Chapter 6
Exemplar Use-Cases for Training Teachers on Learning Analytics

Arvind W. Kiwelekar, Manjushree D. Laddha, and Laxman D. Netak

Abstract Learning Analytics (LA) provides a rich set of methods, techniques, and tools to analyze learners' data. However, educators without a background in data analysis and statistical methods experience difficulty comprehending the potentials and pitfalls of learning analytics based pedagogical practices and Engineering Sciences experience this difficulty. This chapter documents a set of exemplars used to demonstrate learning analytics applications in daily classroom activities. These exemplars have been designed and used mainly to train newly recruited teachers on data analysis methods during faculty induction programs. Exemplars demonstrate the application of statistical methods such as hypothesis testing, analysis of variance (ANOVA), correlation analysis, and regression analysis. Each use case's broad objective is to describe the application's context so that teachers can apply it in a similar situation. The chapter provides ready-to-use examples for conducting teachers training programs on Learning Analytics.

Keywords Learning analytics · Hypothesis testing · Analysis of Variance (ANOVA) · Correlation analysis · Regression analysis

6.1 Introduction

Education domain is currently experiencing two kinds of changes on a large scale. The first change is driven by advancement in digital technologies, which emphasize the deployment of online courses for effective course delivery [1]. Evidence-based

A. W. Kiwelekar (✉) · M. D. Laddha · L. D. Netak
Department of Computer Engineering, Dr. Babasaheb Ambedkar Technological University, Lonere, Raigad 402103, India
e-mail: awk@dbatu.ac.in

M. D. Laddha
e-mail: mdladdha@dbatu.ac.in

L. D. Netak
e-mail: ldnetak@dbatu.ac.in

M. Ivanović et al. (eds.), *Handbook on Intelligent Techniques in the Educational Process*, Learning and Analytics in Intelligent Systems 29,
https://doi.org/10.1007/978-3-031-04662-9_6

reforms drive the second change [2], which suggests adopting data-centric teaching methods. Both changes complement each other because educators can use the data generated from learning platforms for producing evidence.

Mainly such evidence is used to claim the attainment of learning outcomes. Data needs to be analyzed through appropriate methods from Statistics or Machine Learning to extract helpful information demonstrating learning outcomes. However, many teachers lack the required knowledge of these methods from Statistics and Machine Learning. Hence they struggle to adopt such practices and consequently failing to justify the learning outcomes. To overcome this problem, we identify a few exemplar use-cases to demonstrate applications of various statistical techniques.

These use-cases highlight appropriate situations for the use of each statistical method. Four different statistical methods called (i) Hypothesis Testing, (ii) ANOVA, (iii) Correlation Analysis and (iv) Regression Analysis are described in this chapter. These exemplars have been used to introduce Learning Analytics and Statistical methods in various teachers training programs. They have found it helpful and practical to explain the nuances behind applying statistical methods to analyze learning data in different scenarios.

6.2 A Use Case for Hypothesis Testing

Hypothesis testing is one of the most useful statistical methods commonly employed to justify a claim using collected data. Here, a claim or a belief hold by a person (e.g. a teacher, a Head of the Department) or an institute (e.g. a particular University) is formulated as a hypothesis. For example, a piece of news in the Times of Higher Education dated 23rd December 2020 claims that *Students cheating exploding in Covid era.*

The validity of such claims can be tested through the statistical method of hypothesis testing. It consists of formulating two kinds of statements called *Null Hypothesis* and *Alternative Hypothesis.* A null hypothesis assumes that there is no relationship between two variables. An example of the null hypothesis can be *there is no connection between Covid outbreak and students cheating.* The null and alternative hypothesis is formulated using population parameters such as *average* and *proportion.*

6.2.1 Method

The following steps are typically performed during hypothesis testing:

1. Describe the hypothesis in words. It is described in terms of population parameters such as averages and proportions.
2. Define null H0 and alternative H1 hypothesis.

3. Identify the test statistic to be used for testing the validity of the null hypothesis. Usually, the *Z-statistics* is used when we know the standard variance of the population and population has normal distribution. The *t-statistics* is used when we do not know the value of standard variance of the population.
4. Decide the criteria for rejection of a null hypothesis. This is called the significance value or α value. It is typically 0.05 for a population with a normal distribution.
5. Calculate the p-value i.e. probability value which is the conditional probability of observing the test statistic.
6. Take the decision to reject or retain the null hypothesis based on p-value and the significance value α.

