Probability

1. Introduction

Probability is a measure quantifying the likelihood that events will occur. A probability is a way of assigning every event a value between zero and one, with the requirement that the event made up of all possible results is assigned a value of one. The opposite or complement of an event A is the event not A and its probability is given by P(not A) = 1 − P(A).

Experiment – are the uncertain situations, which could have multiple outcomes. Whether it rains on a daily basis is an experiment. Outcome is the result of a single trial. So, if it rains today, the outcome of today's trial from the experiment is "It rained". Event is one or more outcome from an experiment. "It rained" is one of the possible event for this experiment. Probability is a measure of how likely an event is. So, if it is 60% chance that it will rain tomorrow, the probability of Outcome "it rained" for tomorrow is 0.6.

2. Cases

2.1. Independent Events

If two events, A and B are independent then the joint probability is P(A and B) = P(A ∩ B) = P(A).P(B)

2.2. Mutually Exclusive Events

If either event A or event B but never both occurs on a single performance of an experiment, then they are called mutually exclusive events. If two events are mutually exclusive then the probability of both occurring is denoted as P (A ∩ B) which is equal to 0. If two events are mutually exclusive then the probability of either occurring is denoted as P (A ∪ B) which is equal to P(A) + P(B) − P(A ∩ B) = P(A) + P(B) − 0 = P(A) + P(B)

Conditional probability is the probability of some event A, given the occurrence of some other event B. Conditional probability is written as P(A | B) and is defined as P(A | B) = P(A ∩ B) / P(B).

2.4. Bayes' Theorem

Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. For example, if cancer is related to age, then, using Bayes' theorem, a person's age can be used to more accurately assess the probability that they have cancer than can be done without knowledge of the person's age. Bayes' theorem is stated mathematically as the following equation: P(A | B) = (P(B | A). P(A)) / P(B)

3. Probability Distributions

A probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment. A probability distribution is specified in terms of an underlying sample space, which is the set of all possible outcomes of the random phenomenon being observed. The sample space may be the set of real numbers or a set of vectors, or it may be a list of non-numerical values; for example, the sample space of a coin flip would be {heads, tails}.

Probability distributions are generally divided into two classes - Discrete and Continuous. A discrete probability distribution (applicable to the scenarios where the set of possible outcomes is discrete) can be encoded by a discrete list of the probabilities of the outcomes, known as a probability mass function. Well-known discrete probability distributions used in statistical modeling include the Poisson, the Bernoulli, the binomialand the geometric distribution. On the other hand, a continuous probability distribution (applicable to the scenarios where the set of possible outcomes can take on values in a continuous range) is typically described by probability density functions. There are many examples of continuous probability distributions: normal, uniform, chi-squared distribution.

A probability distribution whose sample space is one-dimensional is called univariate, while a distribution whose sample space is a vector space of dimension 2 or more is called multivariate.

3.1. Python

Excel supports a wide range of distribution functions which are mostly used to obtain p-values(INV) or critical values(DIST).

4. Hypothesis Testing

Hypothesis testing involves the careful construction of two statements: the null hypothesis and the alternative hypothesis. The aim of hypothesis testing is to find out if there is sufficient evidence to reject the null hypothesis in favour of alternate hypothesis. If there is not sufficient evidence, then we don't reject the null hypothesis.

Null hypothesis: “x is equal to y.” Alternative hypothesis: “x is not equal to y.”

Null hypothesis: “x is at least y.” Alternative hypothesis: “x is less than y.”

Null hypothesis: “x is at most y.” Alternative hypothesis: “x is greater than y.”

4.1. Null Hypothesis

The null hypothesis reflects that there will be no observed effect in our experiment. This hypothesis is denoted by Ho. The null hypothesis is what we attempt to find evidence against in our hypothesis test. We hope to obtain a small enough p-value that it is lower than our level of significance alpha and we are justified in rejecting the null hypothesis. If our p-value is greater than alpha, then we fail to reject the null hypothesis.

For example, if we are studying a new treatment, the null hypothesis is that our treatment will not change our subjects in any meaningful way. In other words, the treatment will not produce any effect in our subjects.

