Interpreting Data with Statistical Models
by Axel Sirota

Start Course

Bookmark   Add to Channel   Download Course

Table of contents   Description   **Transcript**   Exercise files   Discussion   Learnin

Course Overview

Course Overview

Hi everyone. My name is Axel Sirota. Welcome to my course, Interpreting Data with Statistical Models. I am a machine learning engineer at ASAPP, and I am very excited to present this to you. Data is everywhere, from the newspaper you read on the subway to the report you're analyzing about yesterday's top market performance. Over this course, we will start to give meaning to those charts and actually gaining the ability of making the real questions. We will go from zero to hero, analyzing complex patterns of everyday real-world problems. This journey begins at the very basic of statistics. From there, it will be a roller coaster of distributions and p-values, understanding how to make hypothesis testing, how to actually fit your data to a given hypothesis with chi squares, why ANOVA is such a key tool in comparing metrics, and finally, how to get that model to make your data speak for itself with linear regression. By the end of this course, you will be able to effectively understand how to tackle problems that appear at your work, understand which is the right statistical analysis to use, and how to interpret those results to obtain insights. I hope you will join me on this journey to learn how to interpret data with the Interpreting Data with Statistical Models course, at Pluralsight.

Thinking Like a Statistician

What Is Statistics?

[Autogenerated] we always hear about statistics. Even if we do not realize it, try this friend toothpaste. Nine out of 10. Dentist professor it The latest polls indicated Kanye It X has a 35% positive image. The chance of getting some rain tonight it's 94%. So better shape up or statistical knowledge, right? I thought. So. Then let's die into the start of statistics to get a solid foundation. So what is the statistics? According to so call? We define statistics as the scientific study of Tyra described natural variation. But what does that mean? Let's see the win. An example. Let's imagine we run a pharmaceutical study to understand the effectiveness of our new _____ X to reduce be LDL, which is a type of cholesterol. How do we do that in particular? How do we measure that for study? How do when a nice that data to get a result on? How do we know how accurate those results are? How do I know which confidence we're having them Statistics is a branch of mathematics that tries to give us answers to those questions. It's analyzing every individual in the world. It's super expensive and almost impossible statistics give us those two able to be ableto infer from those people to a larger audience. Statistics. We call that small fraction off all sample, and it's where everything starts like ways. The whole audience toe where we want to get a conclusion. Is that population going back for pharmaceutical study for understanding effectiveness of _____ eggs who need first to understand what the population is? A few examples would be all people in California with problems of cholesterol probably isn't going to be two years full, since there are more types of cholesterol than be at the El, then all people in California with problems off high bld. L probably is when it's about their feet, but we're still including children, which most likely we're gonna study for ethical reasons. 10 or people in California between 18 and 65 years old and problems of High B and E. L. This one appears to be a correct one for now. Usually when they say in the study, we need to sacrifice how many people we can target in the end for our study to be significant, probably in the 1st 2 cases, the population was to the birth for our study to find anything. Now that we have our population, we can take a sample. How do we take it? That's the sample. Look. Okay, select 20 people from some blocking California between 18 and 65 years old and with problems of high, be a deal. There are multiple ways, but in statistics we see to taking random samples on that is not a coincidence. Because if we don't take it that way, then we cannot expand our conclusions for more people than our study. Therefore, how about select 20 people randomly in California between 18 and 65 years old? And we probably took i b l e l. But that isn't enough, right? Because we want to analyze the difference.

Therefore, we need a group to give the _____ and I group to not give the _____, but are on the same conditions so we can avoid variation appeal the I because off diet exercise, he's whatever we call that control or placebo group, and it's extremely important. In the end, we get to select 40 people randomly in California between 18 and 65 years old and with problems of high bld I assigned randomly into tour groups. Control on experiment on this one is the correct one. Final. Yes, we have our assembled What are we going to measure? That usually is given with the intent of the study. And in this case, it could be the concentration in bladder be NTL. I'm How do we measure that? Variable? It haven't defined when we measure be a living blood. Certainly we want to abstract ourselves from the different levels off the individuals. We might take me on the Ellen bladder the start at the end to compare the difference. Using those concentrations with and without the _____ observation or data then would be each difference in me. Really? Ellen Black, because of the trail on the parable we're going to be analyzing is the difference in the LDL in blood. Because of the trial toe broke up. Our population will be all people in California between 18 and 65 years old and with problems of creepy LDL. From there we will take a sample by selecting 40 people randomly in California between 18 and 65 years old, and we probably okay, pl pl We were son randomly into two groups control and experiment. Our parable will be the decreasing be LDL worth a trial period measured by milliliters Birthday See liters off the LDL in black and each observation will be that decrease measured in each individual With statistics we will bury to understand how to analyze this data, take conclusions on with confidence. Extrapolate those to the whole population.

What Are Estimators?

[Autogenerated] now we're brothers. Example. If we've been a little in time, we can suppose the trail ended on. We got all 40 results off. Differences appear d I consideration in blood separated into two groups. What we would like is a way to summarize that data into easy numbers to report on those our estimators if we go back to our example. An example prescription and could see for _____ eggs would be used once every day in the morning. It reduces beyond the alien blood by 34%. Let's think a little bit about how the company arrived at that conclusion. Could it be that all 20 people in Experience group presented 1st 4% increase in bl bl in blood? Exactly. That's really unlikely because each person has natural variation. For example, we don't have the same height. Then how do we write the number? Answer would be they used an average day average is in operation. We use in a set of data to get a number that represents the middle. This is done by some in each Bob. You are then divided by the total number of bad news. So, for example, the average of 2 to 7 free. He is two plus two plus free plus seven debated before, which is Friedman

five. The average represents the center of the data, and it's a center statistic. It tried to say, If we measure the another value, which is the body toward this observation would be more likely to be close. Suppose for a moment we could measure the result of _____ X in every individual in the population took average over that immense data. Would it be the same EST average we took? Most likely, not the average over all. The population is a perimeter, and it's the true value every want to know but cannot reach the average. It's an estimator. That price toe estimate that perimeter in this case the mean. Basically, because it is impossible to take observations on all their population, it is too expensive. Many times we see the media saying that something has a chance of X percent. That usually means that every shot the sample that was taken East X percent, but two samples can have the same average and be very different. For example, check the picture. For that reason, we're statistics off this Persian, which will measure how far that is from the center. The main statistic off this person is the berry ins to tell us how far or that that is from the average, and it's defined by this funky formula. In the previous example, the standard deviation of 2 to 3 and seven is 5.67. Toe dropout. You're in a sample for a variable yourself and statistics to try to estimate it True value. If we could reach all the population they made, two are the average. We will try to estimate the mean and it's an idea off the center of the population on the sample barriers that estimates the population variance. And it's an idea of how far that are from the mean. Sometimes we see in the newspaper that given Candidate has a 30% minus plus 5% intentional vote. That usually involves saying that average on the bearings off the sample. But let's move forward. All this theoretical mambo jumbo Let's seat in a real jam

Demo: Installing Jamovi

[Autogenerated] it was time for them, All right. A lot of talk and no action makes no sense at all. Now we will see everything in action by installing the main suffer for our course, which is sure, Moby we will understand how two important data on we will calculate the mean on the barriers and see how they change in different samples. So let's get it. So I'm already in Germany, that orc slash download Where the webpage out of detective, My operating system on over the correct person to download for our purposes. The version 0.9 point 649 or later will be okay. Also, there were paged out, detected my operating system on offer the DMG because I used my quest in the case you are a Windows user. It will over, or an executable or a C filed to down low. So if we click, it will ask us to down low. So I will really have installed a Marie. I won't open it, but in this case you would just open and then accept handouts of the file. So let's accelerate space time a little bit on. We are already in Giamatti as you can see it's a lot like Excell, but with a lot of nice graphs, so it will be really useful for introductory tool for statistics. Also, your main borrow does not look like

mine in particular. It could not have this one or this one. That basically is because I already added the models, which we will see later on. But it will be done by clicking here and tow import new models. Gemma. We is spilled us. Amen. Self work on dhe. A set off models are plug ins that one can install, so it's really useful for extending it. So I already have a C S B already prepared. So let's go and imported. So we go to the main form we click on Import On On Brows. I already have a model to folder a file called one Sample that CSP. So if we double click it verities. Jamari went out of Detective Heather on it. We have a new Baruchel which actually is called BLD Ellen Blood. Let's remove these free columns because they are quite bothering me. This that I consist of one column with our beloved be LDL consideration in black parable. Inger Marie. It's really easy to make all those calculations we mentioned in the brightest lights. We just go to analysis exploration on descriptive Sze. Here is the main form toe Astra. Mommy, what we want to calculate on on which parables So first we go to statistics on and you're out of selected a lot of things, but what we care about actually is they're very ins on Damien. Maybe the sample size. It's useful to verify that we have 20 that appoints. So we select bl dealing blow, which is our bearable, and we push it to the right on. There it is, Jamal. We outa calculated the statistics and it's older. If we hide the options off the form, we can see on the left the data points on the right, The calculation of the statistics Now that we have already seen ___ __ _____ clay statistics, let's see what happens. You have two samples that I assure you come from the same population. So we going again to the main form. Andi, I said we'll have it in recent. I will in Brockport it from there. Here. Now we have two samples. Is that actually are appealing blood. So but the issues that they were taken from different people. So we go to exploration again. Prescriptives on again. We go to statistics and select what we care about. Which was the Barents Andi mean now with elect both variables? Because we care for both the statistics. Um, what happened? They mean on the variance changed on this is really important. Because if they're estimators change, what can we do in mathematics? We usually like to add levels of obstructions. So one time someone said, Why don't we treat them the average as a new random Maribel? And we can take a sample of Father. She's which come from a sample of different samples. Although this may seem crazy with this on some math, we will be ableto on Lee. Take one sample on be able to predict how this aberration on barriers will behave in your samples.

