Q-1. Which of the following is the correct formula for total variation?

Ans- b) Total Variation = Residual Variation + Regression Variation

Q-2. Collection of exchangeable binary outcomes for the same covariate data are called \_\_\_\_\_\_\_\_ outcomes.

Ans- c) binomial

Q-3. How many outcomes are possible with Bernoulli trial?

Ans- a) 2

Q-4. If Ho is true and we reject it is called

Ans- a) Type-I error

Q-5. Level of significance is also called:

Ans- c) Level of confidence

Q-6. The chance of rejecting a true hypothesis decreases when sample size is:

Ans- b) Increase

Q-7. Which of the following testing is concerned with making decisions using data?

Ans- b) Hypothesis

Q-8. What is the purpose of multiple testing in statistical inference?

Ans- d) All of the mentioned

Q-9. Normalized data are centred at \_\_\_\_ and have units equal to standard deviations of the original data

Ans- a) 0

**Q-10and Q-15 are subjective answer type questions, Answer them in your own words briefly.**

Q-10. What Is Bayes' Theorem?

Ans- Bayes' Theorem, named after 18th-century British mathematician Thomas Bayes, is a mathematical formula for determining [conditional probability](https://www.investopedia.com/terms/c/conditional_probability.asp). Conditional probability is the likelihood of an outcome occurring, based on a previous outcome having occurred in similar circumstances. Bayes' theorem provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence.

In finance, Bayes' Theorem can be used to rate the [risk](https://www.investopedia.com/terms/r/risk.asp) of lending money to potential borrowers. The theorem is also called Bayes' Rule or Bayes' Law and is the foundation of the field of Bayesian statistics.

**KEY TAKEAWAYS**

* Bayes' Theorem allows you to update the predicted probabilities of an event by incorporating new information.
* Bayes' Theorem was named after 18th-century mathematician Thomas Bayes.
* It is often employed in finance in calculating or updating risk evaluation.
* The theorem has become a useful element in the implementation of machine learning.
* The theorem was unused for two centuries because of the high volume of calculation capacity required to execute its transactions.

**Understanding Bayes' Theorem**

Applications of Bayes' Theorem are widespread and not limited to the financial realm. For example, Bayes' theorem can be used to determine the accuracy of medical test results by taking into consideration how likely any given person is to have a disease and the general accuracy of the test. Bayes' theorem relies on incorporating [prior probability](https://www.investopedia.com/terms/p/prior_probability.asp) distributions in order to generate [posterior probabilities](https://www.investopedia.com/terms/p/posterior-probability.asp).

Prior probability, in Bayesian statistical inference, is the probability of an event occurring before new data is collected. In other words, it represents the best rational assessment of the probability of a particular outcome based on current knowledge before an experiment is performed.

Posterior probability is the revised probability of an event occurring after taking into consideration the new information. Posterior probability is calculated by updating the prior probability using Bayes' theorem. In statistical terms, the posterior probability is the probability of event A occurring given that event B has occurred.

**Special Considerations**

Bayes' Theorem thus gives the probability of an event based on new information that is, or may be, related to that event. The formula can also be used to determine how the probability of an event occurring may be affected by hypothetical new information, supposing the new information will turn out to be true.

For instance, consider drawing a single card from a complete deck of 52 cards.

The probability that the card is a king is four divided by 52, which equals 1/13 or approximately 7.69%. Remember that there are four kings in the deck. Now, suppose it is revealed that the selected card is a face card. The probability the selected card is a king, given it is a face card, is four divided by 12, or approximately 33.3%, as there are 12 face cards in a deck.

**Formula for Bayes' Theorem**

​*P*(*A*∣*B*)=*P*(*B*)*P*(*A*⋂*B)​=P(B)P(A)⋅P(B∣A)​*

Where:  
P(A)= The probability of A occurring

P(B)= The probability of B occurring

P(A∣B)=The probability of A given B

P(B∣A)= The probability of B given A

P(A⋂B))= The probability of both A and B occurring​

**Examples of Bayes' Theorem**

Below are two examples of Bayes' Theorem in which the first example shows how the formula can be derived in a stock investing example using Amazon.com Inc. ([AMZN](https://www.investopedia.com/markets/quote?tvwidgetsymbol=amzn)). The second example applies Bayes' theorem to pharmaceutical drug testing.