The following section illustrates the application of these steps with an example.

6.2.2 *Are Students Interested in Higher Education?: An Example of Hypothesis Testing*

Let us consider a situation where a Head of the Department of Computer Engineering at Dr. B. A. Technological University (DBATU) claims that *students at the institute are loosing interest in higher education.*

To support the claim, the head uses data collected from an exit survey which is conducted annually. Students from the final year are the respondents to it. The survey includes a question to judge students' career choices at the beginning of the program. The exit survey consists of a question: *When you took the admission to B. Tech in Computer Engineering program, what was your goal in life?* Only 11.6% of the students respond with an option of higher education, as shown in Fig. 6.1.

Further the head compares institute proportion with the national average of students qualifying the GATE examination. In India, students aspiring postgraduate

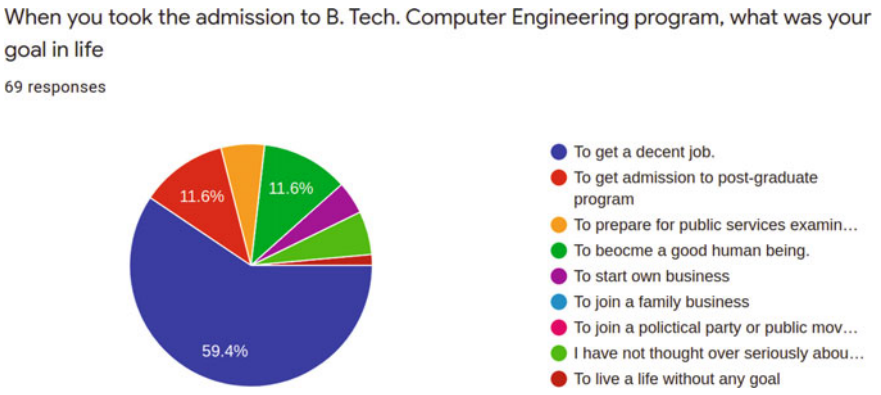

Fig. 6.1 Results of exit survey conducted at DBATU

education need to appear and qualify the Graduate Aptitude Test for Engineering (GATE), a national-level mandatory test. The national average of students qualifying GATE is 17%.

We can validate the head's claim that, *Students at the institute are loosing interest in higher education, through the hypothesis testing.*

1. **Define null H_0 and alternative H_1 hypothesis**: The null and alternative hypothesis are as follows.

   H_0: The proportion of students interested in higher education is below the national average of 17% of students qualifying in GATE.

   H_1: The proportion of students interested in higher education is not below national average 17% of qualifying students in GATE.
2. **Identify the test statistic**: We will use the z-test for proportion as we know the population parameter and distribution. Student's performance in examination typically follows the normal distribution or bell curve. The z-statistic is defined as below

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

   With the values of $\hat{p} = 11.6\%$, p = 17%, and n = 70 i.e. class size the value of z-statistic is −1.20.
3. The critical values for hypothesis testing for population proportion based on the sample are: −1.96 and 1.96. The value of z-statistics is greater than −1.96 (−1.20 > −1.96) and less than 1.96 (−1.20 < 1.96). It tells that z-statistic is not part of rejection region. Hence, we will retain the null hypothesis.
4. We can calculate the p-value using Microsoft Excel function NORM.S.DIST(−1.2). The calculated p-value is 0.11. It is greater than the α-value i.e. 0.05. So also, we will retain the null hypothesis.

Threats to Validity The observation is valid for the particular university as the sample size reflects the trend in the particular institute.

Also experts say that an observation to be valid the value of the expression n $\hat{p}(1 - \hat{p}) >= 10$. In our case, the value of the expression n $\hat{p}(1 - \hat{p})$ is 7.11 which is less then 10. It is because of small n in comparison with students appearing in the GATE. Typically students in the range of 100–200 thousands appear for the GATE.