4.2. Alternate Hypothesis

The alternative or experimental hypothesis reflects that there will be an observed effect for our experiment. This hypothesis is denoted by H1. The alternative hypothesis is what we are attempting to demonstrate in an indirect way by the use of our hypothesis test. If the null hypothesis is rejected, then we accept the alternative hypothesis.

If the null hypothesis is not rejected, then we do not accept the alternative hypothesis. If we are studying a new treatment, then the alternative hypothesis is that our treatment does, in fact, change our subjects in a meaningful and measurable way.

4.3. Comparing sample mean with population mean (T - Test)

It is especially useful to compare a mean from a random sample to an established data source such as census data to ensure you have an unbiased sample. We will understand this with the help of an example. A random sample of 30 incoming college freshmen revealed the following statistics: mean age 19.5 years, standard deviation 1 year. The

In [4]:
```
import numpy as np
import scipy.stats as stats

np.random.seed(6) # fixing the random seed

incomers_ages = np.random.normal(loc= 19.5, scale= 1, size= 30) #sample o

stats.ttest_1samp(a= incomers_ages, popmean= 18) # T-test for comparing t
```

Out[4]:
```
Ttest_1sampResult(statistic=8.828807651463716, pvalue=1.02892601115938e-0
```

Here as we can observe that the p-value is less than 0.05, there is sufficient evidence to reject the null hypothesis being that the population mean= sample mean. The conclusion is that the sample mean is diiferent from the population mean

4.4. Comparing two sample means (Two sample T - Test)

We will understand this concept with the help of an example. An experiment is conducted to determine whether intensive tutoring (covering a great deal of material in a fixed amount of time) is more effective than paced tutoring (covering less material in the same amount of time). Two randomly chosen groups are tutored separately and then administered proficiency tests. Use a significance level of $\alpha < 0.05$, single tail.

Very different means can occur by chance if there is great variation among the individual samples. In order to account for the variation, we take the difference of the sample means and divide by the standard error in order to standardize the difference. Since, we do not know the population standard deviation, we estimate it using the two sample standard deviations from our independent samples.

Null Hypothesis: mean1 = mean2;

Alternate hypothesis: mean1 - mean2 > 0

Data: Sample 1: n= 12, mean1 = 46.31, s1 = 6.44. Sample 2: n= 10, mean2 = 42.79, s2 = 7.52.

$$t = (mean1 - mean2)/\sqrt{(s1^2/n1^2 + s2^2/n2^2)}$$

The number of degrees of freedom for the problem is smaller of n1-1 or n2-1.

With the given data, t is calculated to be 1.66. The degrees of freedom parameter is the smaller of (12 – 1) and (10 – 1), or 9. Because this is a one-tailed test, the alpha level (0.05) is not divided by two. In the t-table, the critical value is found to be 1.833. The computed t of 1.166 does not exceed the tabled value, therefore the null hypothesis cannot be rejected. This test has not provided statistically significant evidence that intensive tutoring is superior to paced tutoring.

In [5]:
```
import numpy as np
import scipy.stats as stats

np.random.seed(12)
score_intensive = np.random.normal(loc= 46.31, scale= 6.44, size= 12) # s
score_paced = np.random.normal(loc= 42.79, scale= 7.52, size= 10) # sampl

stats.ttest_ind(a= score_intensive, b= score_paced) # Two sample t test c
```

Out[5]:
```
Ttest_indResult(statistic=0.5110451757011513, pvalue=0.6149147980342374)
```

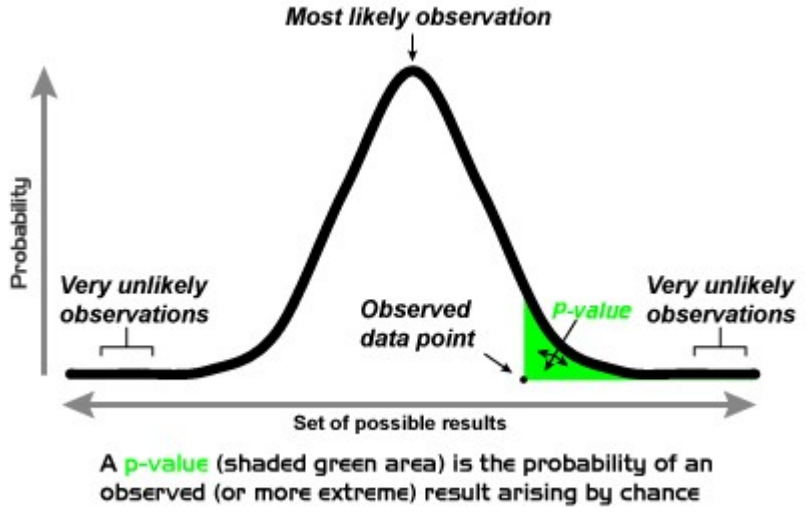
Here, p value is greater than 0.05 and therefore, the null hypothesis cannot be rejected. The conclusion is that there is no sufficient evidence to prove that the two sample means are different.