Q-11. What is z-score?

Ans- A Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score. A Z-score of 1.0 would indicate a value that is one standard deviation from the mean. Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

In finance, Z-scores are measures of an observation's variability and can be used by traders to help determine market volatility. The Z-score is also sometimes known as the [Altman Z-score](https://www.investopedia.com/terms/a/altman.asp).

* A Z-Score is a statistical measurement of a score's relationship to the mean in a group of scores.
* A Z-score can reveal to a trader if a value is typical for a specified data set or if it is atypical.
* In general, a Z-score below 1.8 suggests a company might be headed for bankruptcy, while a score closer to 3 suggests a company is in solid financial positioning.

## How Z-Scores Work

Z-scores reveal to statisticians and traders whether a score is typical for a specified data set or if it is atypical. Z-scores also make it possible for analysts to adapt scores from various data sets to make scores that can be compared to one another more accurately.Edward Altman, a professor at New York University, developed and introduced the Z-score formula in the late 1960s as a solution to the time-consuming and somewhat confusing process investors had to undergo to determine how close to [bankruptcy](https://www.investopedia.com/terms/b/bankruptcy.asp) a company was.1 In reality, the Z-score formula that Altman developed actually ended up providing investors with an idea of the overall financial health of a company.Over the years, Altman continued to reevaluate his Z-score. From 1969 until 1975, Altman looked at 86 companies in distress. From 1976 to 1995, he observed 110 companies. Finally, from 1997 to 1999, he evaluated an additional 120 companies. From his findings, it was revealed that the Z-score had an accuracy of between 82% and 94%.2.In 2012, Altman released an updated version of the Z-score, which is called the Altman Z-score Plus. It can be used to evaluate public and private companies, manufacturing and non-manufacturing companies, and U.S. and non-U.S. companies.3. A Z-score is the output of a credit-strength test that helps gauge the likelihood of bankruptcy for a publicly traded company. The Z-score is based on five key financial ratios that can be found and calculated from a company's annual [10-K report](https://www.investopedia.com/terms/1/10-k.asp). The calculation used to determine the Altman Z-score is as follows:4

## Z-Scores vs. Standard Deviation

Standard deviation is essentially a reflection of the amount of [variability](https://www.investopedia.com/terms/v/variability.asp) within a given data set. Standard deviation is calculated by first determining the difference between each data point and the mean. The differences are then squared, summed, and averaged. This produces the variance. The standard deviation is the square root of the variance.The Z-score, by contrast, is the number of standard deviations a given data point lies from the mean. For data points that are below the mean, the Z-score is negative. In most large data sets, 99% of values have a Z-score between -3 and 3, meaning they lie within three standard deviations above and below the mean.

## Criticisms of Z-Scores

The Z-score should be calculated and interpreted with care. For example, the Z-score is not immune to [false accounting practices](https://www.investopedia.com/terms/c/creative-accounting.asp). Since companies in trouble may sometimes misrepresent or cover up their financials, the Z-score is only as accurate as the data that goes into it.Additionally, the Z-score isn't very effective for new companies with little to zero earnings. Regardless of their actual financial health, these companies will score low. Moreover, the Z-score doesn't address the cash flows of a company. Rather, it only hints at it through the use of the net working capital-to-asset ratio.Finally, Z-scores can swing from quarter to quarter if a company records one-time write-offs. These events can change the final score and may falsely suggest a company is on the brink of bankruptcy.

Q-12. What is t-test?