6.2.3 Use-Cases of Hypothesis Testing from Literature

The hypothesis testing to test the validity of beliefs or opinions has been found useful for many researchers, as listed in Table 6.1. For example, in Ref. [3] authors test the validity of the claim: *In education systems with greater school clustering by past achievement, students have higher academic achievement than otherwise*. Similarly

Table 6.1 A few use-cases of hypothesis testing from literature

Sr. no	Statistic method	Purpose	Source
1	Regression	To improve students' learning achievements by grouping students into section	[3]
2	Correlation statistics Spearman's and Kendall's	To report the quality of ranking and rating of peer assessment using simulation approach	[4]
3	Effect or influence of learning environment on mind wondering and retention in online learning	Regression and Chi-Square test	[5]

in Ref. [4], authors validate the claim: *ranking-based peer assessment produces higher rank-order correlation between the assessment score and true artifact quality as compared to the rating-based peer assessment*. These approaches use the standard method for hypothesis testing, as described in this section.

6.3 Use-Case for ANOVA

The term ANOVA stands for *Analysis of Variance*. It is a statistical method to study the impact of a particular phenomenon on a dependent variable. This systematic method investigates whether there are statistically significant differences between the means of more than two independent groups. In this section, we illustrate the application of ANOVA to answer the question: *Does the choice of reference material affect students' performance?* In this case, the students who refer to different books or reference material form separate groups. The mean of their performance is compared to check the impact of selecting study material on their performances.

6.3.1 An Example: Does the Choice of Reference Material Affect student's Performance?

We have applied the ANOVA technique to check the impact of students' choice of reference material on their performances. In this context, we selected a course on C-Programming offered at DBATU to the students from the first year of B. Tech. in Computer Engineering. Because course on C-Programming is a first introductory course on programming intended to develop programming skills.

The course curricula prescribe two different books on C-Programming authored by E. Balagurusamy (EB) and Herbert Schildt (HS). Some students prefer to study

Table 6.2 Descriptive statistics for whole class and 4 groups formed by reference material used to study

Measures	Whole class	Name of reference books			
		EB	HS	DM	Mix
Count	69	23	13	22	11
Mean	61	59	59	58	72
Std. Deviation	14	12	13	16	13
Min	40	40	40	40	40
Max	95	80	80	95	86
Reference books	EB: E. Balgurusamy				
	HS: Herbert Schildit				
	DM: Digital material				
	Mix: Multiple resources				

either from a third book not prescribed in the syllabus or from the digital material available (DM) on the Internet. Some of the students prefer to study from more than one of the resource (Mix). Hence, there are four different groups in the class as per the choice of reference material used for the study; these are referred to as EB, HS, DM, and Mix, respectively. So the ANOVA test can be used in this context to answer the question *does the choice of reference material affect students' performance?*

The summary descriptive statistics of the whole class and four groups as formed according to their choice of material is given in Table 6.2. It includes the values min, max, mean, and standard deviation. The null and alternative hypothesis for the ANOVA test is formulated as:

$$H_0 : \mu_{EB} = \mu_{HS} = \mu_{DM} = \mu_{MIX}$$

where μ is the mean value of corresponding group. The null hypothesis states that there is no difference between the mean μ values of four different groups formed according to their reference books' choice. All mean values are the same.

The alternative hypothesis states that the mean values of all four groups are different.

$$H_A : \mu_{EB} \neq \mu_{HS} \neq \mu_{DM} \neq \mu_{MIX}$$

In the ANOVA case, the F-test is the test-statistic used to check whether the results are statistically significant or obtained by chance. The F-test is defined as:

$$F - Test = \frac{\frac{\sum_{i=1}^{k} n_i(\mu_i - \mu)^2}{k-1}}{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_j}(Y_{ij} - \mu_i)^2}{n-k}}$$

where,

K = Number of groups,

n_i = Number of observations in a group i

n = Total number of observations

Y_{ij} = Observation j in group i

μ_i = Mean of group i

μ = Overall mean.

We performed the ANOVA test using Python's *StatsModel* API. The ANOVA table obtained after the test is shown in Table 6.3. The value of F-statistic is 2.81. The critical value of F-statistic (Fc) can be calculated from Microsoft Excel function *FINV (α, degreeOfF reeedom, Sample size)*. The degree of freedom is 3 in our case, as four groups are present. With sample size of 69 and α equal to 0.05, we get F_c as 2.71. The F-statistic value is slightly greater than the critical value F_c. Hence, we reject the null hypothesis and the alternative hypothesis is retained (Fig. 6.2).