4.5. P - Value

Test statistics are helpful, but it can be more helpful to assign a p-value to these statistics.

A p-value is the probability that, if the null hypothesis were true, we would observe a statistic at least as extreme as the one observed. In simple terms, p-value is the probability of your null hypothesis actually being observed. For example, p=0.05 implies that one in a twenty observations follow your null hypothesis.

Typically, before we conduct a hypothesis test, we choose a threshold value. If we have any p-value that is less than or equal to this threshold, then we reject the null hypothesis. Otherwise we fail to reject the null hypothesis. This threshold is called the level of significance of our hypothesis test, and is denoted by the Greek letter alpha. There is no value of alpha that always defines statistical significance. However, the most common used criteria is that P < 0.05 is statistically significant and P < 0.001 is statistically highly significant.



A p-value (shaded green area) is the probability of an observed (or more extreme) result arising by chance

4.6. Test Statistics

A test statistic is a statistic (a quantity derived from the sample) used in statistical hypothesis testing. A hypothesis test is typically specified in terms of a test statistic, considered as a numerical summary of a data-set that reduces the data to one value that can be used to perform the hypothesis test. In general, a test statistic is selected or defined in such a way as to quantify, within observed data, behaviours that would distinguish the null from the alternative hypothesis. There is a wide range of test statistics used and we will look into a few of the important ones.

Z Value

Simply put, a z-score is the number of standard deviations from the mean a data point is. The basic z score formula for a sample is: $z = (x – \mu) / \sigma$, where μ is the mean of the population and σ is its standard deviation. If you pick any point in a normal distribution and desire to know its location in relevance to the mean of the population, z-score comes in handy. Z-score gives the point's location in terms of how many standard deviations away from the mean.

Sample problem: In general, the mean height of women is 65″ with a standard deviation of 3.5″. What is the probability of finding a random sample of 50 women with a mean height of 70″, assuming the heights are normally distributed?

$z = (x – \mu) / (\sigma / \sqrt{n}) = (70 – 65) / (3.5/\sqrt{50}) = 5 / 0.495 = 10.1$

We know that 99% of values fall within 3 standard deviations from the mean in a normal probability distribution. Therefore, there’s less than 1% probability that any sample of women will have a mean height of 70″.

CHI Square

Chi Square is highly useful in comparing the observed results and expected results of an experiment. The chi-square value can be obtained with the formula $chisquare = sum(i = 1tok)((oi - ei)^2/ei)$, where o is the observed value, e is the expected value and i is i'th position in the contingency table. Lower the chi-square value, higher is the similarity between observed and expected phenomena. You could also take your calculated chi-square value and compare it to a critical value from a chi-square table. If the chi-square value is more than the critical value, then there is a significant difference.

Consider a standard package of milk chocolate M&Ms. There are six different colors: red, orange, yellow, green, blue and brown. Suppose that we are curious about the distribution of these colors and ask, do all six colors occur in equal proportion? This is the type of question that can be answered with a goodness of fit test.

The null and alternative hypotheses for our goodness of fit test reflect the assumption that we are making about the population. Since we are testing whether the colors occur in equal proportions, our null hypothesis will be that all colors occur in the same proportion. More formally, if p1 is the population proportion of red candies, p2 is the population proportion of orange candies, and so on, then the null hypothesis is that p1 = p2 = . . . = p6 = 1/6. The alternative hypothesis is that at least one of the population proportions is not equal to 1/6.

