that

$$Cov(n_{t+k}, n_t) = Cov(x_{t+k} + D_{t+1,t+k} + S_{t,t+k}, n_t) = Cov(S_{t,t+k}, n_t) \ ,$$

Since there is no correlation between objects which have entered the system after time t with those that were in the system at time t.

Letting $\Pi_{t,t+k} = \prod_{i=0}^{k-1} (1 - p_{t+i})$, from Result 5 we have

$$\begin{aligned} Cov(S_{t,t+k}, n_t) &= E_{n_t}(n_t E(S_{t,t+k} \mid n_t) - E(n_t) E_{n_t}(S_{t,t+k} \mid n_t) \\ &= E_{n_t}(n_t^2 \Pi_{t,t+k}) - \mu_t E_{n_t}(n_t \Pi_{t,t+k}) \\ &= (\sigma_t^2 + \mu_t^2)\Pi_{t,t+k} - \mu_t^2 \Pi_{t,t+k} \\ &= \sigma_t^2 \Pi_{t,t+k} \end{aligned}$$

Therefore,

$$Cov(n_{t+k}, n_t) = \sigma_t^2 \prod_{i=0}^{k-1} (1 - p_{t+i}) \ ,$$

From which the correlation can be computed by dividing by the product $\sigma_t \sigma_{t+k}$.

Result 7

Since $\dfrac{\dfrac{(\lambda+\Delta_\lambda)}{p}-\dfrac{\lambda}{p}}{\dfrac{\lambda}{p}}=\dfrac{\Delta_\lambda}{\lambda}$,

it follows that the percentage relative change in the average new enrollment is the same as the percentage relative change in the average total enrollment.

Result 8

Assume that the up till and including time *t* the process has parameters λ *and* p**, and that for all time** $t+k, k \geq 1$ **the process has parameters** $\lambda+\Delta_\lambda$ *and* p**, then it follows that**

$$\mu_{t+1} = \lambda + \Delta_\lambda + \mu_t(1-p) = \lambda + \Delta_\lambda + \frac{\lambda}{p}(1-p)$$

$$\mu_{t+2} = (\lambda+\Delta_\lambda) + (\lambda+\Delta_\lambda)(1-p) + \frac{\lambda}{p}(1-p)^2$$

$$\vdots$$

$$\mu_{t+k} = (\lambda+\Delta_\lambda)\sum_{i=0}^{k-1}(1-p)^i + \frac{\lambda}{p}(1-p)^k$$

$$\Rightarrow \mu_{t+k} = \frac{\lambda+\Delta_\lambda}{p}[1-(1-p)^k] + \frac{\lambda}{p}(1-p)^k$$

$$\Rightarrow \mu_{t+k} = \frac{\lambda+\Delta_\lambda}{p} - \frac{\Delta_\lambda}{p}(1-p)^k \quad QED$$

Result 9

Assume that the up till and including time *t* the process has parameters λ *and* p**, and that for all time** $t+k, k \geq 1$ **the process has parameters** λ *and* p-Δ_p**, then it follows that**

$$\mu_{t+1} = \lambda + \mu_t [1-(p-\Delta_p)] = \lambda + \frac{\lambda}{p}(1-p+\Delta_p)$$

$$\mu_{t+2} = \lambda + \lambda(1-p+\Delta_p) + \frac{\lambda}{p}(1-p+\Delta_p)^2$$

$$\vdots$$

$$\mu_{t+k} = \lambda \sum_{i=0}^{k-1} (1-p+\Delta_p)^i + \frac{\lambda}{p}(1-p+\Delta_p)^k$$

$$\Rightarrow \mu_{t+k} = \frac{\lambda}{p-\Delta_p}[1-(1-p+\Delta_p)^k] + \frac{\lambda}{p}(1-p+\Delta_p)^k$$

$$\Rightarrow \mu_{t+k} = \frac{\lambda}{p-\Delta_p} - (1-p+\Delta_p)^k \left(\frac{\lambda}{p-\Delta_p} - \frac{\lambda}{p} \right)$$

$$\Rightarrow \mu_{t+k} = \frac{\lambda}{p-\Delta_p}[1-(1-p+\Delta_p)^k \frac{\Delta_p}{p}] \qquad QED$$

Result 10

Since $$\frac{\frac{\lambda}{p-\Delta_p} - \frac{\lambda}{p}}{\frac{\lambda}{p}} = \frac{\Delta_p}{p-\Delta_p} = \frac{\frac{\Delta_p}{p}}{1-\frac{\Delta_p}{p}}$$

a relative decrease in attrition of Δ_p / p, results in a relative increase in the average total enrollment of $(\Delta_p / p)/[1-(\Delta_p / p)]$.