Ans- A t-test is a statistical test that compares the means of two samples. It is used in hypothesis testing, with a null hypothesis that the difference in group means is zero and an alternate hypothesis that the difference in group means is different from zero. A t-test is an inferential [statistic](https://www.investopedia.com/terms/s/statistics.asp) used to determine if there is a significant difference between the means of two groups and how they are related. T-tests are used when the data sets follow a normal distribution and have unknown variances, like the data set recorded from flipping a coin 100 times.The t-test is a test used for hypothesis testing in statistics and uses the t-statistic, the [t-distribution](https://www.investopedia.com/terms/t/tdistribution.asp) values, and the degrees of freedom to determine statistical significance.

**KEY TAKEAWAYS**

* A t-test is an inferential statistic used to determine if there is a statistically significant difference between the means of two variables.
* The t-test is a test used for hypothesis testing in statistics.
* Calculating a t-test requires three fundamental data values including the difference between the mean values from each data set, the standard deviation of each group, and the number of data values.
* T-tests can be dependent or independent.

## Understanding the T-Test

A t-test compares the average values of two data sets and determines if they came from the same population. In the above examples, a sample of students from class A and a sample of students from class B would not likely have the same mean and standard deviation. Similarly, samples taken from the placebo-fed control group and those taken from the drug prescribed group should have a slightly different mean and standard deviation.Mathematically, the t-test takes a sample from each of the two sets and establishes the problem statement. It assumes a null hypothesis that the two means are equal.Using the formulas, values are calculated and compared against the standard values. The assumed null hypothesis is accepted or rejected accordingly. If the null hypothesis qualifies to be rejected, it indicates that data readings are strong and are probably not due to chance.The t-test is just one of many tests used for this purpose. Statisticians use additional tests other than the t-test to examine more variables and larger sample sizes. For a large sample size, statisticians use a [z-test](https://www.investopedia.com/terms/z/z-test.asp). Other testing options include the chi-square test and the f-test.

## Using a T-Test

Consider that a drug manufacturer tests a new medicine. Following standard procedure, the drug is given to one group of patients and a placebo to another group called the control group. The placebo is a substance with no therapeutic value and serves as a benchmark to measure how the other group, administered the actual drug, responds.After the drug trial, the members of the placebo-fed control group reported an increase in average life expectancy of three years, while the members of the group who are prescribed the new drug reported an increase in average life expectancy of four years.Initial observation indicates that the drug is working. However, it is also possible that the observation may be due to chance. A t-test can be used to determine if the results are correct and applicable to the entire population.Four assumptions are made while using a t-test. The data collected must follow a continuous or ordinal scale, such as the scores for an IQ test, the data is collected from a randomly selected portion of the total population, the data will result in a normal distribution of a bell-shaped curve, and equal or homogenous variance exists when the standard variations are equal.

## T-Test Formula

Calculating a t-test requires three fundamental data values. They include the difference between the mean values from each data set, or the mean difference, the standard deviation of each group, and the number of data values of each group.This comparison helps to determine the effect of chance on the difference, and whether the difference is outside that chance range. The t-test questions whether the difference between the groups represents a true difference in the study or merely a random difference.The t-test produces two values as its output: t-value and [degrees of freedom](https://www.investopedia.com/terms/d/degrees-of-freedom.asp). The t-value, or t-score, is a ratio of the difference between the mean of the two sample sets and the variation that exists within the sample sets.The numerator value is the difference between the mean of the two sample sets. The denominator is the variation that exists within the sample sets and is a measurement of the dispersion or variability.This calculated t-value is then compared against a value obtained from a critical value table called the T-distribution table. Higher values of the t-score indicate that a large difference exists between the two sample sets. The smaller the t-value, the more similarity exists between the two sample sets. Degrees of freedom refer to the values in a study that has the freedom to vary and are essential for assessing the importance and the validity of the null hypothesis. Computation of these values usually depends upon the number of data records available in the sample set.

Q-13. What is percentile?

