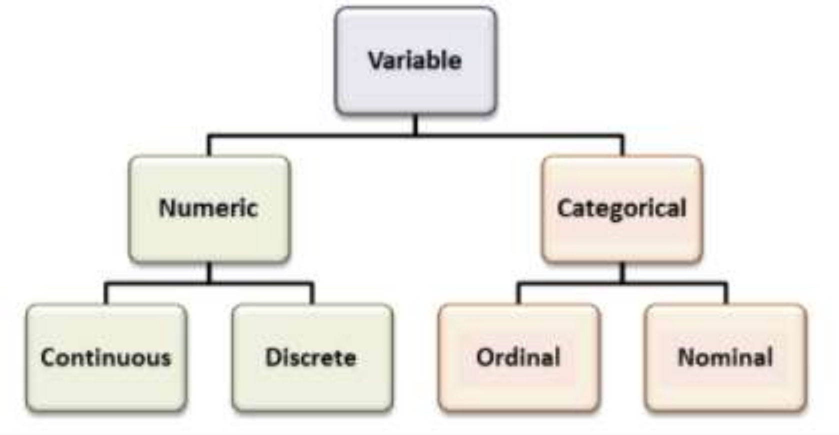


Categorical Data:

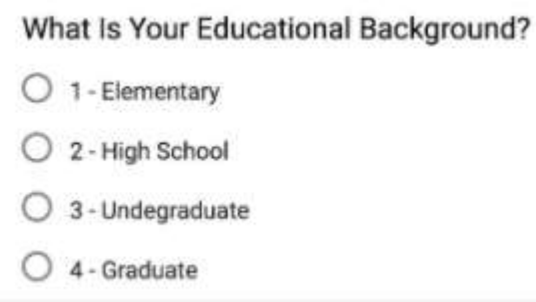
Categorical data represents characteristics. Therefore it can represent things like a person's gender, language etc. Categorical data can also take on numerical values (Example: 1 for female and 0 for male). Note that those numbers don't have mathematical meaning.

Nominal Data:- Nominal values represent discrete units and are used to label variables, that have no quantitative value. Just think of them as "labels". Note that nominal data that has no order. Therefore if you would change the order of its values, the meaning would not change. You can see two examples of nominal features below:

What is your gender?   What languages do you speak?   etc

The left feature that describes a persons gender would be called "dichotomous", which is a type of nominal scales that contains only two categories.

Ordinal Data:- Ordinal values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that it's ordering matters. You can see an example below:



What Is Your Educational Background?

○ 1 - Elementary

○ 2 - High School

○ 3 - Undegraduate

○ 4 - Graduate

Note that the difference between Elementary and High School is different than the difference between High School and College. This is the main limitation of ordinal data, the differences between the values is not really known. Because of that, ordinal scales are usually used to measure non-numeric features like happiness, customer satisfaction and so on.

Numerical Data:-

Discrete Data-: We speak of discrete data if its values are distinct and separate. In other words: We speak of discrete data if the data can only take on certain values. This type of data can't be measured but it can be counted. It basically represents information that can be categorized into a classification. An example is the number of heads in 100 coin flips.

Continuous Data:- Continuous Data represents measurements and therefore their values can't be counted but they can be measured. An example would be the height of a person. You can only describe them by using intervals on the real number line

Why Data Types are important?

Datatypes are an important concept because statistical methods can only be used with certain data types. You have to analyze continuous data differently than categorical data otherwise it would result in a wrong analysis. Therefore knowing the types of data you are dealing with, enables you to choose the correct method of analysis.

z-score

A z-score (aka, a standard score) indicates how many standard deviations an element is from the mean. A z-score can be calculated from the following formula.

$z = (X - \mu) / \sigma$ [the formula is actually $z = (X - \mu) / (\sigma/\sqrt{n})$ where n is sample size, n=1 if only one observation is talked about]

where z is the z-score, X is the value of the element, μ is the population mean, and σ is the standard deviation of population.

Here is how to interpret z-scores.

- A z-score less than 0 represents an element less than the mean.
- A z-score greater than 0 represents an element greater than the mean.
- A z-score equal to 0 represents an element equal to the mean.
- A z-score equal to 1 represents an element that is 1 standard deviation greater than the mean; a z-score equal to 2, 2 standard deviations greater than the mean; etc.
- A z-score equal to -1 represents an element that is 1 standard deviation less than the mean; a z-score equal to -2, 2 standard deviations less than the mean; etc.
- If the number of elements in the set is large, about 68% of the elements have a z-score between -1 and 1; about 95% have a z-score between -2 and 2; and about 99% have a z-score between -3 and 3.