Basic Probability and Random Variables

[Autogenerated] take the umbrella tomorrow. There's a 94% privilege of frame. We've heard it a lot of time in our lives, right? And if it's not that one, surely, is this one? How can it be raining? I check the forecast, and there was only 5% growth. Refrain. What's going on? Brutality is all

around us, and it's the basis of false statistics. One of the problems with Stadium in the demo was that the average changed over samples, and we would like to control that. In a sense, we would like to answer over which values the average tends to move around or in a more formal setting, which is seen terrible over the average has a high probability of staying on the rest of the model. We will die in tow. What statistics can bring to these questions. So back lap. And don't worry, I will guide you from this attorney of fire off course. When may I ask, why would I ever want to know that? And it's a completely valid question. They made a bandage of trying to answer those questions is that after some prep work that goes up to _____ when encounters a lot of tests to answer almost any question our that I may have on being ableto infer from the sample we took to the whole population. If we think about it, taking a sample imply some randomness. So off course there must be some morality laying around on the subject. The important thing is to keep in mind that the destination is important, which is being able to make a lot of tests but also the route. Because if you take this inning one, you also get to understand the tests on a boy. Some policies. What is priority? What has the concept of chance in eight? But the way I like to see it is an ____ IG nation. Let's see when an example Suppose we're rolling a dice twice. We can create a set off all the possible outcomes off that nice 11121 free, etcetera A. Broil it. In this case, it's no other thing than assigning every outcome in number between serum one that will indicate how likely you will get that option. But we cannot go row. We have some rules. The problem is always between zero and one. We have some wine toe all on Syria to man on the union of Old cans has to some off each one's probability. He seems super abstract, right? In the case of the dice, if we say the chance of any outcome is one over 46 pieces, which is the total number of outcomes, that basically is that the deceased violence, then depravity off all would be to sum over all the outcomes. One hour 46 which is 46 over 46 which is one. So we're good. Then, for example, liberality that the first dies is even on the second ice is too will be there some. The probability of getting to two for two and 62 us. Each one has won over 46. That counselor's 0.0 83 and we're done. But why is this useful? Because then we may ask something like we would know am be such that the average of sample has a priority being between a young B off syrup with 95 and that's awesome. And remember, our Barry goes, If they can change from individual to individual, we call them random variables. So in summary around the barrel is some miracle we care about that changes pair in the middle. Examples of random barrels could be the result of a dice, for example, here a possible outcome. It's free. On important thing is that we can only have interesting results, so we call them discreet random barrows. Another random variable could be the concentration of the LCL in blood. Remember that one here? A possible outcome is 14 to 45 and we can go one around so we can actually any decimal result. Then we call them continues random variables.

Continuous and Discrete Distribution Functions

[Autogenerated] some time in our lives. We always heard about distributions. For example, the law of diffusion of Innovation says that all people are distributed like a bunker and splits the population innovators, early adopters, early majority late, majority on bloggers, they the other distribution. Then it's a way of describing how around the barrel behaves, what are the most likely values will take. So, for example, about if we would measure how innovative some random person is most likely here, she would be in early early majority. Hence the name majority. Now that we have introduced random Baracus Emperor realities, we will immerse them to get the best way of analyzing, um, this quiet around the bearable if distribution function. Let's imagine we want to understand how a number of identical twins per family through it in our state. So we go to a statistics and found out that in our state the proportions are syrup pair of identical twin 75% 1 17 to 5 on three or more free. Therefore, if we cut around the barrel assigning the number of Peru identical twins to some family, we could guess that the Prelate of experience zero it's European 75 the relative ex being wine syrup 750.17 probably expecting to its surprise your five on the product if x mean free or more, it's Sierra points. You're free or in a chart. It would look like this if we call F in a given value X Toby. The probability off our random barrel Explain that value. Then, for example, of our identical twins, then F off zero would be syrup on 75 if one would be Sierra 10.17. If off to would be 0.5 on ever free would be syrup. Points here are free, as the image says in this case, as we cannot have 1.5% identical twins. X was a discreet from the Margaux. Therefore, we call F a discreet distribution function for X extrusion function for X would be a function that, for each possible value that run memorable, assigns the probability for random barbell being that value. Now, if we can see her again, the concentration the LDL in black for people, what is the probability that, for example, given in the usual have 12 milligrams per milliliters off the LDL in blood? If we measure with super tool, then a measurement off 12 would be 12.44 on we're Even Better Tool, he could be 12.44 56 82. Therefore, in reality, we can have infinite possibilities for X X can have any decimal value. Therefore, us one hour Infinity kind of looks like zero is really, really, really small. The probability that X equals 12 0 and that is the same for every value. It would be useless to have a situation function like the one before for this case, because if off X would be zero for every value, seems that probably d off every value he zero. But if we consider quality of being in a range that it's a narrow tail in our example appeal dealing blood, there is some chance of Pierre de LB in between 12 and 14. It could be low, but there exists in this case. What statistics do is to call a continuous distribution function for X. It functions such that the area under read is the probability in that range. I know they seem super abstract, but it's more common that it sounds going to the law of diffusion of innovation. When this law says that people the truth in a

male, let's try to crack how this works. What this law is saying is that if we said ex Toby the degree of innovation, some person cast and check the distribution function for X, then this function, especially aped, meaning that the prelate of some person being an early majority or late majority, is actually the area under that curb. That means the barely. This type of _____ parable is usually called normal or counter distribution, and it's the Messias statistics. We'll encounter that bad boy again in a little bit. For now, let's do a recap to review what we learned this last videos, which I admit we're kind of abstract, a theoretical. But they are useful, I promise. A brutality is an assignment off a possible outcome. Toe a number a random barrel is something I care about the changes over each individual of my sample. In the case of that barrel, being discreet, we can have its destruction function a step reality off that run unbearable being that value. But in the case off continues ones. The extrusion function is a function such that the area, and there it is, the probability of the random bearable being in that range with all this prep work, who will be back to analyze the average being understood around the Maribel and get to tame it?

Does the Distribution Change in a Sample

[Autogenerated] and the best videos we've seen that the average can change over different samples of the same population. That is the end of the world for our beloved farmer. Tried, right? How in the world can I say anything? If the baddest I measure actually are not understand? Don't panic. We're here to help. Here is where the central image fear, um comes into play to save the day. Yeah, we won't be able to say the decreases for the 4% for example, but we will be able to say that increases at least 23% for example, with 95% probability on that rocks. By the way, it's great that now we know what that means, right? So let's understand that that's imagine we gather free samples off 20 people for our former trialled on all free take drugs. In the end, we measure the average percentage of the crease of the idea in blood on, we have pre values. For example, for simple one, we have 14% for sample to 24%. I'm for simple free 32% If we go together, the knowledge we've seen the last to be Dios than the average on an estimator. Actually, it's a random Baruchel where my different observations are actually sample. I know. Mind blowing, right. Let's do a quick overview for the parable observation sample and population here. Here. Each number we're interested in is actually the average percentage of the crease of the ideal in blood. Therefore, that is our bearable. We're moving up one layering obstruction, each observation or that would be a single value for her parable, which in this case is an average percentage of the crease of pl pl in blood. Then assemble of 20 observations would be actually 20 samples off 20 people each from the original population. This means that it our sample has 20 observations then, actually, we're checking the trail on 20 times 2400 people as we're taking one

value per sample of people from the regional population. Then the population now is all the possible samples off 20 people in California between 18 and 65 years old, with problems of high V a. T. L. As always, we're moving up one layer obstruction about from all these three questions Can meteorites. What is the average and variance off this new random. Maribel, what is the distribution? What do I actually gain from all this nonsense? Let's answer them little by little. But the main idea I start with all these, we will be able to actually make an inference on the regional variable. As we know, average members assemble barons, our estimators off mean and population variants. Actually, mathematics tell us that they mean off the new random bearable, which is the possible averages is the same. We don't lose anything on. Actually, the parents is reduced by a factor of the sample size. This is awesome because it means that if our sample size 20 nor example is large enough, then there is a very high chance that the Irish ease off averages. Actually, sweetie, close to the original mean this is why it's useful. All this obstruction from here, we can actually in for the population mean that true average of the population from one sample on Dhe then know if our track trial actually reduces Pierre deal in blood. But there is one thing Mason, right? You're already thinking if it I'm smelling it, how do I calculate the probability you and I were? It's what is the distribution off in our case, the average percent of the crease of the LDL in blood. And here comes the boom because it is with deepest pride on great displeasure that I present the central limit. Fear, um, which basically says, if you're some besides is large enough for us will be more than 20 then the average off a random Maribel. Anyone has a normal distribution. But, Axl, what is a normal distribution? Remember the love fusion of innovation that bell shaped distribution wants the normal. The normal distribution is the bell shaped distribution that has to perimeters mean I'm variance. That is why we take so much interest in them where the mean controls, Where is the center? That means where it's more likely to have values on the barriers. Control how wife the population ins on. What about sums on differences of normals? Because we always take differences off different samples. Well, mathematics also tell us that the same on differences are also normals where they mean off the Sam's. Some of the means, the mean of the difference is actually the difference off the means. On always the variance is the stain is the sum of the barons is we don't care on the formula. But what is crucial is that the probabilities for normal distributions are known and easy to calculate when the minimum variance are known. So we have a recipe for success. Suppose we have now our all too samples drag on control and we want to test that our drag actually reduces. Be a deal in, lad, I said. Extra tip. We already know that the parents of pl TL in blood for the population is actually one milligram permanently. How we do all these, it's clear. Now we have a recipe. First, we calculate the mean for both groups. If the sample size you speak more than 20 for us, then we already know that both overseas are random variables on our normal because of the central image. Fear him.