Ans- A percentile is a comparison score between a particular score and the scores of the rest of a group. It shows the percentage of scores that a particular score surpassed. For example, if you score 75 points on a test, and are ranked in the 85 th percentile, it means that the score 75 is higher than 85% of the scores. In [statistics](https://en.wikipedia.org/wiki/Statistics), a k-th percentile (percentile score or centile) is a [score](https://en.wikipedia.org/wiki/Raw_score) below which a given [percentage](https://en.wikipedia.org/wiki/Percentage) k of scores in its [frequency distribution](https://en.wikipedia.org/wiki/Frequency_distribution) falls (exclusive definition) or a score at or below which a given percentage falls (inclusive definition). For example, the 50th percentile (the [median](https://en.wikipedia.org/wiki/Median)) is the score below which (exclusive) or at or below which (inclusive) 50% of the scores in the distribution may be found. Percentiles are expressed in the same [unit of measurement](https://en.wikipedia.org/wiki/Unit_of_measurement) as the input scores; for example, if the scores refer to [human weight](https://en.wikipedia.org/wiki/Human_weight), the corresponding percentiles will be expressed in kilograms or pounds.

The percentile score and the [percentile rank](https://en.wikipedia.org/wiki/Percentile_rank) are related terms. The percentile rank of a score is the percentage of scores in its distribution that are less than it, an exclusive definition, and one that can be expressed with a single, simple formula. Percentile scores and percentile ranks are often used in the reporting of [test scores](https://en.wikipedia.org/wiki/Test_score) from [norm-referenced tests](https://en.wikipedia.org/wiki/Norm-referenced_test), but, as just noted, they are not the same. For percentile rank, a score is given and a percentage is computed. Percentile ranks are exclusive. If the percentile rank for a specified score is 90%, then 90% of the scores were lower. In contrast, for percentiles a percentage is given and a corresponding score is determined, which can be either exclusive or inclusive. The score for a specified percentage (e.g., 90th) indicates a score below which (exclusive definition) or at or below which (inclusive definition) other scores in the distribution fall.

The 25th percentile is also known as the first [quartile](https://en.wikipedia.org/wiki/Quartile) (Q1), the 50th percentile as the median or second quartile (Q2), and the 75th percentile as the third quartile (Q3).

Applications[[edit](https://en.wikipedia.org/w/index.php?title=Percentile&action=edit&section=1)]

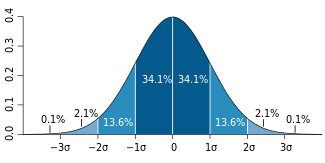
When [ISPs](https://en.wikipedia.org/wiki/Internet_Service_Provider) bill ["burstable" internet bandwidth](https://en.wikipedia.org/wiki/Burstable_billing), the 95th or 98th percentile usually cuts off the top 5% or 2% of bandwidth peaks in each month, and then bills at the nearest rate. In this way, infrequent peaks are ignored, and the customer is charged in a fairer way. The reason this statistic is so useful in measuring data throughput is that it gives a very accurate picture of the cost of the bandwidth. The 95th percentile says that 95% of the time, the usage is below this amount: so, the remaining 5% of the time, the usage is above that amount.

Physicians will often use infant and children's weight and height to assess their growth in comparison to national averages and percentiles which are found in [growth charts](https://en.wikipedia.org/wiki/Growth_chart).

The 85th percentile speed of traffic on a road is often used as a guideline in setting [speed limits](https://en.wikipedia.org/wiki/Speed_limit) and assessing whether such a limit is too high or low.[[1]](https://en.wikipedia.org/wiki/Percentile#cite_note-1)[[2]](https://en.wikipedia.org/wiki/Percentile#cite_note-2)

In finance, [value at risk](https://en.wikipedia.org/wiki/Value_at_risk) is a standard measure to assess (in a model-dependent way) the quantity under which the value of the portfolio is not expected to sink within a given period of time and given a confidence value.

The normal distribution and percentiles[[edit](https://en.wikipedia.org/w/index.php?title=Percentile&action=edit&section=2)]

[](https://en.wikipedia.org/wiki/File:Standard_deviation_diagram.svg)

Representation of the [three-sigma rule](https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule). The dark blue zone represents observations within one [standard deviation](https://en.wikipedia.org/wiki/Standard_deviation) (σ) to either side of the [mean](https://en.wikipedia.org/wiki/Mean) (μ), which accounts for about 68.3% of the population. Two standard deviations from the mean (dark and medium blue) account for about 95.4%, and three standard deviations (dark, medium, and light blue) for about 99.7%.