Result 11

In order to test the hypothesis

$$H_0 : \mu_{t+1} = \mu_t = \mu$$
$$H_A : \mu_{t+1} \neq \mu_t$$

assume the series is in equilibrium with constant mean μ, so that from equation (7) it follows that

$$\mu = \frac{\lambda}{p}$$

$$E(n_{t+1} - n_t \mid H_0) = \mu - \mu = 0$$

$$Var(n_{t+1} - n_t) = \frac{2\lambda}{p}[1-(1-p)^1] = 2\lambda = 2\mu p$$

Since both n_{t+1} and n_t follow the Poisson Distribution, and can be approximated by the Normal Distribution, it follows that the difference $n_{t+1} - n_t$ can be approximated by a Normal Distribution which under H_0 would have mean *0* and variance $2\mu p$. It then follows that the statistic

$$z = \frac{n_{t+1} - n_t}{\sqrt{2\lambda}}$$

approximately follows the normal distribution with mean *0* and standard deviation *1*. Unfortunately, λ is usually unknown and must be replaced by an estimate $\hat{\lambda}$. Accordingly, the practical statistic to be used to test H_0 is

$$z = \frac{n_{t+1} - n_t}{\sqrt{2\hat{\lambda}}}$$

which, based on the large size of n_t, should still approximately follow the normal distribution with mean *0* and standard deviation *1*.

Several possible estimates of $\hat{\lambda}$ can be made, depending on what information is available. For example is the number of new enrollees x_t and x_{t+1} are available, then

$$\hat{\lambda} = \frac{x_{t+1} + x_t}{2},$$

is the minimum variance unbiased estimate of $\hat{\lambda}$, under the assumption that λ is the same at time period *t* and *t+1*.

If only x_{t+1} is available, then one could take

$$\hat{\lambda} = x_{t+1}.$$

If neither x_t nor x_{t+1} is available, then one can compute z as

$$z = \frac{n_{t+1} - n_t}{\sqrt{2\hat{\mu}p}}.$$

Under the assumption that the series is in equilibrium, it is natural to look at an estimator of the form

$$\hat{\mu} = qn_t + (1-q)n_{t+1} \quad with\ 0 \le q \le 1.$$

Such an estimator is unbiased and has variance

$$\begin{aligned} Var(\hat{\mu}) &= q^2\sigma^2 + (1-q)^2\sigma^2 + 2q(1-q)\sigma^2(1-p) \\ &= \sigma^2 - 2q(1-q)\sigma^2 p \end{aligned}$$

The variance is minimized by taking $q = .5$, so that the test statistic becomes

$$z = \frac{n_{t+1} - n_t}{\sqrt{(n_{t+1} + n_t)p}} \quad .$$

Finally, if p is not available, the statistic

$$z_C = \frac{n_{t+1} - n_t}{\sqrt{n_{t+1} + n_t}}$$

can be used as a conservative test of H_0. This is conservative in the sense that if z_C indicates a significant chant, then so also will z (but not vice versa).

<u>Result 12</u>

The terms $\lambda_{t+1} + n_t(1-p_t)$ and $[n_{t+1} - \lambda_{t+1} - n_t(1-p_t)]$ are uncorrelated.