Therefore, the differences off pull flourishes isn't normal. Where they mean off the difference run memorable is actually the difference of the means. And we know the barrens because if the sum of the parents is which we know which waas one over the sample size 20 now we go to Tamami. We calculate the probability that the samples were equal. Therefore, that means we're the same. We got the value that we got That is the same. To calculate the probability that we got a value greater than the difference off the actual sample averages when we are using it, this normal distribution off means zero on barriers to over 20. If that brutality is low dim, that means that if the samples were equal, we were so lucky that we got a very different value. Then we can conclude safely that assembles we're not equal because we don't believe we're like on Dan. Let's see all these in a demo, which is far easier than it sounds.

Demo: Making Our First Statistical Test

[Autogenerated] on. We got here After a lot of coming and go in some abstract theory on a lot of hearing me talk, we finally got to the real demo. The first test recrimination off a lot of understanding in a single real world application Congress to yourself forgetting here it's a lot in this demo. We won't see anything. Actually knew. However, what? Why? We will take em to end our bl bl concentration in blood. Example on crack it all the way through. We will identify our first hypothesis to test embarrassing play. We will perform our first test for difference of means and understand the results. It was a great journey and he would come even better. Let's go. So I'm already in Germany again and I already opened and knew that a file called the Memorial to as always, you will have all these available in the course materials. We have our two samples. They're increasing the LDL in blood of percentage on the control on the experiment. If we look at them, it appears that experiment east different. But we will do our first test. So if we go by our recipe, the first thing we need to do to create a new barrel, which is actually the difference off off parables. How do we do that tomorrow? If we go to Tara Weekly gone ad, we can add a new computed variable computed variable is nothing more than a computation of different functions on different variables. So here, if we click on effects, we can see both parables and then we go decrease militarily blood with double click it. We will do the difference with B I E I for the experiment. So if we click Ender were then then we can hide again the completed parable and we can see that our new barrel, which now is called C we will change it. It has the difference off the creases of the LDL for both samples. We double click, See, we can get back to the form and I will change these two difference if we hide it again. So we go to analysis exploration scripted and we want to know for our variable what is the mean and barriers. So before we select mean and variance, this is like the other ones we go to difference and we can check actually that they mean is very negative on

differences to 0.48. So we have to continue on our test or this deal now have to calculate the durability off carrion minus 17.2. Supposing these two barrels are actually the same. If they come from the same publish in, they would have the same mean. Therefore, the parable difference should be a normal off mean zero on barriers. Two over 20 for that is that we have these model destruction. I will tell you how to import destruction will be a plug in that allows us to calculate the probabilities are normal and other illustrations, and it will be really useful for installing a plugin. Giamatti, we click on plus models show Mommy library on we goto available in available here. Issue scroll down. You will find destruction here. I already have it installed. As you can see, it looks like these destruction. Quentin said parolees for continuous and this redistribution flashers. So once you installing, you click it and we wait. We can hide and we can use the distraction library. If we click on destruction, you will offer us how to calculate priority. Some Paris institutions, including the normal one. So now we go to a normal distribution on the first being that destruction will ask us is a parameter rise the normal distribution we think we have in our supposition, the mean 10 on dhe here it asks for the standard deviation. The standard deviation is the square root of the barriers on for some recently movie likes to talk in terms of standard division instead of parents. But if the stains just taking a square root. Therefore, if the variances two over 20 the sandal aviation is Sierra 200.348 Already the miscalculation on what we want is to compute the probability off. Karen, our about you minus 17.2 or lower on, we can check in the graphic. Actually that well, iwas extremely infrequent. The Bryant actually get anybody we got. Supposing the two samples were the same is zero. Now we have two ways them the 1st 1 we can think. Okay, the breath iwas zero. But we were like on They are the same, but that is not what we're going to do. We're going to say, as this product is very, very low on, I don't want to believe I'm lucky The superstition was wrong. So the two samples are actually different. On that is the test. As you can see, it was way simpler in Germany than when I told you on the slides. And this is what we'll do all the time. We will suppose something you see, quote, then based on the position, we will get extradition with that distribution, and that's a position we will get a probability. And if that priority is low enough, we would talk about it later. Module About what? Slow enough we will say. Okay, now I reject that supposition. I was wrong. Therefore, this supposition waas not true on that's it. Now let's go back, make a hole summary of what we've learned in this model and stay tuned for the next one.

Summary

[Autogenerated] we covered a lot for a model. I was there that this last half hour, actually lot off weeks in a standard statistic Course so good for you. We started actually the very basics to

define what is a population observation sample. And Maribel, we saw that variable. It's actually what I care about on observation is each value measure in a sample, which is a small part of the whole population Over a given sample. We can describe it by estimators that try to resume to numbers information about our sample. Thes estimators actually changed over time, while their true value, the perimeters don't. So we took the long route. Understand that a probability is an assignment off any outcome toe a number on that. These different outcomes were our old random variables, combining both. We got to the concept of distribution, understand how the random Maribel actually takes its different values and predict which ones are most likely with all that. I'm full power when Alyce estimators us random barrel themselves on a ride to the conclusion that the average off any Maribel with any distribution can have a normal distribution. If the sample size is speaking up that was the key to create our very first statistical test in the case that the parents is known. But what will happen if we don't know the barriers? If we don't assume anything, stay tuned for the next model where we will uncover T tests.

Testing a Hypothesis

Introducing Case Study

[Autogenerated] on the brilliance model. We did a lot of work and could make our first test. We evaluated, if two samples wearing it from the same population or not. But we suppose that Paris was known. What if it's not her? We lost off course not. This will be a deep dive into statistical tests. Toby Ableto make us many comparisons as one one's Sky's the Limit. If the sample size is big enough, let's imagine. Now we're working at an investment bank on. We're taking the decision for other people's money. Let's make sure we're making the correct one right. We were assigned to the side or working best next between telco hospitality with one simple move Yet we must decide based entirely on driving you. We want to look at more things stadium for next models. Therefore, we grab revenue for last quarter from 20 random public companies in each particle. I would like to know which particle gets the best revenue, how in the world we do that. We certainly don't know the bearings off any of the bird occurs, so we can't use the same recipe with last model. Here is where the T tests will come into play

T Distribution: When the Variance Is Unknown

[Autogenerated] as we've seen the case in the last mother was kind of odd because we knew beforehand the parents of the population. In normal cases, we don't know which perimeter, however most paralysis treated normally. So we would love to say that even if the parents isn't know, the simple average is normal. But can that be the case in this gaze, where the Berries is not known, there's an extra certainty that has to translate in institution that allows for more extreme values. That means more flat. Also, if our sample size grows bigger and bigger, that decision needs to be close to a normal because of the central limit theorem. That destruction is called the restriction, and it looks like this. Basically, it's the normal, but with bigger probably descend more extreme values because we have more uncertainty since we don't know the barriers, and if we check the image, we can actually see free things first. It has to stay in shape as a normal distribution. Secondly, is a family of curves on us assemble size is bigger. That's the sub script. It's more like the normal. Finally, we can actually calculate probabilities using the mowing and this is great because with this knowledge were back in track, toe or investment bank problem. Let's talk a little bit about the history should mean a family of curves. That is because it is perimeter. It's also by the degrees of freedom, which tell us how many observations are independent and give us an idea of how sure we are off our estimators. In that sense, let's notice that as the number of degrees of freedom increase, our tea must be closer to the normal, because our estimators are more accurate and we don't have that much parents anymore. In our case or our comparisons, acres of freedom will be the number off that appoints minus one. Going _____ example of investment bank. We now know that averages off each particle. Actually, this really like the student off, maintain the wish of freedom. Then again, this is because the sample size was 20. Then we will use our problem from comparing this party guts toe a problem, how to compare two and possibly with different degrees of freedom, he stealing distributions. The correct way of comparing statistics is using hypothesis testing, and this is where we're heading

Statistical Hypothesis: What Am I Missing?