Further, the value of $eta_sq(\eta^2)$ is 0.11 signify that the choice of reference material students' performances is a moderate one.

Table 6.3 ANOVA table: the output of statmodels in python

	sum_sq	*df*	*mean_sq*	*F*	*PR(>F)*	*eta_sq*	*omega_sq*	F_c
C(Ref)	1713.35	3.0	571.11	2.81	0.04	0.11	0.07	2.71
Residual	13,196.34	65.0	203.02	NaN	NaN	NaN	NaN	

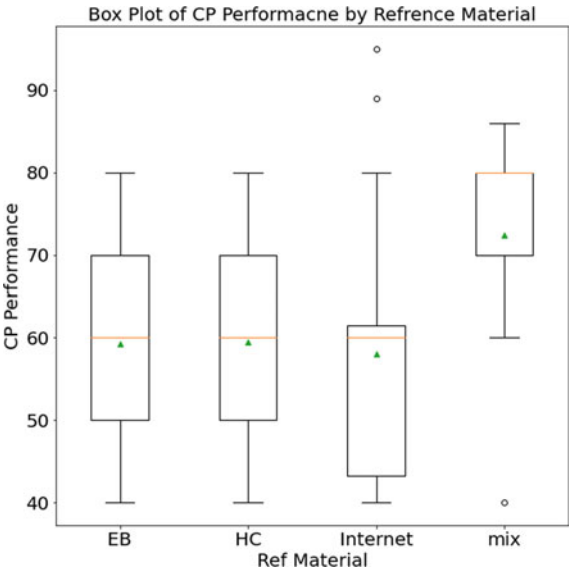

Fig. 6.2 Box plots for 4 groups performed according to reference material

6.3.2 Use-Cases of ANOVA from Literature

As observed from Table 6.4, the technique of ANOVA has been applied by many researchers to investigate the impact of different phenomenon on learning activities. In Ref. [6], authors have used ANOVA to study the impact of various kinds of background music on reading comprehension. The impact of student's choice about sequencing the completion of assignment on their performance has been analyzed using ANOVA. Further, an improvement in student's performance has been observed when such choice is offered.

6.4 A Use-Case for Correlation Relation

The correlation analysis is a powerful statistical method and it is quite useful from the learning analytics point of view. The purpose of correlation analysis is to check the association between two random variables. The correlation analysis plays an essential role during machine learning model development mainly to build predictive models. Correlation analysis is useful to investigate the strength of the relationship between the predicted variable and predictors. It is also helpful to check collinearity between two independent variables to eliminate redundant independent variables and thus reduce the number of features required to build predictive models.

The strength of the relationship between two random variables X and Y is measured by Pearson correlation coefficient when both X and Y are numeric and continuous variables. The Pearson correlation coefficient r is given by

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sigma_x \sigma_y}$$

where, (i) X, Y are mean values of X and Y, and

(ii) X_i, Y_i individual values of X and Y.

Table 6.4 A few use-cases ANOVA from literature

Sr. no	Techniques	Purpose	Source
1	ANOVA	To interpret patterns of students skill growth from learning curves for group of students and correlates with performance of groups of students	[7]
2	ANOVA	The impact of different types of background music on reading comprehension	[6]
3	ANOVA	To investigate the effect of choice on student assignment completion and learning gains. To investigate better performance between choice group and prescribe group	[8]

	CP	DSA
count	69	69
mean	61	61.56
std	14.80	15.67
min	40	40
max	95	95

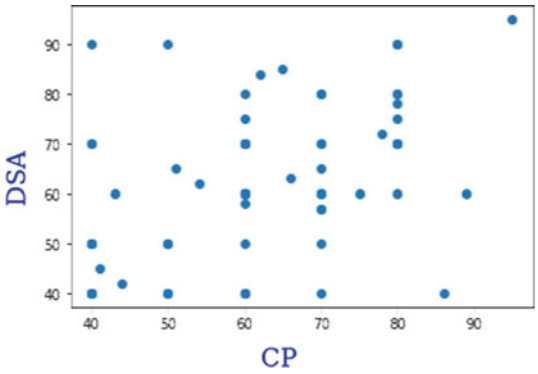

Fig. 6.3 Summary statistics and scatter plot for correlation analysis

The value of r is always between -1 to 1. A high absolute value of r, i.e. $|r|$ indicates a strong relationship between X and Y. When the value of r is positive, the relationship between X and Y is positive. Similarly, X and Y negatively related when r is negative (Fig. 6.3).