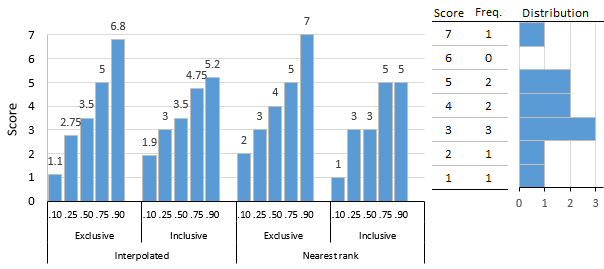
The methods given in the definitions section (below) are approximations for use in small-sample statistics. In general terms, for very large populations following a [normal distribution](https://en.wikipedia.org/wiki/Normal_distribution), percentiles may often be represented by reference to a normal curve plot. The normal distribution is plotted along an axis scaled to [standard deviations](https://en.wikipedia.org/wiki/Standard_deviation), or sigma ({\displaystyle \sigma }) units. Mathematically, the normal distribution extends to negative [infinity](https://en.wikipedia.org/wiki/Infinity) on the left and positive infinity on the right. Note, however, that only a very small proportion of individuals in a population will fall outside the −3σ to +3σ range. For example, with human heights very few people are above the +3σ height level.

Percentiles represent the area under the normal curve, increasing from left to right. Each standard deviation represents a fixed percentile. Thus, rounding to two decimal places, −3σ is the 0.13th percentile, −2σ the 2.28th percentile, −1σ the 15.87th percentile, 0σ the 50th percentile (both the mean and median of the distribution), +1σ the 84.13th percentile, +2σ the 97.72nd percentile, and +3σ the 99.87th percentile. This is related to the [68–95–99.7 rule](https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule) or the three-sigma rule. Note that in theory the 0th percentile falls at negative infinity and the 100th percentile at positive infinity, although in many practical applications, such as test results, natural lower and/or upper limits are enforced.

Definitions[[edit](https://en.wikipedia.org/w/index.php?title=Percentile&action=edit&section=3)]

There is no standard definition of percentile,[[3]](https://en.wikipedia.org/wiki/Percentile#cite_note-hyndman-3)[[4]](https://en.wikipedia.org/wiki/Percentile#cite_note-4)[[5]](https://en.wikipedia.org/wiki/Percentile#cite_note-pottel-5) however all definitions yield similar results when the number of observations is very large and the probability distribution is continuous.[[6]](https://en.wikipedia.org/wiki/Percentile#cite_note-schoonjans-6) In the limit, as the sample size approaches infinity, the 100pth percentile (0<p<1) approximates the inverse of the [cumulative distribution function](https://en.wikipedia.org/wiki/Cumulative_distribution_function) (CDF) thus formed, evaluated at p, as p approximates the CDF. This can be seen as a consequence of the [Glivenko–Cantelli theorem](https://en.wikipedia.org/wiki/Glivenko%E2%80%93Cantelli_theorem" \o "Glivenko–Cantelli theorem). Some methods for calculating the percentiles are given below.

Calculation methods[[edit](https://en.wikipedia.org/w/index.php?title=Percentile&action=edit&section=4)]

[](https://en.wikipedia.org/wiki/File:Frequency_histogram_and_exclusive_and_inclusive_percentiles_2.png)

Interpolated and nearest-rank, exclusive and inclusive, percentiles for 10-score distribution.