Source (https://stattrek.com/statistics/dictionary.aspx?definition=z-score)

Why do we need Standard Normal Distribution?

A standard normal distribution is used to make comparisons between different variables. Lets see this with an example.

Lets say Student 1 got 75 out of 100 in English.

Student 2 got 60 out of 100 in Maths.

Question:- Who is a better Student, 1 or 2.

Solution-

Now this is a tricky question as the Students are being compared between 2 different subjects and we do-not have their marks in the same subject. If we had both Student1 & 2's marks in the same subject, comparing would be easy. But here since its in 2 different subject it makes it a tricky question.

We will try to find how much better (or worse they are with respect to the average class). Or in other words, How much better is Student 1 in English with respect to the mean marks of the class and similarly how much better is Student 2 in Maths with respect to the mean marks of the class. Recollect this is given by the z values.

Z score gives us how far are you from the mean.

Of course we will assume that the marks in both subjects follow a Normal Distribution.

Lets say,

English follows a ND with Mean=65 and sd=4

Maths follows a ND with Mean=54 and sd=2

Lets find z score for Student 1 in English

z_student1= (75-65)/4 =2.5

So, Student 1 is 2.5 sd's away from the average class.

Lets find z score for Students 2 in Maths

z_student2= (60-54)/2 =3

So, Student 2 is 3 sd's away from the average class.

This gives us a conclusion that Student 2 is better than Student 1 as she/he scored (more) better than the rest of the class than Student 1.

Note: Instead of finding how much Better the Students are from the mean class in terms of z value, we could have also used Proportion as in How much better are they from the rest of the class (example they scored better than 70% of the class etc). In that Case we will calculate the area on the

LHS of the curve for both Students. (The conclusion would be the same in both approaches)

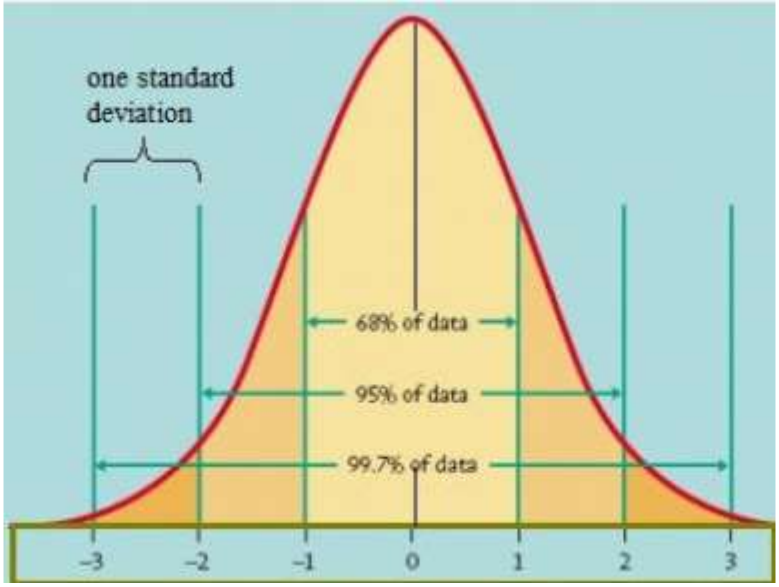
More examples here

https://statistics.laerd.com/statistical-guides/standard-score.php

Key Ideas_Normal Distribution

Monthly rents in a neighborhood have an average of $900 and an SD of $600. Could the distribution of rents be approximately normal?

Solution:



Normal distribution says that beyond 2 SD's away on both side, you have 5% of the Area. So that means beyond 2 SD on one side you have 2.5% area. So in this case If the rents followed a ND beyond 2 SD on the lower side their should be approx 2.5% rents. Lets calculate 2SD below the mean in the question.

Mean- 2* SD= 900-2*600= A negative number.

Now since Rents can never be negative, the above cannot be TRUE.

This gives us a conclusion that this distribution of rents is not normal, not even approximately.
The big spread is most likely coming from a long right hand tail.

(Note: A negative z is possible in this case, Negative z only means the rent is less than mean, but we cannot go less than z=-1.5 as that would mean Negative rents which is not logical.)

Key_Ideas_Binomial Theorem

Binomial distribution can be used in any task which requires repeating the same experiment more than once and calculating the probability of a specified number of outcomes.

What is the probability of getting upto 17 Heads when you toss a coin 20 times?