The correlation coefficient is also used to formulate the hypothesis. With r as a correlation coefficient of a sample and ρ as the population correlation coefficient the null and alternative hypothesis is formulated as:

$$H_0 : \rho = 0 \text{ There is no correlation between two random variables.}$$

$$H_1 : \rho \neq 0 \text{ There is correlation between two random variables.}$$

The t-statistic is typically used to check the statistical significance of correlation analysis. It is given by

$$t - value = r\sqrt{\frac{n-2}{1-r^2}}$$

6.4.1 An Example: Does a Student's Performance in a Course Depend on a Pre-requisite Course?

As an example of correlation analysis we would like to answer the question: *Does a student's performance in a course depend on a pre-requisite course?* A teacher normally assumes that a student's performance in a course depend on knowledge and skill acquired in a pre-requisite course. This assumption can be validated through the application of correlation analysis.

We consider two courses offered at DBATU at the first and second year of B. Tech. in Computer Engineering program. Courses are *C-Programming (CP)* and *Data Structure and Algorithm (DSA)* offered at first and second year respectively. C-programming is a pre-requisite course for Data Structure and Algorithm. Further, we use the marks obtained by students in both the courses to analyze the association between pre-requisite course and the main course. Table 6.3 shows the descriptive or summary statistics of both courses. The null and alternative hypothesis can be formulated in this case as given below:

$H_0 : r_{cp,dsa} \leq 0.50$: There is no correlation between CP and DSA performance of students.

$H_1 : r_{cp,dsa} > 0.50$: There is a correlation between CP and DSA performance of students.

We have calculated the Pearson Correlation coefficient for the data described in Table 6.3, and the value for $r_{cp,\,dsa}$ is 0.47. Pearson Correlation Coefficient can be calculated either in Excel sheet or in Python. Corresponding t-statistic value is 3.87. Critical value of t-statistic is 1.99 which can be calculated using Excel function *TINV (0.05, 67)*. Value t-statistic is greater than the critical value of t-statistic; so the null hypothesis H0 is rejected. Also, the p-value is 0.0002 (calculated using Excel function *T.DIST.2 T(3.87, 67)*), which is less than the α value of 0.05, indicating that the results are statistically significant. Hence, we can claim that students' performance in DSA depends on his/her performing in a pre-requisite course on C-Programming.

6.4.2 Use-Cases of Correlation Analysis from Literature

Researchers have used the correlation analysis to investigate the association between various parameters affecting learning. Few examples of such use cases are listed in Table 6.5. In Ref. [9], the authors have checked the dependency between students' confidence level and their performance in a course. An association between recurrently generated words by learners and their reading comprehension performance has been explored in [10]. In Ref. [11], researchers use correlation analysis to study the association between academic performance with learning strategies and course feedback in flipped classroom settings.

6.5 Use-Case for Linear Regression Analysis

Regression analysis is a powerful yet simple technique used to build predictive models. Predictive models attempt to predict the value of a dependent variable or an output variable from a set of independent or input variables as accurately as possible. While doing so, the model development methods make certain assumptions. Method

Table 6.5 A few use-cases of correlation analysis from literature

Sr. no	Techniques	Purpose	Source
1	Correlation	It revealed that recurrence indices were significantly related to the students comprehension score at both surface and deep levels	[10]
2	Pearson correlation	While writing the journals, grade and word count are significantly related. No significant correlation was found between journal grade and any sentimental levels (positive, negative and overall sentiment)	[12]
3	Pearson correlation	To correlate overconfidence and under confidence with students' overall course performance	[9]
4	Kruskal Wallis test	To examine association between learning strategies and academic performance	[11]
5	Spearman's correlation	To understand student use of digital data, online services to their own practices of privacy self management and their relation and concern about use of their data in the context of learning analytics	[13]

of Linear Regression Analysis assumes a linear relationship between an output variable and input variables. Hence, this kind of relationship is represented as a linear equation between input and output variables as given below

$$\mathscr{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

where, (i) x_1, x_2, …, x_n are input variables, (ii) β_0, β_1, …, β_n are coefficients of input variables, (iii) β_0 a Y-intercept, and (iv) Y is an output variable.