[Autogenerated] people ready, then intestine last model. But now let's leverage that time frame it in a way that we can compare anything. So what is really a statistical best? It's actually a process to ballet, some assertion over the population based on a single sample. It's an assertion over the perimeters, and it's a decision rule based on the distribution of given statistic that sounds super abstract, right? Let's go over it with an example in the spirit of finding work. In best were Reader The New York Times, that Romany for hospitality we're less quarter decreased in average $23,000,000 as we want to check that, we take a sample of 20 different hotels and gather their revenue changes over the last quarter getting the following statistics on average off minus two

point $6,000,000 and the parents of minus 2.5. That's a statistical tests need to be over the perimeters that this would be the average change in prison. You in the population is minus $23,000,000 in the common for a work what we just stayed. It is called the No hypothesis or hypothesis off. No change. It basically says I will suppose whatever it's known until now. In our example, the North hypothesis is the average changed in revenue is minus $23,000,000. The other hypotheses the one we want to prove. It's called the alternative hypothesis, or hypothesis of Action. It is what we want to prove, for example, in our case, that every shred we knew for our hospitality wasn't actually minus $24,000,000. Ham the news in New York, Times or force. And how do we prove that? We suppose that the North hypothesis is room on the supposition We calculate the probability of getting the sample we got. If that rate is very low, we're rejected all hypothesis because we don't believe we're lucky. Does it sound the bell? It was the same. We did last model, and we will keep on doing it until it becomes automatic. That quality we calculate under than one hypothesis is called be value on the fresh hole. We used to compare the P value and decide if we reject is called significance. This is the big guns, actually comes with the test, and if they find out the stark, so when we see the conclusions we read. The average revenue was minus $23,000,000 with significance off 5%. Usually significant sees 15 or 10% on the compliment of the significance is called the confidence. So you better significance was 5%. Then the confidence is 100 minus 5 95%. That was a lot, right? Let's go back to our example and check all these. We took a random sample and got miners two point $6,000,000. What does that mean? We could be lucky. But the nice thing about all we know is that we can calculate the probability of how lucky we would have been if actually the ever changing the population was minus 20 free. And there they know hypothesis. Then, from the first video, we know that the average this reels like a T student off mean minus 20 free on barriers five over 20. Remember, disease, because the sample size is 20 ham with 19 degrees of freedom. With those values, we go to jail Marie and calculate the P value off actually getting minors 2.6 or more extreme on in this case, Jamal, we will tell us that this problem is almost no. So as this brutality is below the significance that we set a syrah 0.5 we're rejected all hypotheses. Therefore, with 95% confidence, the average revenue last for hospitality is not minus $23,000,000.

T Test: Estimate the Chance of Equal!

[Autogenerated] But in our example of the Miss Mom Bank, we wanted to compare two samples. It seems like we got close enough, but not there yet. Let's fix that problem right now. Checking how what we saw actually already solves Comparing to sample salsa the two most confidence We may want to compare our averages off your miracles. For example, average revenue off

different verticals or proportions for candidate John Doe. We want to know if Rotary support him more than non retirees. For example, let's start with them with what we saw. If we have to population still hospitality, then each has its sound mean and we want to check if they're different. Therefore, the North hypotheses needs to be that they are equal. So then the hypothesis will be that it means our equal on Delta on the hypothesis that they are not in the progress. Pdo We compare the barrel against given value, and that's a good idea. So if we rearrange these hypotheses we can get, there's no head office is being. The difference of means equals to zero on the alternative hypothesis that the difference off means it's not zero. And how can I estimate the difference off population all means with averages, of course. So our estimator needs to be the difference off our feast between tell hospitality. Always take care off the order if we put always Telco Minds Hospitality list so that all along the way. Finally, then we need to know the distribution of that estimate. We have three possibilities. The 1st 1 which is extremely rare, is that we actually know the Barents is off. Welcome hospitality In the case, as we're supposed from the South, our variables of normal then both averages are normal on the perimeters. Then would be those ones were really so that in the Bruce Mother. But we don't care too much about the math, because in this case, we go to Germany on we calculate the priority off getting the difference we got with these perimeters in this normal distribution. That's it. Let's check all these with an example with significance off 10%. Suppose we got a sample off size Tim for Terry Go. The average revenue was 14 on the variance five on hospitality. We got 15 samples. Actually, never trailing you was 12 on the variance was 7.5 then under the null hypothesis, we want to know the probability off getting 14 mine was 12 to within a normal distribution off means zero I'm variance five worth 10 plus 7.5 over 15 In Germany, we can calculate the probability has beens 0.0 23 therefore has be value is less than the significance of a test. We can reject the hypothesis that the birth occurs we're equal. Therefore, we proved that telco provides better revenue and hospitality off course. In this case, it was symbol because we need extradition was normal and we can calculate that probability then using the normal installation calculation In other examples, we will rely more on so forth to do these on those are the desks so we don't have to actually know the variance and difference of means of the disillusion But the mechanism is the same for that We will do the demos after these video going back to the three possibilities. The next one would be that the both parities are not known. Then we have to know in which case we are if the birds are similar or not like we can see on the picture differences are similar. Dim our barrows and normal than the average is not actually our estimator city student on it has these perimeters off course for the variance I put it. Who cares? Because we have software. Therefore we are great because we can do that easy in this offer. Actually, we will check in the m o that that it's time for us. So we don't need to plug the barriers. Hurry Mind that this is another case where viruses are in similar. Then our estimator will

be it a student, but with different diameters more complex. This case is not that common. Therefore we will not cover It were time right now I promise you that we were comparing means on proportions. Proportions actually are not easier because if the sample size is big enough to remember that for us is more than 20 on the proportions have not so close to one. That means that there is not a vast majority which we translate that be times and is more than five on cue time. Sam is more than pipe work. Thank you are to proportions. We compare then estimator difference of proportions. Actually it's normal with these perimeters. Um, also, it's taken care of by general without having to play all the perimeters. So now with all this information that's poor of this in the demons and let's do some rock and roll and compare.

Demo: Comparing Means

[Autogenerated] on. We got here again to Damon time the good time where we will revisit the investment bank case with all we've learned over the following day, Mom, we will make an average comparison of revenue for telecom hospitality. We will validate assumptions off the test on analyze further what doesn't mean their results. So let's go. So what's always I'm here in Jamadi. Andi, I already prepare that set for these. _____. So we go to the main Freidel. Andi status that is called Memorial Free means I always will find that in the course materials we will have to comes. The 1st 1 with our bargain is actually the revenue for each company we assembled, Onda Stig unbearable is going to be our categorical variable that actually separates the revenue which will have to levels telco on hospitality. Now it's really important that we understand why this is disposed this way because actually, we have only one barrel revenue issues that we are measuring it in different samples and we want to compare if they're from this same population with all these in the movie. We go on, we click t tests on independent samples, detests there to our parents of the 1st 1 when some hypothesis isn't bad But this is not the case. So we go to independent samples to tests On first, it would ask us a lot of things. So what is our hypotheses? Were hypotheses actually east? That telco on hospitality are different? We cannot actually say one is bigger from the other one from a sample because they help others is need to be stated before taking December we will do a STI student test These other two We don't have to take care into that. Those are minor details that we were covering Next courses We don't have any missing values on additional statistics. We will ask for the mean difference on we were less for some descriptive sze. So we understand what is going on. So we go We say that our barrel it's actually the revenue on the group. Emeril is the particle on. We're done. If we hide, we can check. We have our result how he sees interpreted. We can see that we did an independent samples to test the statistic. It's minus 4.87 degrees of freedom is 48. So that means that Actually, the barren seas

were similar. And then the rigors of freedom are the summer. Each one's on the P value. The suffer says it's below Cyril going 01 Therefore, with the significance that we were talking about 0.5 we know we can reject the new hypothesis. We already know that. Tell corn hospitality our end the same. Actually, the main difference. We got waas minus seven times $97,000,000 Andi. But the sample error we got was 1.64. Also, we have some group that's treated so we can understand for their own with we have the mean 1,000,000 is another statistic like mean? But we don't care about that on the sample error and we can see those are kind of similar off course. We didn't do a test to validate. The similarity of the parents is we will do that in later models where everything will come together on on the plots. We can understand that telco indeed, waas greater hospitality and reverence is wearing that different. Also, the plot is really informative because it enforced immediately mean within 95% confidence interval, which we will understand that the end of the model what it means. But it will give an idea where the average revenue for each of the verticals actually is. That is really important what we've seen because we took on Lee one sample per vertical and we made a conclusion on their populations.

What if I Am Mistaken?

[Autogenerated] Now we have a clear picture. What that's on doesn't mean to reject or not reject The new hypothesis. To reject all help offices means that there is enough proof to say that the null hypothesis is false. It doesn't mean that alternative hypothesis is true on the other side. Not rejecting all hypotheses means that there isn't enough evidence to say that in our hypotheses falls, it doesn't mean that I know hypotheses isn't false. Northern alternative hypotheses is in force. Also, the test is inconclusive. That's off. And what if I am mistaken? We can split the world of possibilities in the following two times to table the four. Chances are that I did not reject the more hypotheses on. Actually, this one was true. So I was right that I did not reject an old head offices on. Actually, this one was falls. I was wrong and this is called type to error that I did rejecting all high purposes and this one was actually true. So I was wrong and this was called Tiguan error on that I did reject an ally purposes on. Actually, this one was false, So I waas right. Let's think what we're doing with significance in this context, we're setting that. We rejected no hypotheses. If the value is below the significance, therefore, the probability off me being lucky on actually the null hypothesis is right. It's very low. Therefore, we're actually setting the type one error. Toby. The significance on why is that? Now we're investment bank is difficult to picture the difference. So let's suppose for now we're in a trial. We have to decide for the life of someone else. General purposes them would be that the person is innocent. Therefore they take one error means that someone was innocent and went to death row and died. That's really bad. The Ted to

error means that the person was actually guilty, but we set him free. Now we can see that on the first scenario there is no backing down. The person is now dead on. The sentence was made on the 2nd 1 We can make another trial. Therefore, we try to control the significance of type one error because it can be catastrophic. And can we do improve them the type to error by improving the sample size So the parents of the sample reduces