There are many formulas or algorithms[[7]](https://en.wikipedia.org/wiki/Percentile#cite_note-7) for a percentile score. Hyndman and Fan [[3]](https://en.wikipedia.org/wiki/Percentile#cite_note-hyndman-3) identified nine and most statistical and spreadsheet software use one of the methods they describe.[[8]](https://en.wikipedia.org/wiki/Percentile#cite_note-nist-8) Algorithms either return the value of a score that exists in the set of scores (nearest-rank methods) or interpolate between existing scores and are either exclusive or inclusive.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Nearest-rank methods (exclusive/inclusive) | | | | | |
| PC: percentile specified | 0.10 | 0.25 | 0.50 | 0.75 | 0.90 |
| N: Number of scores | 10 | 10 | 10 | 10 | 10 |
| OR: ordinal rank = PC × N | 1 | 2.5 | 5 | 7.5 | 9 |
| Rank: >OR / ≥OR | 2/1 | 3/3 | 6/5 | 8/8 | 10/9 |
| Score at rank (exc/inc) | 2/1 | 3/3 | 4/3 | 5/5 | 7/5 |

The figure shows a 10-score distribution, illustrates the percentile scores that result from these different algorithms, and serves as an introduction to the examples given subsequently. The simplest are nearest-rank methods that return a score from the distribution, although compared to interpolation methods, results can be a bit crude. The Nearest-Rank Methods table shows the computational steps for exclusive and inclusive methods.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Interpolated methods (exclusive/inclusive) | | | | | |
| PC: percentile specified | 0.10 | 0.25 | 0.50 | 0.75 | 0.90 |
| N: number of scores | 10 | 10 | 10 | 10 | 10 |
| OR: PC×(N+1) / PC×(N−1)+1 | 1.1/1.9 | 2.75/3.25 | 5.5/5.5 | 8.25/7.75 | 9.9/9.1 |
| LoRank: OR truncated | 1/1 | 2/3 | 5/5 | 8/7 | 9/9 |
| HIRank: OR rounded up | 2/2 | 3/4 | 6/6 | 9/8 | 10/10 |
| LoScore: score at LoRank | 1/1 | 2/3 | 3/3 | 5/4 | 5/5 |
| HiScore: score at HiRank | 2/2 | 3/3 | 4/4 | 5/5 | 7/7 |
| Difference: HiScore − LoScore | 1/1 | 1/0 | 1/1 | 0/1 | 2/2 |
| Mod: fractional part of OR | 0.1/0.9 | 0.75/0.25 | 0.5/0.5 | 0.25/0.75 | 0.9/0.1 |
| Interpolated score (exc/inc) = LoScore + Mod × Difference | 1.1/1.9 | 2.75/3 | 3.5/3.5 | 5/4.75 | 6.8/5.2 |

Interpolation methods, as the name implies, can return a score that is between scores in the distribution. Algorithms used by statistical programs typically use interpolation methods, for example, the percentile.exc and percentile.inc functions in Microsoft Excel. The Interpolated Methods table shows the computational steps.

Q-14. What is ANOVA?

Ans- An Analysis of Variance (ANOVA) is an inferential statistical tool that we use to find statistically significant differences among the *means* of two or more populations. We calculate variance but the goal is still to compare population mean differences. The test statistic for the ANOVA is called F. It is a ratio of two estimates of the population variance based on the sample data. Experiments are designed to determine if there is a cause and effect relationship between two variables. In the language of the ANOVA, the factor is the variable hypothesized to cause some change (effect) in the response variable (dependent variable).

An ANOVA conducted on a design in which there is only one factor is called a **one-way ANOVA**. If an experiment has two factors, then the ANOVA is called a *two-way ANOVA*. For example, suppose an experiment on the effects of age and gender on reading speed were conducted using three age groups (8 years, 10 years, and 12 years) and the two genders (male and female). The factors would be age and gender. Age would have three levels and gender would have two levels. ANOVAs can also be used for within-group/repeated and between subjects designs.  For this chapter we will focus on *between subject one-way ANOVA*. In a One-Way ANOVA we compare two types of variance: the variance between groups and the variance within groups, which we will discuss in the next section.

### Observing and Interpreting Variability

We have seen time and again that scores, be they individual data or group means, will differ naturally. Sometimes this is due to random chance, and other times it is due to actual differences. Our job as scientists, researchers, and data analysts is to determine if the observed differences are systematic and meaningful (via a hypothesis test) and, if so, what is causing those differences. Through this, it becomes clear that, although we are usually interested in the mean or average score, it is the variability in the scores that is key.