The actual counts are the number of candies for each of the six colors. The expected count refers to what we would expect if the null hypothesis were true. We will let n be the size of our sample. The expected number of red candies is p1 n or n/6. In fact, for this example, the expected number of candies for each of the six colors is simply n times pi, or n/6.

We will now calculate a chi-square statistic for a specific example. Suppose that we have a simple random sample of 600 M&M candies with the following distribution:

212 of the candies are blue. 147 of the candies are orange. 103 of the candies are green. 50 of the candies are red. 46 of the candies are yellow. 42 of the candies are brown.

If the null hypothesis were true, then the expected counts for each of these colors would be (1/6) x 600 = 100. We now use this in our calculation of the chi-square statistic.

We calculate the contribution to our statistic from each of the colors. Each is of the form (Actual – Expected)^2/Expected.:

For blue we have (212 – 100)^2/100 = 125.44

For orange we have (147 – 100)^2/100 = 22.09

For green we have (103 – 100)^2/100 = 0.09

For red we have (50 – 100)^2/100 = 25

For yellow we have (46 – 100)^2/100 = 29.16

F Value

F values are generally used to compare the variances between two samples.

The F value is given by between-group variability/ within group variability. Between group variability = $sum(i = 1tok)Ni((Yi - Y)^2/k - 1)$, where Ni is the number of observations in the i'th group, k is the number of groups, Yi is the sample mean of the i'th group and Y is the overall mean. Within group variability = $sum(i = 1tok)sum(j = 1toNi)((Yij - Yi)^2/Ni - k)$, where Yij is the jth sample in the i'th group.

The F-statistic follows the F-distribution with degrees of freedom d1 = k-1 and d2 = Ni-k. The F-statistic compares the effects of all the variables together whereas the T-statistic only considers one variable.

F-value should always be used along with the p-value. If F-value is greater than the F-critical value and p-value is less than p-critical value then the null hypothesis can be rejected in favour of the alternate hypothesis. When there is a conflict, p-value gets more importance.

Sample problem: Conduct a two tailed F Test on the following samples: Sample 1: Variance = 109.63, sample size = 41. Sample 2: Variance = 65.99, sample size = 21.

Step 1: Write your hypothesis statements: Ho: No difference in variances. Ha: Difference in variances.

Step 2: Calculate your F critical value. Put the highest variance as the numerator and the lowest variance as the denominator: F Statistic = variance 1/ variance 2 = 109.63 / 65.99 = 1.66

Step 3: Calculate the degrees of freedom: The degrees of freedom in the table will be the sample size -1, so: Sample 1 has 40 df (the numerator). Sample 2 has 20 df (the denominator).

Step 4: Choose an alpha level. No alpha was stated in the question, so use 0.05 (the standard “go to” in statistics). This needs to be halved for the two-tailed test, so use 0.025.

Step 5: Find the critical F Value using the F Table. There are several tables, so make sure you look in the alpha = .025 table. Critical F (40,20) at alpha (0.025) = 2.287.

Step 6: Compare your calculated value (Step 2) to your table value (Step 5). If your calculated value is higher than the table value, you can reject the null hypothesis: F calculated value: 1.66 F value from table: 2.287. 1.66 < 2 .287. So we cannot reject the null hypothesis.

alt text

.jpg)

A type I error is the rejection of a true null hypothesis (also known as a "false positive" finding or conclusion), while a type II error is the non-rejection of a false null hypothesis (also known as a "false negative" finding or conclusion). Much of statistical theory revolves around the minimization of one or both of these errors, though the complete elimination of either is treated as a statistical impossibility.

The likelihood of type I error, is equal to the level of significance, that the researcher sets for his test. Here the level of significance refers to the chances of making type I error. Often, the significance level is set to 0.05 (5%), implying that it is acceptable to have a 5% probability of incorrectly rejecting the null hypothesis. The likelihood of making type II error is analogous to the power of the test. The power of the test is the probability that the test will find a statistically significant difference between men and women, as a function of the size of the true difference between those two populations. As the sample size increases, the power of test also increases, that results in the reduction in risk of making type II error. Power = 1 - risk of Type II error.

Consider a sample problem.

Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. Assume in a random sample 35 penguins, the standard deviation of the weight is 2.5 kg. If actual mean penguin weight is 15.1 kg, what is the probability of type II error for a hypothesis test at .05 significance level?

We begin with computing the standard error estimate, SE.

n = 35 # sample size

s = 2.5 # sample standard deviation

SE = s/sqrt(n); SE # standard error estimate

0.42258

We next compute the lower and upper bounds of sample means for which the null hypothesis μ = 15.4 would not be rejected.

alpha = .05 # significance level

mu0 = 15.4 # hypothetical mean

I = c(alpha/2, 1-alpha/2)

q = mu0 + qt(I, df=n-1) * SE; q

14.541 16.259

Therefore, so long as the sample mean is between 14.541 and 16.259 in a hypothesis test, the null hypothesis will not be rejected. Since we assume that the actual population mean is 15.1, we can compute the lower tail probabilities of both end points.

mu = 15.1 # assumed actual mean

4.8. KS Stat

4.8.1. One sample KS Test:

The Kolmogorov-Smirnov Goodness of Fit Test (K-S test) compares your data with a known distribution and lets you know if they have the same distribution. Although the test is nonparametric — it doesn't assume any particular underlying distribution — it is commonly used as a test for normality to see if your data is normally distributed.It's also used to check the assumption of normality in Analysis of Variance. More specifically, the test compares a known hypothetical probability distribution (e.g. the normal distribution) to the distribution generated by your data — the empirical distribution function.

The hypotheses for the test are:

Null hypothesis (H0): the data comes from the specified distribution.

Alternate Hypothesis (H1): at least one value does not match the specified distribution.

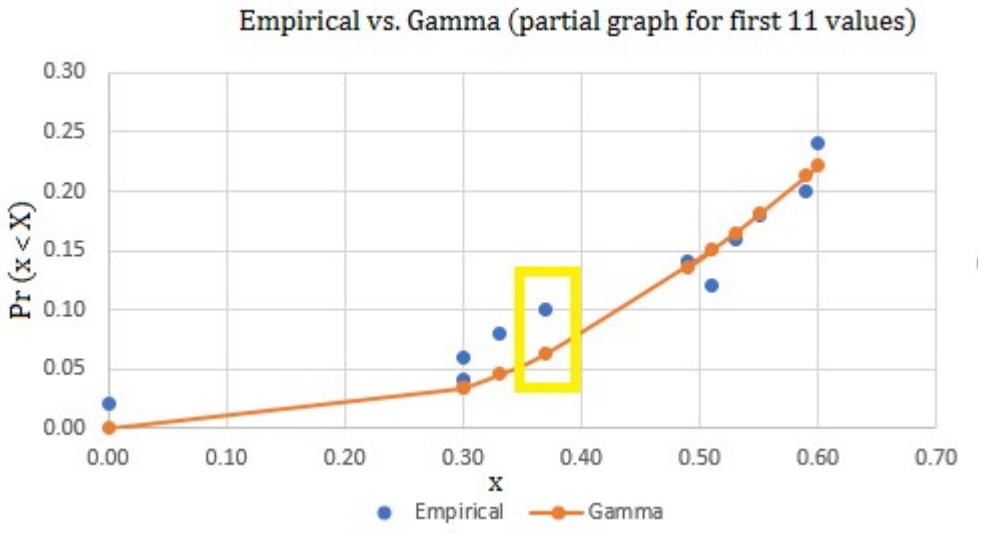
That is, H0: P = P0, H1: P ≠ P0. Where P is the distribution of your sample (i.e. the EDF) and P0 is a specified distribution.

Example:

Step 1: Create an EDF for your sample data.

Step 2: Specify the parent distribution. In this example, we will compare the emperical distribution with the gamma function.

Step 3: Graph the functions together. A snapshot of the scatter graph might like this:



Step 4: Measure the greatest vertical distance. Let's assume that the graph is of the entire sample and the largest vertical distance separating the two graphs is .04 (in the yellow highlighted box).

Cross Validation

Cross-validation, sometimes called rotation estimation or out-of-sample testing, is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (called the validation dataset or testing set). The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem).

One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set).