Confidence Interval: The Other Face of the Coin

[Autogenerated] Have you ever checked the blood analysis one usually has to take. Sometimes on example result could be your glucose is 80 plus minus five milligrams permitted leader with 95% confidence. What does that mean and how that correlates with the confidence we were talking about? Let's analyze that because it's the other face of the kind of detests and it's called parameter estimation. Sometimes we just don't want to compare things. We want to estimate some value. For example, what is my glucose? What is the intention both for candidate X? What is the improvement in sales by this new feature? All these times we want a number that give us a way of understanding a perimeter. That problem is what we call estimation. There are mainly two forms of estimation. Foreign destination. The perimeter is estimated by a single value. I take a sample of blood and the glucose was 80. Therefore, your glucose must be 80. Regardless of new samples. I'm confidence Inderal estimation. The perimeter is estimated. I am in terrible number plus minus. Some error such that the resemble reality. The true value is there on example would be I take a sample and say that the glucose is 80 plus minus five with 90% probability. This one is the preferred way because it says something about any possible other sample. Mmm mmm. Confidence. I remember that name. Yes, it was one minus the significance, and we did not use it much before him. However, let's see why all is connected. If we check the type one error off estimating evaluation in terrible, no matter how we calculated it would be. I rejected that the parameters within that in terrible. However, that wasn't true. Therefore, the privilege of that event is the significance. The opposite of that is that the probability that the perimeter actually is within the terrible, which was the priority we mentioned about what we call confidence. That is why confidence intervals are the other face of the time. Instead of putting our minds in comparisons where we want to reject something that me and Sam per meter is not in a given distribution. And we look at the significant We checked the opposite when some perimeter is in the exclusion and therefore we look at the confidence looking forward. Um, what does he mean for the glucose to be 80 plus minus five with the confidence of 90%. It would mean that if we sample 100 times blood 90 times the average book of sample would be within that in terrible on their 10 times. It wasn't EMP, really? And how do we calculate them? The upper and lower limits will always have

the same form, which will be it. Be constant times the square root off the Barry and Sarah of this amber. We're B stands for a constant. There will be to the upper constant of the lower constant on B E stands for the Baron's era of example, which will be the parents off the sample, the birds of the sample. We already know what it looks like this ever cases from what we've seen, if we tried to estimate an average, we know that in both cases it involves a Barents relation Alert a sample over the sample size because we tried to estimate the proportion were in the same case why it is important to know that the confidence interval has that form. Why do I care? As we know that entire being smaller is good, we won't waste of reducing it with you singing terrible, then translate to reducing the baron's off the sample. Since the other part is constant, I'm from the table. We can understand forgiven confidence how having a bigger sample size or lower bearings of the sample means a Peyton's parents error found toe a final super tweak. What does it mean in this context, for two means to be different with a certain significance. It means for the confidence intervals for that significance to not overlap and everything is connected and makes sense now.

Summary

[Autogenerated] on Dhe as fast as a Ferrari. We got to the end of these model. We cover the lot, and now a lot makes more sense. I bet you now you can go all bragging about all the significant differences, and coffee is intervals to your teammates. That one was a model huge of content, supercritical toe. Our mission. Robbie Nap, A statistical test consists of a set of hypothesis Nolan mentality on the significance. We also come on estimator, the one that estimates the perimeter that the new hyper offices talks about and uses what is known to be the distribution that we can fit from knowledge or software. With the distribution and under the no hype offices, we calculate a P value of the test. That means the probability of getting the value we cut in now assembled, if then all hope of is, is what's true. In case this P value is lower than the significance, we can say that we reject the North had purposes. If it's not the case, we don't say anything we may need a new sample on this process is the same for every comparison. If you look for tables, you will find a distribution perimeter and deaths for almost any estimator that you may want. After that, it's just playing into the software. Compare on by the 18 assumptions to reject an all hypotheses when it was true is very serious, and it's cold type one error. That is why we control it with the significance but the accidental head offices. When it wasn't true, it's not that serious, and we call it type to error. We try to improve it by sampling size on better tests. Finally, we gave a new twist to understand that now we also can estimate the actual perimeters with the same knowledge and connected to tests with confidence intervals. We understood how these are made on. Then we

can we let himself or for calculating them. But with the clear understanding of how this works, in case we need them to get better and a new sample, I am so brother fast we got so far now in our future comes a quick but very useful topic. How to test if some data actually follows a given distribution

Comparing Categorical Values with Frequency Analysis

Introducing Demo: Salary Fairness

[Autogenerated] and now we know how to test things. But there is a simple issue. Right sentence. We have more than one person left. Um, tow an experiment. For example. We want to decide the best brought in for our dog between fish, chicken and pork. We want to know if our diet is balanced. We want to understand the boat intention for different candidates. You know, this case is we faced with the problem of multiple comparisons. The 1st 1 is a continuous parable. Split it by categories, while the other two variables is a number one for 10. The following models will help us understand how to tackle these problems. Excited? Let's go. Before we were at Investment Bank, remember now with Sweet Rose again, let's suppose we're on HR Compensations manager at a company called Romantics. We were assigned to analyze the income distribution for our company. Apparently, the job market changed in this crazy times on our scouts were out that the fair income distribution should be the following 1 22% of the company should be in a salary back it off under 40,042% between 40 and 50 for 27% between 50 and 60 11% between 60 and 80 6% between 80 and 102% over when hungry. So we take a sample of 5000 and nurseries for a company, and we want to know if we are being fair. Sounds like something every company should do every year, right? Let's figure out how glad romantics is doing.

Goodness of Fit: Let's Test if This Follows The Pattern!

[Autogenerated] the broom where we want to verify some categorical latter follows a certain extradition is called Goodness of feet and extremely useful as this is a statistical test, we already

know the steps. This would be a recurrent pattern. From now on we saw all the theory. So now everything is following our recipe. Therefore, we need an all another alternative hypotheses on estimator to know the distribution off that estimator and finally took a greater people You on the side and what would they know? Hypothesis local here. One possible would be the human distribution is correct Basically since that's the only thing we know in more there if we could rewrite it US all probability certificate, they always are the one stated, in which case they'll turn. The hypothesis needs to be There exists at least one morality that it's close Now We have said that the hypotheses we have to go to our estimator We want to cover our frequencies with some theoretical probabilities. For example, they probably had a salary below for decay. Waas Sierra 0.22. That's a theoretical probability. What do we do? One way of doing so under the new head offices if the product of some categories, for example, 0.22 on we got assemble off five pounds and then expected number of cases for that category would be 5000 times 0.22 equals 1000 and 100. That number is called expected frequency. And how can we compare off served and expected frequencies with Cheese square? The square is defined by summing over all the values that differences between the Sarah expected frequencies it counts. How different are they expected on a separate frequencies as they are more different that she square will be bigger. But what's the solution? What I'm actually assuming to make some decision over the statistic on we're back to another name distribution. In this case, the T score statistic will follow a CI Square distribution. Let's not fear it because it's a lot like the tea. The T Square decision is the following one and is the usual decision for the barriers off a normal variable experiences positive we can see them that that she square is positive is a symmetric on like the T depends on the degrees of freedom. But as we know how to treat its distribution. We are not scared. So going back to your problem for calculating the P values, we need the degrees of freedom. In goodness, if it tests the degrees of freedom, follow this equation we're in is the total number of data points on M is the number off perimeters we needed to estimate, for example, if we need to verify normality and we need for that to estimate mean I'm variance, then em with equal to let's look at an example, Suppose we want to compare these absurd frequencies for different blood types to those expected frequencies. That, of course, we come from expected probabilities. But we already have expected frequencies, then what we have to do first, calculate a statistic. This is the formula for that statistic, and we get nine doug 49. Then we have to get a P value. So the number off the grease of freedom would be four minus one, minus the number of parameters we need to estimate she's zero. Therefore, they are free. So we have calculated operability that our chief square statistic is greater than mine 0.9 49 and that reality is equal for free degrees of freedom. Sierra 0.0 23 off course. This morality was calculated in Germany. Therefore, ask that priorities below syrup 0.5 We can reject those expected frequencies where the correct ones and we made

our first goodness of three tests. Well, these were then we go to the software on what we're telling us what we will see. And the some shells assumptions of these tests are the total sample size has to be more than 50 old Expected frequencies should be positive on the amount of expected frequencies that are below five should be less than 20%. All these assumptions are in order for this statistic, Toby well approximated by the Chief Square distribution so we can calculate the probability with all these, Let's help our HR compensations manager.