Solution:-

```
(binom.cdf(17,20,0.5))*100
```

Key_Ideas_CLT

The central limit theorem says that if you sample randomly from a population repeatedly, and for each sample you compute an average value over that sample, that the distribution of the averages is approximated by a ND.

It also tells us that the larger the sample, the better the approximation, and it most crucially, tells us that the mean of the normal distribution that is approximated is equal to the true population average that you're trying to estimate.

For instance, suppose we are pollsters trying to guess how an election will turn out. We take a poll and find that in our sample, 58% of people would vote for candidate A over candidate B.

Of course, we have only observed a small sample of the overall population, so we'd like to know if our result can be said to hold for the entire population, and if it can't, we'd like to know how large the error might be.

The central limit theorem tells us that if we ran the poll over and over again, the resulting guesses would be normally distributed around the true population value.

The central limit theorem is also used in surveys, clinical trials, randomized experiments, longitudinal studies and all kinds of empirical research.

Key Ideas Hypothesis Testing 1

What is a Hypothesis?

A hypothesis is an educated guess about something in the world around you. It should be testable, either by experiment or observation. For example:

- A new medicine you think might work.
- A way of teaching you think might be better.
- A possible location of new species.
- A fairer way to administer standardized tests.

It can really be *anything at all* as long as you can put it to the test.

Their purpose is to help you learn whether Random chance might be responsible for an observed effect.

What is Hypothesis Testing?

Hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results. You're basically testing whether your results are valid by figuring out the odds that your results have happened by chance. If your results may have happened by chance, the experiment won't be repeatable and so has little use.

Hypothesis testing can be one of the most confusing aspects for students, mostly because before you can even perform a test, you have to know what your null hypothesis is. Often, those tricky word problems that you are faced with can be difficult to decipher. But it's easier than you think; all you need to do is:

1. Figure out your null hypothesis,
2. State your null hypothesis,
3. Choose what kind of test you need to perform,
4. Either reject or fail to reject the null hypothesis.

What is Null Hypothesis?

The null hypothesis is a construct reinforcing the notion that Nothing special has happened, and any effect you observe is due to random chance.

What is Alternative Hypothesis?

Alternative Hypothesis the opposite of a null hypothesis and assumed as per the observation in the sample. This is counterpoint to the NULL (What you hope to prove)

P Values.

A p value is used in hypothesis testing to help you support or reject the null hypothesis. *Assuming null to be True, the probability of getting a data like the result in the sample or more in the direction of alternative is the P value.*

The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

Remember this poem--

P Low, Null Go ---> reject the null

P High, Null Fly --->fail to reject Null

Usually a 5% threshold is taken to reject a Null Hypothesis. This 5% is called significance level.

More reads here

https://www.analyticsvidhya.com/blog/2015/09/hypothesis-testing-explained/

Key Ideas Hypothesis Testing 2:

Court trials:

The process of testing hypotheses can be compared to court trials. A person comes into court charged with a crime. A jury must decide whether the person is innocent (null hypothesis) or guilty (alternative hypothesis). Even though the person is charged with the crime, at the beginning of the trial (and until the jury declares otherwise) the accused is assumed to be innocent. Only if overwhelming evidence of the person's guilt can be shown is the jury expected to declare the person guilty--otherwise the person is considered innocent.

In a jury trial the person accused of the crime is assumed innocent at the beginning of the trial (NULL HYPOTHESIS), and unless the jury can find overwhelming evidence to the contrary, should be judged innocent at the end of the trial. Likewise, in hypothesis testing, the null hypothesis is assumed to be true, and unless the test shows overwhelming evidence that the null hypothesis is not true, the null hypothesis is accepted.

Errors:

In the jury trial there are two types of errors:

(1) the person is innocent but the jury finds the person guilty, and

(2) the person is guilty but the jury declares the person to be innocent.

In our system of justice, the first error is considered more serious than the second error. These two errors along with the correct decisions are shown in the next table where the jury decision is shown in bold on the left margin and the true state of affairs is shown in bold along the top margin of the table.

With respect to hypothesis testing the two errors that can occur are:

(1) the null hypothesis is true but the decision based on the testing process is that the null hypothesis should be rejected, and

(2) the null hypothesis is false but the testing process concludes that it should be accepted.

These two errors are called Type I and Type II errors. As in the jury trial situation, a Type I error is usually considered more serious than a Type II error.

Another Popular example of Type 1 & Type 2 Error is -

A Type 1 error is also known as a false positive*, and a Type 2 error is also known as a* false negative*.*

Source: https://www.csus.edu/indiv/j/jgehrman/courses/stat50/hypthesistests/9hyptest.htm