To reduce variability, in most methods multiple rounds of cross-validation are performed using different partitions, and the validation results are combined (e.g. averaged) over the rounds to give an estimate of the model's predictive performance. In summary, cross-validation combines (averages) measures of fitness in prediction to derive a more accurate estimate of model prediction performance.

Types

Two types of cross-validation can be distinguished: exhaustive and non-exhaustive cross-validation.

Exhaustive cross-validation

Exhaustive cross-validation methods are cross-validation methods which learn and test on all possible ways to divide the original sample into a training and a validation set.

Leave-p-out cross-validation

Leave-p-out cross-validation (LpO CV) involves using p observations as the validation set and the remaining observations as the training set. This is repeated on all ways to cut the original sample on a validation set of p observations and a training set. LpO cross-validation requires training and validating the model nCp times, where n is the number of observations in the original sample, and where p is the binomial coefficient. For p > 1 and for even moderately large n, LpO CV can become computationally infeasible. For example, with n = 100 and p = 30 = 30 percent of 100, 100C30 ≈ 3 × 10^25.

Questionnaire

1. A couple has two children, of which at least one is a boy. If the probabilities of having a boy or a girl are both 50%, what is the probability that the couple has two boys?

At first glance, We might reason as follows: “We know that one is a boy, so the only question is whether the other one is a boy, and the chances of that being the case are 50%. So, the answer must be 50%. But, it happens to be incorrect.

Define two events, A and B as follows:

A = Both children are boys. B = One of them is a boy.

We have to find P(A|B)

From Bayes' Theorem,

P(A|B) = P(B|A)P(A)/P(B) = (1)(1/4)/(3/4) = 1/3

2. Jackie wants to determine the probability of a Type I Error before she performs hypothesis testing on her sample of 40 individuals. Prior to testing, she decides on a significance level of 2%. What is the probability of a Type I Error in this case?

Rejecting the null hypothesis Ho when in fact Ho is the truth is known as Type 1 error. Type 1 errors depend on the selection of significance level. In this problem, a significance level of 2% is selected. Therefore, the probability of a Tpe 1 error happening = 0.02

3. A man was able to complete 3 files a day on an average. Find the probability that he can complete 5 files the next day.

The following formula can be used for computation of Poisson probability:

P (x, μ) = $\frac{(e^{-\mu})(\mu^{x})}{x!}$

Here, ‘x’ represents the actual number of occurring successes that are resulting from the Poisson experiment, the value is ‘e’ is 2.71828 approximately, 'μ ' is the average of the number of successes that are within a specified region.

Here we know this is a Poisson experiment with following values given: μ = 3, average number of files completed a day

x = 5, the number of files required to be completed next day

4. Consider the following situation. You are given a task of analysing the weights, heights and number of students who attend a class on a particular day. Which distribution will you use for the analysis and why?

In order to choose a distribution for study, we have to first understand the data and the target requirements.

Normal distribution describes continuous data which have a symmetric distribution, with a characteristic 'bell' shape.
Binomial distribution describes the distribution of binary data from a finite sample. Thus it gives the probability of getting r events out of n trials.
Poisson distribution describes the distribution of binary data from an infinite sample. Thus it gives the probability of getting r events in a population.

In this example, consider the weights and heights of the students. The weights and heights will more or less form a symmetric distribution, and we might be looking into the mean, median or mode of these features so as to understand the data. Normal distribution is a bell shaped curve characterised by two parameters mean and standard deviation. Thus, for the analysis of these two variables, normal distribution will be the best choice.

Consider the attendance of a class on a particular day. We will be looking into the probabilities of different number of students attending the class on a particular day. The Poisson distribution is one such distribution whose area under the curve at a particular point gives the probabilities of that many events happening. Thus, Poisson distribution will be an apt choice for these kinds of analysis.

5. If a group of students are treated using a new technique "Smart Class" and you want to check the proportion of students who will successfully pass the exam conducted at the end. Which distribution will you use and why?

Data which can take only a binary (0 or 1) response, such as fail or pass in an exam, follow the binomial distribution provided *the underlying population response rate does not change*. Under these conditions, binomial distribution will be of the best use. In this example, if we can assume that the student passing rate will be independent of the attempt, then we can use binomial distribution for the required analysis.