Demo: Salary Fairness

[Autogenerated] we are super prepare for these or the following them are We will categorize the different sellers for global man ticks. To get himself frequencies, we will make the goodness of protest against the probabilities and finally to Bali. Assumptions off the test with all these will be able to help our hr compensations manager, Bring it on. We are ready for another day, Mama. I won't really in pretty They knew that a set for romantics where we have on the first column the frequency off salaries so that you see here 1550 means that there were 1550 salaries that well below $40,000 on so on and so forth with the other packets. So we go to frequencies hand Emery off her ____. The NL comes one sample proportion test, which is the cheese square Goodness off the test, which is what we saw. So we take that, um we have to input us always the parable on DDE in the case that we have categorical badges us now the frequencies So they were really Is the salary back it heated? It defines what are the different outcomes on the counts are actually the frequency with that way, see on the right that for example, between 50 and $60,000 there were 1350 persons which wants the 27% of the population off course. Those proportions were in the ones we wanted to test against. Therefore, what we want to do is actually goto expected proportions and feel on the _____ that we wanted to test against. I will accelerate space time a little bit. Toefield those _____ right now. Um, there it goes. I already into the alteration on as the ratio summed up 100. We're good to go on the right. We can see that already. That she square goodness off It was field for us where the chief square was calculated US free family and 68 with five degrees of freedom. The five degrees of freedom off course is because they because of freedom, are m this case six minus one minus the number of parameters we estimate which in this case was zero therefore, six minus one minus zero equals fire the 3 68 We're a build, basically, by comparing the off the count that we observed with expected frequencies. How do we get those? Well, we can select on expected towns. I'm here. We have them. So, for example 3 68 We wouldn't do it right now, but it would be 1550 miners, 1100 to the square, divided by 1100 summing up all those different categories And that sums up to 368 now, also tomorrow we already gave us the P value off our tests, which waas below 0.1 That means

that for the 5% significance that we set for our test, we are not following the given proportions that our scalp found out where the flare proportions. So we need to start helping romantics to find out what happened. We can see that the proportion for the first category under $40,000 is actually a lot bigger. That will need to be so. For example, that can be an issue now, off course. We're getting a fair for ourselves because we haven't evaluated assumptions yet. That's high form. And remember the assumptions for the goodness of fruit test. The 1st 1 What's that? The total sample size was about 50. We are good too. Though we have 5000. The 2nd 1 wants that the number of packets that Hal if expected, frequency below five, 120%. We have no back it below five. So we are good to go. And finally, all expected frequencies needed to be positive. Also, we're good to go. Therefore, as we accomplished all assumptions, we can be sure that our statistic she square is well approximated by the square distribution. And then this p value is welcome, lady. So now we are sure that we don't follow the proportions that the scouts actually told us that we needed to have in our cellar estimations.

Normality Tests

[Autogenerated] and I really come on. An important goodness of the test is goodness off it to a normal distribution. And why? Because normality is an assumption in almost all tests as a normally detest, it's a goodness of fit test. We'll really know the hypotheses then your hypothesis would be the parable distributes. Normally on alternative hypothesis would be the parable dust industry. Normally. Therefore, what we may want is to have a high p value, not reject the new hypothesis. There are Freeman ways that complement each other very fine normality. Parametric tests none. Parliamentary tests on graphical verifications on the perimeter. Excited. There are too many ways she square, as we've seen on the force. Both need to first estimate the mean and variance off the normal distribution to test against. Therefore, they're called Parametric because they need to estimate a parameter inches square, one beans the data and then calculates the expected frequencies. It that being to use the same test we used before. However, it was a studied that type to error of this test is really high. That means the chance of saying something is normal when it is not on your hand. Lily Force that's the same but compares the community frequencies instead of separate. However, it has the same issue. Both have high type to ever. On the non Parametric side, the most common is shopping a week and actually is the most powerful of all. That means it yields the lowest type to error. It's formula. It's really complicated, but we don't care as we have the software and we will use it on the next models for chicken normality every time, however, should be a week has a flow when the sample size is really beauty it it's really susceptible to normal values and it easily will reject and all hypothesis they come and

accepted weeks do. This is checking the graphic on methods they killed You plot, which is the blood on the left. It is a plot where one compares the actual community frequencies with expected ones in the normal distribution. As one is closer to the line, the points are close out of the eternal values are closer to a normal distribution, and what happens before that is not normal. Usually our tests are quite resistance. It's lightning normality, people. You're close to syrup. It's verifiable. The smaller However, if that is not the case, we need to transform all the data. But that is a topic for another time because it opens the sea of possibilities.

Summary

[Autogenerated] one more on another model is gone. This one was short, but not less important in this model, when allies goodness of feet, which is crucial to evaluate assumptions in almost any desk, we checked the general case against some expected priorities for or categories in the sacre. For that purpose, we learned about a whole new name distribution. The Cheese Square using that we help ourselves US HR managers to bury five romantics was fair against the market. Finally, in the normal case, we distinguish different ways of checking normality. We will demo of those ones in the next morning when we will need to verify the submissions for an over. And speaking of which we finally got to a point where we will understand how to compare mutable continues parables. One of the hot topics of the course analysis off parents stay tuned because what comes this war of hearing me

Analyzing Experiments with ANOVA

Introducing ANOVA

[Autogenerated] on. We got here to the big moment. The one we were waiting for. A Noah. I must say I'm excited because I know what it's both the end of a great journey starting a statistical testing but also the start of something great. When we start making our model more, more, more complex. I envy of you that this is your first time on. At the same time, I'm honored to present this topic to you. So without some many mambo jumbo, let's trillion.

Introducing Demo: Selecting the Best Coffee

[Autogenerated] our company, where bring coffee labs customers and wants to help them figure out what is the best player in the coffee for them. In that spirit, we will the machinery application that after some questions, it's as just a coffee baron to be your favorite. Our state's BP is confident that this lunch would be extraordinary to the next year sales. So we'd better have the best algorithm on board. Our engineering team developed free different algorithms to do the task one involving a logistic regression one involving an S area unit on the final one involving a CNN. And it depends on us to decide which one is better be a complex A B test in production. I learned that line with sample in 15 different countries that free algorithms randomly to get the average sentiment. The suggestion in Bo's, where each sentiment was in the scale of 0 to 5. How can we compare the sentiment for each algorithm on decide which one we will choose

Null Hypothesis of ANOVA

[Autogenerated] Unova or a nice appearance, he said. Technique that will allow us to answer these questions. It is a test where it will be able to compare two or more means off the same, Maribel explained. By one or more explanatory variables, it's used when there is one quantitative variable that can be explained by one or more quality parables. In our example, the quantitative variable is the average sentiment per country they call it. The barrel that explains, is the type of algorithm CNN s are you or a lower where l. R will stand for Jesus Regression, it's are you will stand for the S R. U unit on CNN will stand for CNN. Our sample size per category is 15. They consider the burial is called Dependent Barrel. On the qualitative ones are called explanatory variables. Each different experience. Very variable are called factors on each different value. Off the explainer bearable is called level. For example, in the case before we've had one factor, I'm free levels. If also we differentiated between male and female, we will have two factors with two on three levels each. The algorithm on the gender over these course, we will cover one way an hour or another off one factor. So what does a number do? It is a way of splitting the variability off the direct into two types between treatments. That means the variability that comes from different algorithms and inside the treatments, which is the morality inside each particular algorithm. So by doing that for the variability between the treatments is high enough respect to the virility inside the treatment. Let's not worry right now how we check that. Then we can say that there is a statistical significance off some treatment. What? Let's look at the picture on the left. There could be some difference between the treatments, however. There is a lot of operability within each treatment itself, so the plot becomes Larry, and it's not easy to distinguish if they are actually different. Compare house par cities inside its treatment compared to between the treatment. This is a typical case where I know I would say there is no difference. On the other

hand, on the right, each treatment is quite uniform on the data points don't differ that match. However, each Friedman is really different from the other. This is the case where we say that the variability between it's much larger than inside andan over will say there is different in some treatment. We know these. Maybe it's clear what the North hypothesis for another will be right. If we imitate music I to be the mean off the ice treatment, then the new hypothesis needs to be that all the means are equal music, one with equal music, too, with equal music. L were I stands for the number of levels on the alternative hypothesis them would be some mean is different on Lee that we cannot say more than that for now.

The Output of ANOVA

[Autogenerated] and now that we cover a hypothesis, we can go to our estimator. In this case, estimator will be a lot funkier, but the guidelines are the same as we've been telling. Another tries to compare the variability between treatments from our ability in sight treatments In our world, operability equals variance. So what we have to do is actually compare variances, remembering more than if we were talking about comparing variances and we let that hang for a bit. Let's recover that point. I continue further analyzing a new name distribution F. Fisher Snedecor. The F distribution comes from the division off to cheese squares as we want to compare variances. What better way of comparing them than debating the greater by the smaller. If the variances are equal, then they're accordion. Must be one, as Paris is behave like to square their quota and master street like an F. The effort solution is the one on the picture. We can see a lot of things that it is positive, since parents is are always positive. It is parliament raised by two perimeters, the degrees of freedom of the numerator and denominator and in this case, the perimeters air on s in the picture at the degrees of freedom of the numerator and denominator, respectively. I'm always trust me. We can calculate probabilities off them with Jamal. We therefore, how do we compare variances? Well, they're not. Hypothesis need to be that they are equal that can be translated into our estimator. We were talking about dividing them under the known hypothesis This estimator, this one Therefore we could play the P value. You send the degrees of freedom of the numerator and denominator off getting the value we got against one. For that. We used the f extrusion we compare against the significance and we know how to compare barons is now and that they are put off family for another he now. But we will see on F score, which is the result of the binding the barriers between the treatments by the various inside the treatments. They're degrees of freedom between the treatments on the degrees of freedom inside the treatments. And finally, the p value we care about which now we know how to interpret is the P value getting their score. We got those two degrees of freedom by the distribution f off Fisher's medical.