Take a look at figure 1, which shows scores for many people on a test of skill used as part of a job application. The x-axis has each individual person, in no particular order, and the y-axis contains the score each person received on the test. As we can see, the job applicants differed quite a bit in their performance, and understanding why that is the case would be extremely useful information. However, there’s no interpretable pattern in the data, especially because we only have information on the test, not on any other variable (remember that the x-axis here only shows individual people and is not ordered or interpretable).

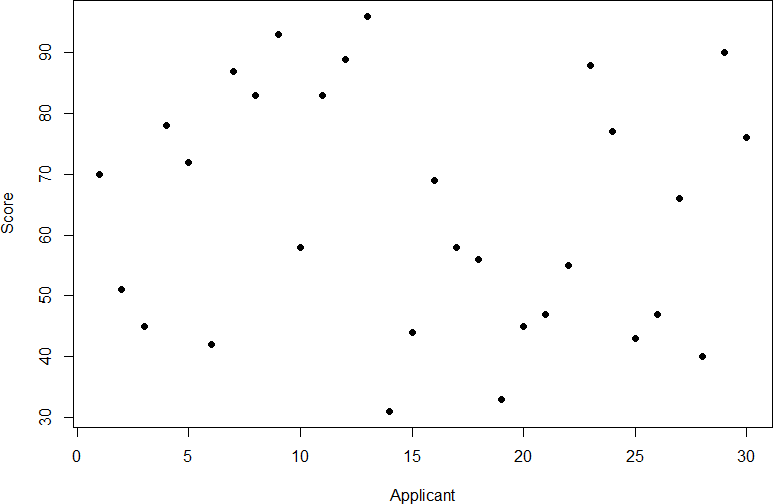


Figure 1. Scores on a job test

Our goal is to explain this variability that we are seeing in the dataset. Let’s assume that as part of the job application procedure we also collected data on the highest degree each applicant earned. With knowledge of what the job requires, we could sort our applicants into three groups: those applicants who have a college degree related to the job, those applicants who have a college degree that is not related to the job, and those applicants who did not earn a college degree. This is a common way that job applicants are sorted, and we can use ANOVA to test if these groups are actually different. Figure 2 presents the same job applicant scores, but now they are color coded by group membership (i.e. which group they belong in). Now that we can differentiate between applicants this way, a pattern starts to emerge: those applicants with a relevant degree (coded red) tend to be near the top, those applicants with no college degree (coded black) tend to be near the bottom, and the applicants with an unrelated degree (coded green) tend to fall into the middle. However, even within these groups, there is still some variability, as shown in Figure 2.

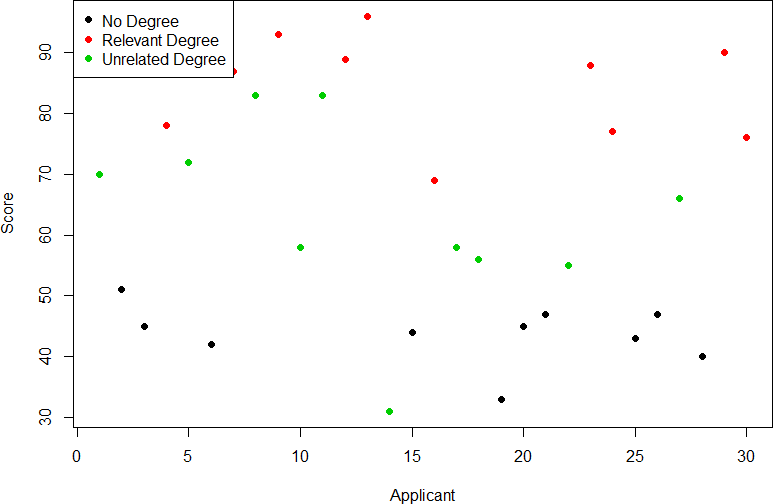


Figure 2. Applicant scores coded by degree earned

This pattern is even easier to see when the applicants are sorted and organized into their respective groups, as shown in Figure 3.