Proof:

Since λ_{t+1} is a constant, we have from Result (6)

$$\begin{aligned} &Cov(\lambda_{t+1} + n_t(1-p_t),\ [n_{t+1} - \lambda_{t+1} - n_t(1-p_t)]) \\ &\quad = Cov(n_t(1-p_t),\ [n_{t+1} - n_t(1-p_t)]) \\ &\quad = Cov(n_{t+1}, n_t)(1-p_t) - Var(n_t(1-p_t)) \\ &\quad = \sigma_t^2(1-p_t)^2 - \sigma_t^2(1-p_t)^2 = 0 \end{aligned}$$

QED

Result 13

From Results 3 and 12, it follows that

$$\begin{aligned} Var(n_{t+1}) &= \sigma_{t+1}^2 = \lambda_{t+1} + \mu_t p_t (1-p_t) + \sigma_t^2 (1-p_t)^2 \\ &= Var(\lambda_{t+1} + n_t(1-p_t)) + Var(n_{t+1} - \lambda_{t+1} - n_t(1-p_t)) \\ &= \sigma_t^2 (1-p_t)^2 + Var(n_{t+1} - \lambda_{t+1} - n_t(1-p_t)) \end{aligned}$$

Which by matching terms, implies that

$$Var(n_{t+1} - \lambda_{t+1} - n_t(1-p_t)) = \lambda_{t+1} + \mu_t p_t (1-p_t) .$$

Result 14

If one takes $\hat{n}_{t+1} = \hat{\lambda}_{t+1} + n_t(1-p_t)$**, then**

$$\begin{aligned} Var(n_{t+1} - \hat{n}_{t+1}) &= \lambda_{t+1} + \mu_t p_t (1-p_t) + Var(\hat{\lambda}_{t+1}) \\ &\quad -2Cov(n_{t+1}, \hat{\lambda}_{t+1}) + 2(1-p_t) Cov(n_t, \hat{\lambda}_{t+1}) \end{aligned}$$

Proof:

Since $n_{t+1} - \hat{\lambda}_{t+1} - n_t(1-p_t) = n_{t+1} - \lambda_{t+1} - n_t(1-p_t) - (\hat{\lambda}_{t+1} - \lambda_{t+1})$**,**

it follows that

$$\begin{aligned} Var(n_{t+1} - \hat{n}_{t+1}) &= Var(n_{t+1} - \lambda_{t+1} - n_t(1-p_t)) + Var(\hat{\lambda}_{t+1}) \\ &\quad -2Cov(n_{t+1}, \hat{\lambda}_{t+1}) + 2(1-p_t) Cov(n_t, \hat{\lambda}_{t+1}) \end{aligned}$$

Then using Result 13, the result follows. QED

Result 15

If one takes $\hat{\lambda}_{t+1} = x_t$**, then**

$$Var(n_{t+1} - \hat{n}_{t+1}) = \lambda_{t+1} + \mu_t p_t (1 - p_t) + \lambda_t$$

Proof:

Since $Cov(n_t, x_t) = Var(x_t) = \lambda_t$, and by an argument similar to that used in Result 6, $Cov(n_{t+1}, x_t) = \lambda_t (1 - p_t)$, it follows that the two covariance terms of Result 14 cancel each other out. QED

Results 16

If $m_+ = \sum_{i=1}^{K} m_i$ and $\bar{p} = \sum_{i=1}^{K} p_i / m_+$, then

$$m_+ \bar{p}(1 - \bar{p}) = \sum_{i=1}^{K} m_i p_i (1 - p_i) + \sum_{i=1}^{K} m_i (p_i - \bar{p})^2 .$$

Proof:

$$m_+ \bar{p}(1 - \bar{p}) = m_+ \bar{p} - m_+ \bar{p}^2$$

and

$$\sum_{i=1}^{K} m_i p_i (1 - p_i) = \sum_{i=1}^{K} m_i p_i - \sum_{i=1}^{K} m_i p_i^2 = m_+ \bar{p} - \sum_{i=1}^{K} m_i p_i^2 ,$$

From which it follows that

$$m_+ \bar{p}(1 - \bar{p}) - \sum_{i=1}^{K} m_i p_i (1 - p_i) = \sum_{i=1}^{K} m_i p_i^2 - m_+ \bar{p}^2$$

$$= \sum_{i=1}^{K} m_i (p_i - \bar{p})^2$$

QED.