Therefore, supposing the significance is Sierra 0.5 If that value is less than 0.5 That means that the 95% confidence that variability between the treatments is significantly more than the burrowing inside the treatments. Because our analysis in the previous video that translates that there exists some treatment that is different, although we don't know which one yet. What I really want you to take from here is why there is an f Can I put off a nova? What does it mean that the value and finally, how that translates to difference in means?

Demo: Selecting the Best Coffee

[Autogenerated] And just like that, we are right to our demo. Let's see all this in action. Take a peep on what all these are Noah is really about. We will make this critic plots of sentimental algorithm to analyze variability. We will calculate a one way and over and try to interpret the results. Let's go on. We're backing Jamadi here. Already loaded the data set for these new demo. We have the sentiment for each country for each Calvary. Remember each that happen here was the our sentiment in a scale of 0 to 5. Oh, Verity. Dr. Bones in each off the 15 countries. Therefore, these 4.5 to 6 is actually the ever sentiment in one country for the algorithm, Jesse or aggression. We didn't mind. Let's go on, understand what is going on with our parables. So we go to exploration on descriptive sous usual. We'll try to remove a lot off the things that we don't care about, like me seeing an end and minimum maximum entree to put minimum arians which we always take care. But now we will look at the plots, a really useful plot that we will look. It's the data plot they let a plug will allow us to understand how they vary. Well, actually, this purse on that it's really useful. Now we hide these two forms on we select the power barrel is sentiment, and it's pleated by algorithm on that seat. So if we hide the form, if we look at the descriptive, we can see that actually, CNN appears to have a better sentiment that luscious Russian and I serve you. So we hope that another will try to find that holds on the bar. Inside it appears the CNN as much less variance that the other two, which are kind of similar on that, translates to the plot. It would check under the other blood. We can see that actually logistic regression, and it's for you. I don't have a really clear sentiment here. This purse between serum five, however, CNN appears to be almost everything about free. So we hope that I know I will try to find that in one unique number to do the Noah test, we go and select a nova. I'm one way in over on one way and over. First, we will assume equal variances because we're doing Fisher's another. Oh, so we won't take their submission checks for now, and we'll we'll ask for their descriptive plots. Of course, part of those were already them before, but they don't hurt then, on sentiment with elect beats, the dependent variable on groupie Maribel is algorithm on. That's it. Now, on the one way and Nova result, we can see that the F score for these and over waas six point well, the degrees of freedom

between the treatments waas, too. On inside, the treatments were 42. The value of this test is syrup 0.5 as this is below the significance. That means that I know I found that at least one treatment was different off course from this critic blood that we made before. It makes sense that will be CNN, but we don't know it yet. There is critic plots on the group's also tell us the same thing with you. Before that, CNN had a mean off four below. We also have the box floats. A brook flowed. It's a nice representation off the mean on the comfy and seen the root for that mean where we see again that it's a real electricity progression seemed to be kind of the same on CNN appears to have a better sentiment. With all these, we finally analyzed a one way an hour all the way. So let's go and try to understand water we assume in. And how do we validate US assumptions so we can trust that that syrup on Syria 05 Israel.

Assumptions of the Model: Hetero-what?!

[Autogenerated] and now that we know the test, we go to the assumptions. The assumptions of another are quite simple, but sometimes tricky to validate. Samples should be random and independent. It's treatment should distribute normally on the treatments should be almost a drastic What? Don't worry, we will revisit the last 21 at the time, as we can take for granted. When we have our data set that assembles were correctly, they can become a forward to normality now. The complicated part is that each treatment should be normal. But usually we don't have many at the points per treatment. So what we will do is first, let's define the residue. That's the difference between a given data point asked on the mean off that treatment, then this is where math is our ally because there's a theorem that says that normality for every level, is a Greenland to normality off their residues off a whole. And now we're back on track because we already know for the test normality off one population Shapira week and kick it plus. And that is what drama with us when calculating and over Jamal, we will ask us if we want to include the Shapiro with test on the tickle plot against the residues. So we're good, because if the P value off the Shapiro Week test is over, the significance on the scatter plot off the gift applause is close to the original. We can safely assume our treatments are normally distributed. And what is the other world? Almost elasticity. It's an equivalent toe acumen is off. Parents is which actually mean the guarantees off all treatments should be equal. Remember the detest and this assumption is super super important. I understand why. Let's check this picture. If the guarantees of the treatments are not equal, then we can have an undescended relationship between variance on exploratory variable. Also, that means that in the calculation of Paris is inside treatments, there is more pool on the treatment with different variances. In this case, CNN without more pool on the inside treatment durability. That logistic regression, therefore we cannot infer from difference in Paris is

that we have a difference in means and that was the whole point. Therefore, I know what I cannot say anything if we don't have almost elasticity

Checking Homoscedasticity: Levene and ANOVA Inception

[Autogenerated] and how do we check on those elasticity Liking normality? There are two ways that complement each other graphical methods on leave and test the gravity of method is to calculate the recipes for each value on then, to applaud off. The rest of you is trying to say uniformity of distribution against hero like in the picture, a common example of the taking their side elasticity, Which means not almost elasticity by this method is the graph on the left, where we can see that L. R. Is really, really, really known. Dispersed on CNN, it is a classic example. Off almost elasticity is the picture on the right, where they're really similar on the expression against Hero. On the other side, we have 11 Test, which cast a really nice street to the Tech Thomas Elasticity. The main idea come from the pictures on the left. We can check a plot off the recipes for each level. It is easy to see that the second group has a higher barons at the 1st 1 so we should be able to detect that right on the right. What we do is send every negative body of the rest of you two. It's positive, Counterpart, and check what happened before everything. Terry had a means center close to zero. However, after the movement, the mean changed even more if the variance was bigger, the mean of the new barrel speaker on dhe. That's the cool trick. Because now we know how to compare means and different categories on Nova. And that is why I like to call it. Then. I know inception. Because if I never find differences with the means of the new parable, we can infer that is because off it'll okay, elasticity incredible, right. But what about all the sanctions off this new and over? We are eating our tastes, right, Nat? Because math also tell us that normality comes from Shapiro weeks from before on this new parable it there. Okay, elasticity in 11 off the new rest of you is the same as difference. Off means magic. So to give an example and I put a little Aaron looks like the one in the left. It's a new nova. So after no hypothesis of a nova is that it means are equal. That translates to all parents. This is Ali are equal on the alternative hypotheses would be some variance is different. Therefore, if the P value is over the significance, we are good to go. We look for a high P valley hoof. Such a mass, right? Let's do a quick recap before them. A essentials off a nova are normality and almost elasticity for normality. We shop for two weeks and dig a plus for the rest of use to verify the normality. That translates to normality of the treatments For almost estes iti, we support a phrase used to verify on then live in, which is a new and nova on a new parable of the rescuers to take. Thomas Elasticity involved statistical tests that no hypothesis is actually what we want, so we look for a high P value.

Demo: Revisiting Selecting the Best Coffee Video

[Autogenerated] Now let's revisit the same example. But now we will check for normality, which appear weeks and kick it plots. We'll check for almost Estes City. With Lauren on, we will calculate their CDs to make the blood of the restive use. Um, we're back again to the same data set in this data. Set off the sentiment for algorithm. We will now analyze if the assumptions, often over we're actually taken care off. So first in first we go and we try to run again the one way and over. However, now we will take a look at the assumption checks again. We go when we assume equal. We can ask for this with the plots, but we will ask for the free assumption checks. Quiet if parents these chemical plant on Shapiro weeks we selected the dependent variable is sentiment on the explanatory barrel is algorithm. So then again, the F score is six point Well, we already knew that. However, we can see that actually, the Shapiro Week test gave a high P value syrup 0.45. That means that the population off race views actually, we cannot say that it doesn't this route normally on because off our fear, um that translate that we cannot reject the hypothesis that each treatment behaves normally. So we're good to go. Then we consume live in off course. Living is another know us we've seen and we were really close. So it depends here on the significance. As usual, we're taking 5% significance. Therefore, the P value 0.51 is overseer points here are five. It's quite close, but we can assume therefore, that we have equality of currencies and we're good to go. If we check the plots, then we go to a cute A plot on a Shapiro week said everything was okay. We have to verify because we know Shapiro weeks can fail in the case where some precise is big enough here. 15 it's never be on the give Blood says that all the other bones align with Lionel. So we're good to go. Assumptions off normality and almost elasticity appear to be evaluated. So the final thing that we need to go on do is to do the rest of your blood ____ off the versions here 0.94 point six for nine. Family doesn't include per se the recipe plot. So I will show you how to calculate the rest of us so we can actually create a plot as any blood. So to calculate the rest of us, we go back again and we have to create a new computed variables. So we go to data on we goto. Other on this have would be to upend a new computer parable we select on Dad and with double click on C. See, now you can see it's a computed parable. That means that we can put a formula to select which its its value. We will name it racy views on what would be the value. Let's remember the receipts were each data point against the mean of the treatment. So if we go here, we have our barrel sentiments that we would be sentiment my nose, them enough treatment. So here we have a function called them in which me, which is the mean on be goodbye So we don't leak thes. And it says that the first Maribel needs to be there very well. We want to group and the second is the experience very variable. So we put that we want the mean off sentiment on We want to group by algorithm. We click center on. We're good to go, Jamal B.