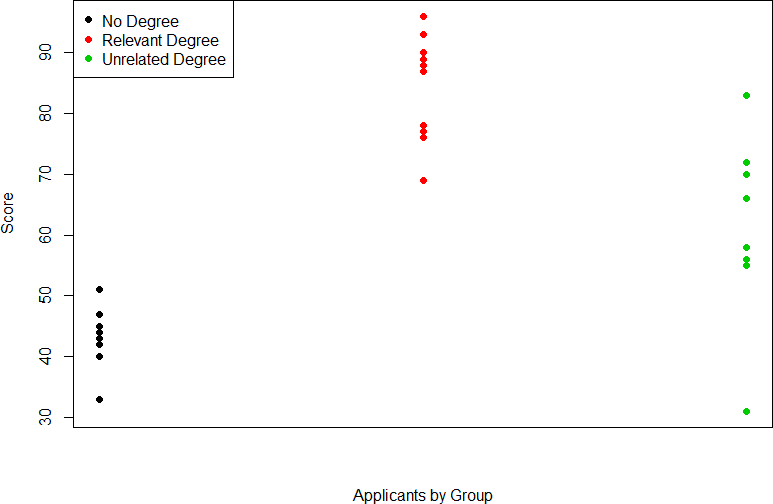
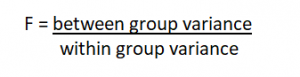


Figure 3. Applicant scores by group

Now that we have our data visualized into an easily interpretable format, we can clearly see that our applicants’ scores differ largely along group lines. Those applicants who do not have a college degree received the lowest scores, those who had a degree relevant to the job received the highest scores, and those who did have a degree but one that is not related to the job tended to fall somewhere in the middle. Thus, we have systematic variance between our groups.

We can also clearly see that within each group, our applicants’ scores differed from one another. Those applicants without a degree tended to score very similarly, since the scores are clustered close together. Our group of applicants with relevant degrees varied a little but more than that, and our group of applicants with unrelated degrees varied quite a bit. It may be that there are other factors that cause the observed score differences within each group, or they could just be due to random chance. Because we do not have any other explanatory data in our dataset, the variability we observe within our groups is considered random error, with any deviations between a person and that person’s group mean caused only by chance. Thus, we have unsystematic (random) variance within our groups.

The process and analyses used in ANOVA will take these two sources of variance (systematic variance between groups and random error within groups, or how much groups differ from each other and how much people differ within each group) and compare them to one another to determine if the groups have any explanatory value in our outcome variable. By doing this, we will test for statistically significant differences between the group means, just like we did for t– tests. We will go step by step to break down the math to see how ANOVA actually works.

ANOVA (analysis of variance) breaks down to:                                                                                                                                                                             where F is the new statistic reported for ANOVAs

Q-15. How can ANOVA help?

Ans- A researcher might, for example, test students from multiple colleges to see if students from one of the colleges consistently outperform students from the other colleges. In a business application, an R&D researcher might test two different processes of creating a product to see if one process is better than the other in terms of cost efficiency. The type of ANOVA test used depends on a number of factors. It is applied when data needs to be experimental. Analysis of variance is employed if there is no access to statistical software resulting in computing ANOVA by hand. It is simple to use and best suited for small samples. With many experimental designs, the sample sizes have to be the same for the various factor level combinations. ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample [t-tests](https://www.investopedia.com/terms/t/t-test.asp). However, it results in fewer [type I errors](https://www.investopedia.com/terms/t/type_1_error.asp) and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources. It is employed with subjects, test groups, between groups and within groups.

## One-Way ANOVA Versus Two-Way ANOVA

There are two main types of ANOVA: one-way (or unidirectional) and two-way. There also variations of ANOVA. For example, MANOVA (multivariate ANOVA) differs from ANOVA as the former tests for multiple dependent variables simultaneously while the latter assesses only one dependent variable at a time. One-way or two-way refers to the number of independent variables in your analysis of variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether all the samples are the same. The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents. For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as salary and skill set. It is utilized to observe the interaction between the two factors and tests the effect of two factors at the same time.