In the case of *Simple Linear Regression (SLR)*, we predict an output variable's value from a single input variable. The *Multiple Linear Regression (MLR)* predicts an output variable's value from more than one input variables.

Two different kinds of model development methods exist to build linear regression models. The first one is called *Ordinary Least Square* method and the second is known as *Gradient Descent Method*. The purpose of both methods is to estimate the values of linear coefficients i.e. β_0, β_1, β_2, …, and β_n.

The *Ordinary Least Square* method uses a statistical approach to build the regression model that minimizes the error between predicted and actual value. The *Gradient Descent Method* is an iterative method that guesses linear coefficients in each iteration values to find the best optimal combination of linear coefficients.

6.5.1 Predicting the Students Performance from Pre-requisite Courses

As an example to illustrate the application of linear regression analysis, we use the task of predicting the student's performance in a course from the prerequisite courses. We use the same set of courses in the previous section, i.e. C-Programming and Data Structure and Algorithm. The course C-Programming (CP) is a pre-requisite course for Data Structure and Algorithm (DSA). In this case, our objective is to estimate the marks in *Data Structure and Algorithm* from the marks of C-Programming. Hence, this is the case of applying simple linear regression. The predictive model can be described using the following equation.

$$DSA\_Marks = \beta_0 + \beta_1 \text{X} CP\_Marks$$

DSA_Marks and *CP_Marks* are the marks obtained in the course on Data Structure and Algorithm and C-Programming respectively.

We have used the Ordinary Least Square method provided in *Stats_Model* API of Python programming language to find the values of β_0 and β_1. The calculated values of β_0 and β_1 are 31.00 and 0.50 respectively as shown in Fig. 6.4. Thus the equation for the predictive model becomes

$$DSA\_Marks = 31.00 + 0.50 \text{X} CP\_Marks$$

The model has a residual prediction error of 13 means that predicted marks would be in the range of ± 13. This high expected range may be due to weaker linear strength observed between DSA marks and C-Programming, i.e., Pearson Correlation

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Data Structure and Algorithm   R-squared:                       0.224
Model:                                   OLS   Adj. R-squared:                  0.212
Method:                        Least Squares   F-statistic:                     19.34
Date:                       Thu, 07 Jan 2021   Prob (F-statistic):           4.01e-05
Time:                               09:59:49   Log-Likelihood:                -278.55
No. Observations:                         69   AIC:                             561.1
Df Residuals:                             67   BIC:                             565.6
Df Model:                                  1
Covariance Type:                   nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const           31.0038      7.149      4.337      0.000      16.735      45.273
C-Programming    0.5010      0.114      4.398      0.000       0.274       0.728
==============================================================================
Omnibus:                        1.668   Durbin-Watson:                   2.166
Prob(Omnibus):                  0.434   Jarque-Bera (JB):                1.100
Skew:                           0.290   Prob(JB):                        0.577
Kurtosis:                       3.215   Cond. No.                         268.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Fig. 6.4 A summary report of simple linear regression model

Table 6.6 A few use-cases of regression analysis

Sr. no	Technique	Purpose	Source
1	Regression	To find out strong relationship between pre and post requisite skill	[14]
2	Linear regression	To predict learning group productivity under the given restrictions or privacy constraints	[15]
3	Regression and root mean square error	From the instructor dashboard prediction of low scored students and identify students who need help from peer tutors or instructor	[16]

coefficient of 0.473 only (as seen in previous section). A Course instructor can use this predicted range of marks to identify students at risk at the beginning of the course.

6.5.2 Use-Cases of Regression Analysis from Literature

Researchers have used regression analysis for multiple purposes as listed in Table 6.6. One of the everyday use cases is to predict the performance on post-requisite skill set from the pre-requisite skill set. For example, in [14] authors use the data collected from the adaptive testing platform and multiple linear regression techniques to predict the performance on post-requisite skills.

The authors in [15] use simple linear regression to identify the factors contributing to the formation well-functioning group when learners share no personal data. The Ref. [16] use the logistic regression technique to identify the student's loss of interest in a MOOC at an early stage from the various engagement factors collected such as time spent watching videos and submission of assignment.

6.6 Conclusion

A large amount of data is being generated when educational institutes start using digital technologies such as the Learning Management System. Knowledge and skills to analyze this data are required to use this data meaningfully. Teachers can use various statistical methods to extract meaningful information from this data. The chapter illustrates applications of a set of such statistical methods in day-to-day learning activities.

Four statistical methods viz, hypothesis testing, ANOVA, correlation analysis and regression analysis are described with one example of each method. The use of these methods depends on the context of application. For example, hypothesis

testing is useful to conclude by comparing the means of a sample with a population or known accepted value. At the same time, ANOVA is used to measure the impact of a common phenomenon on multiple groups. The correlation analysis is used to check the association among various variables, while regression analysis is useful for predicting a dependent variable's values.

We describe a typical example highlighting the context in which a particular method is appropriate. Further, a few examples from the existing literature have been briefly described. The illustrated examples in the chapter and use-cases from the current literature lay the foundation upon which teachers can devise novel use-cases or experiments to investigate various learning theories' and their implications.

Appendix: A Primer on Statistical Terms and Definitions

The following are most commonly used terms from Statistics. Here we reproduce their definitions from [17]

1. **Population** It is the set of all possible data for a given context.
2. **Sample** It is the subset taken from population.
3. **Types of Data** A Data can be broadly classified into four categories. These are: (i) Continuous, (ii) Discrete, (iii) Ordinal, and (iv) Nominal. Discrete and continuous data is numeric type data. A continuous random variable such as temperature may take any value from the number space. A discrete random variable such as color can take a fixed set of values. For example, red, blue green etc. An implicit order of hierarchy is understood in case of ordinal type of variable such performance level may be excellent, very good, good and fair. No such implicit order is assumed in case of nominal type of random variable
4. **Mean** is the arithmetical average value of data and is one of the most frequently measures of central tendency. It is defined as:

$$\mu = \sum_{i=1}^{n} \frac{x_i}{n}$$

5. **Mode** is the most frequently occurring value in the data set. Mode is the only measure of central tendency which is valid for qualitative (nominal) data since the mean and median for nominal data are meaningless. In the bar chart (and histogram), mode is the tallest column.
6. **Median** It is the value that divides the data in to two equal parts, that is the proportion of observations below median and above median will be 50%. Median is much more stable than the mean value that is adding a new observation may not change the median significantly. However the drawback of median is that it is not calculated using the entire data like in the case of mean.
7. **Variance** It is the average of the squared differences from the Mean.

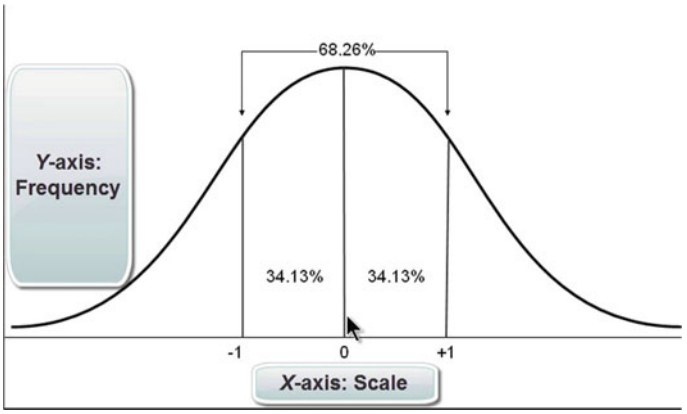

Fig. 6.5 Normal distribution

$$\text{var} = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$$

8. **Standard Deviation** It is a measurement of how far data is spread out from the mean, or average. The Standard Deviation is a measure of how spreads out numbers are. It is typically defined as

$$\sigma = \sqrt{\text{var}}$$

9. **Normal Distribution** It is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side. The area under the normal distribution curve represents probability and the total area under the curve sums to one. It is symmetrical bell-shaped graph as shown in Fig. 6.5.

10. **Co-Variance** It signifies the direction of the linear relationship between the two variables. By direction we mean if the variables are directly proportional or inversely proportional to each other. Increasing the value of one variable might have a positive or a negative impact on the value of the other variable.

$$Cov_{x,y} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{n - 1}$$

11. **Correlation** It is a measure of the strength and direction of relationship that exists between two random variables. It is a measure of association between two variables.
12. **Pearson Correlation Coefficient** measure the strength of the linear association relationship using numerical measure

$$PCC = \frac{Cov_{x,y}}{\sigma_x \sigma_y}$$

13. **Mean Square Error** It measures the average of the squares of the errors i.e., the average squared difference between the estimated values ($\hat{y}$) and actual value (y).

$$MSE = \sum_{i=1}^{n} \frac{\left(y_i - \hat{y}\right)^2}{n}$$

14. **t-test** is used when the population follows a normal distribution and population standard deviation is unknown. It shows how significant the differences between groups are.
15. **z-test** is a statistical test to determine whether two population means are different when the variances are known and the sample size is large. It can be used to test hypotheses in which the z-test follows a normal distribution.
16. **Chi-Squared test** It is used for testing relationships between categorical variables.

References

1. M. Kiseleva, N. Kiseleva, E. Kiselev, Changes in education under the influence of digital technologies: main problems and risk of division. KnE Soc. Sci. 22–28 (2020)
2. R.E. Slavin, How evidence-based reform will transform research and practice in education. Educ. Psychol. **55**(1), 21–31 (2020)
3. M.M. Chiu, B.W.-Y. Chow, S.W. Joh, How to assign students into sections to raise learning, in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (ACM, 2017), pp. 95–104
4. D. Babik, S. Stevens, A.E. Waters, Comparison of ranking and rating scales in online peer assessment: simulation approach, in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (2019), pp. 205–209
5. T.L. Varao-Sousa, C. Mills, A. Kingstone, Where you are, not what you see: the impact of learning environment on mind wandering and material retention, in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (2019), pp. 421–425
6. X. Hu, F. Li, R. Kong, Can background music facilitate learning? Preliminary results on reading comprehension, in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (2019), pp. 101–105
7. L. Chen, A. Dubrawski, Learning from learning curves: discovering interpretable learning trajectories, in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (ACM, 2017), pp. 544–545
8. S.A. Adjei, A.F. Botelho, N.T. Heffernan, Sequencing content in an adaptive testing system: the role of choice, in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (ACM, 2017), pp. 178–182
9. A. Aghababyan, N. Lewkow, R. Baker, Exploring the asymmetry of metacognition, in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (ACM, 2017), pp. 115–119

10. L.K. Allen, C.A. Perret, A.D. Likens, D.S. McNamara, What'd you say again?: recurrence quantification analysis as a method for analyzing the dynamics of discourse in a reading strategy tutor, in *LAK* (2017), pp. 373–382
11. W. Matcha, D. Găsević, N.A.A. Uzir, J. Jovanović, A. Pardo, Analytics of learning strategies: associations with academic performance and feedback, in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (2019), pp. 461–470
12. Y. Chen, B. Yu, X. Zhang, Y. Yu, Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals, in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (ACM, 2016), pp. 1–5
13. S. Slade, P. Prinsloo, M. Khalil, Learning analytics at the intersections of student trust, disclosure and benefit, in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (2019), pp. 235–244
14. S.A. Adjei, A.F. Botelho, N.T. Heffernan, Predicting student performance on post-requisite skills using prerequisite skill data: an alternative method for refining prerequisite skill structures, in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (ACM, 2016), pages 469–473
15. T. Hecking, D. Doberstein, H.U. Hoppe, Predicting the wellfunctioning of learning groups under privacy restrictions, in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (2019), pp. 245–249
16. M.L. Bote-Lorenzo, E. Ǵomez-Śanchez, Predicting the decrease of engagement indicators in a mooc, in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (2017), pp. 143–147
17. U.D. Kumar, *Business Analytics: The Science of Data-Driven Decision Making*. Wiley (2017)