Outer Calculating all the residues off course. This line and this formula will be on the provided material in, so you don't have to actually memorize it. We hide it, and now what we can do is to create the final barrel, the one that 11 goes and makes the test because it's really good toe. See Levin working. Therefore, we are a final computed parable without a click on this completed valuable Remember it Waas. If the recipe was positive, it's the same. If it's negative, it goes to positive. Therefore the function is us, so we will look for it. And here it is and its ads off residues and we will name it ABS recipes. So we have all the variables we want. We can check that 1.878 went to the same value, but minus 1.769 went to the positive one. So we're good to go. Let's make their plots to understand what was going on. We go back to analyses, exploration, the script IDs, and now the barrows. We take care off our residues and ABS recipes with split also by algorithm. But we don't want the statistics, although it may be useful to see that actually mean. We can see that actually, the mean off each of their issues is really zero. That is because off the absolute off the machine that he doesn't say Syria, but it's zero. But here is the important thing. We can check that actually asked the barren seas off logistic regression. And Esther, you were bigger than CNN in the ___. Recipe is variable. The mean off CNN is lower than the other two. See, this is exactly what we were talking about. So if we go to plots and we ask for it, that a plot, we will see exactly that on the residues. We see that everything is center at zero. As we've seen, CNN appears to be with a smaller barons. Although we were good to go because Levin said that it was a close call. But there was almost Estes City on on the abs. Rest of you we see that actually the mean on logistic regression test a real now are is bigger. That mean on the after I see you on CNN. So with all these were evaluated assumptions off an hour. We understood howto calculate the rest of us. In case that we need to understand what is going on. A ski jackboot. We calculated the ___ residues to understand. What did Levin do? So we see with our own eyes what I've told you that it difference. Embarrasses translates on is equivalent to a difference. It means in the abs rest of you. Awesome, Right? Let's do a little bro Pop off all the concepts we've learned over thesis model.

Summary

[Autogenerated] That was a great model. I enjoyed a lot telling you about an hour because I think it's one of the things that are most needed analysis made by companies today to prop up. We discover what a number that's to compare me and seeing a single test. We understood why it works, and from there we check how it looks like. From there we stayed and checked how to body eight assumptions on why they are like that. In the meantime, we learned a lot about almost elasticity, which is a crucial concept in statistic. In general. I know what you think. This cannot get any better, right? But he can, because until now we only know there exists differences. But it

would be awesome to know which ones on DDE. That is what contrasts bring us. Let's discover them.

Comparing Groups and Effects with ANOVA

Introducing Demo: Mixing All Together

[Autogenerated] after this point, we were able to discover if there is any difference in the means of one. Dependent continues parable by one explanation. Every variable, however, we want more. We want to know exactly how they behave. Let's remember a little bit the previous model example. So we have free algorithms we need to choose from Georgie. See, Regression is aerial unit and combination and narrow network This waas because these algorithms suggested coffee baron Toby, our favorite, and it was crucial in the sails off next year along that line with sample 15 different countries, the free algorithm friend only to get the ever sentiment were discovered in the last month. Of that there were differences on dhe from the plot. It may appear that CNN was better, but how can we know these?

Why Don't You Use Multiple T Tests?

[Autogenerated] from less models. We know how to compare to means, right? We could do it to test for each pair of means and were Then we could compare CNN's versus El Our CNN purses are you And it's our uber sailor on with these retests were then we didn't need another right. Mmm mmm mmm. Not so much because there is a cover it. Remember that each detest has associative an error. That means the probability of making an error of type one, which let me find now at Sierra 0.5 Therefore, we have pre experiments, each one with its own possible error of type one off syrup in 05 What's the role of making any error? That one? In the whole decision, all the possibilities are on a tight in the detest found differences with the D on that it didn't with an M these ones and Andy and and and and so on. And so far, there are eight different possibilities. Took Calculate the probability off one of them. For example, TDN We can suppose each test was independent. Therefore, liberality off ddn will be the property of the poorest one getting a d times the property of the state on one getting a D times the protein for the 1st 1 Getting an N and this one's off course are Sierra 0.50 point 05 on syrup I 95. Therefore, the property for DDM is

Sierra 0.97 If the PM wants, then the reality the product, if at least one error of one is one minus, the priority is getting it right. That is one my the Sirah points here. 976 to 5. That is almost over Sierra 50.9. Incredible, right? So this is a really, really problem on Noah's contrasts provide a way of doing these comparisons but controlling the global type one error to the significance. There are two intense off contrasts. Plant or a prairie. The hypothesis are set beforehand on before the data, said Israeli. They can detect difference when I know what cannot, but they are limited on known plant or a possibility they run on Lee. If I know what war significant, the year for exploration knows where we want to understand further theirselves stuff on. Over. On the most common one, it's too okay. First place analyzed Turkey since this one, very close to what we tried to do before with the tests

Tukey Comparisons: I Want It All

[Autogenerated] took a comparisons are great for your understanding. With unlimited. The main goal is to compare everything one against the other. Why controlling the type one error? So in our example, we have free levels we will have free. Took it tests CNN against all our CNN against this are you and l are against us, are you? It's no hypotheses has to be that there is nothing out there. So music. I has to be able to music J where I m j r iterating, our all the different levels. It's statistic them. It's the difference of means. And we will have medical hypothesis. As tests on we would have as many tests as possible. One verses, one comparisons. The difference of Tokay against multiple detests is that idea to find one value that is called minimum significant difference for all tests that value is calculated. Think it account everything on white that changes everything because then the tests are not independent, as we stated with the tests. For example, if the difference of means between L. A and CNN, it's 0.2 on between l A and S. R. U s European Sierra one on the minimal significant difference would be Sierra 10.1. Then, in that case, today would say that l R sequel Trust or you. But dinner is different from CNN with 5% significance injure Malloy on output of Tokay. Looks like this where we will have all the comparisons. It's me indifference and then all the calculations to get a P values for every comparison. In this case, we can safely say that CNN is significantly different from El are honest. Are you? However, L R is not sick statistically different from s Are you all with 95% confidence? However, life is not simple sometimes and we may have something like this where we have that CNN is not different from us. Are you and s Are you? It's not different from a lower our luscious them would say If a equal speak on b equals C, then egg will see right from city Vitti. However, statistics is tricky on these things don't happen. Equality in significance. It's not transitive in our case. Check It happened CNN east different from a lower. So always beware off assuming without testing

Planned Contrasts: When You Need More Potency

[Autogenerated] But sometimes we're just on the other side of the road. We were sure there was going to be some difference based on brilliant studies. And I know I said nothing. Are we lost off? Course not. This is where plan contrasts coming to play. These have to be stating even before and over there a way of specify in what we suspect of our data. But here is a little trick. There are a few of them. We cannot go row. Usually they come in two flavors one versus one, like in two. Okay, where we cannot have them all. So it depends on us. Which tohave or one person? Some group of levels on these ones are usually to know where one category is better than a group. This is where demonic comes handy because it offers a good operative possible brown contrast without having to think match. We will review all the option in there must for now. Let's check an example. So, a contrast in Jamari looks like this where it tell us that CNN is significantly different. That both logistic regression honest Are you by separate? However, we don't know anything regarding this. Are you and Ella. Also, we don't know if maybe a combination of woe is even the best on. Why would I want is Let's check this in this example Plan Contras will save us. In this case, both a nova on tookie returned non significant differences. Check the p value. Their numbers are not important. However, find two contrasts. We see that actually CNN all those not better than any of the other alone. It is the better than the other two as a whole. This is not the most common example. But it may happen, so we must beware. Plan, contrast. Save us when we think we're lost.

Demo: Mixing all Together

[Autogenerated] And now we are ready to revisit last month as example to take all the differences and actually finally decide what our company must do. We will make clam conference on the algorithms and discuss the different ways of doing so and make took comparisons to get a good idea of what is going on. With all that, we will make a decision for our company. Wild brain coffee on one's more We got to the same that set. But now we will go all the way. On the first time we're visited, we went, did a Nova and found there were some differences. Then we went once that more validated assumptions. So now we want to understand more about what differences we can actually go and smell. Let's go on, do it for that we clicking and over. But now we go to the full and Norma, where we will be able to do a lot more fees that will cover it. Actually, So before this dependent variable is sentiment on the fixed factors is algorithm. So here we have our I know what it says that there are significant differences. This is every Venus we knew. However, now we have two different rose. We don't have to take care about that right now. What what that means is algorithm tries to stress, to say what is expressed by the more what is

the variability that the mother explains. We will take over that over the next modules. So now what we have, we can go and go. The assumption checks. Although there is nothing new here, we can go and check that give you plus the same on Live in gives the same result, and we go to contrasts. So we have an idea. Our idea was that actually CNN was better than the other two. So let's go on holiday that if we go to contrast, these are planned contrasts. There are different types. Let's review some what each one of those go to the next one. East Aviation Division means that it will take the 1st 1 against the group off All so basically it means doing actually think one is better than all those working together. We're not so interesting in that one, but it's important to take into account. They're one, which is the wonder. It's important. It's symbols, simple actually goes on and takes the first against the second on, then the first against the further, in this case, logistic regression against the other two. When didn't have. Actually, that's a position about the position regression. We actually take care about CNN. One thing we could do actually within Jamal Reese to just change the order off the levels. Both CNN, as the 1st 1 on Simple would do what we want, which is to validate CNN against the other two. Finally, let's cover the last one that it's mostly used, which is repeated basically re pretty takes the first against a second and second against effort and so on and so forth. It's important to understand that in this case, we don't want to actually do 200 of us and take one plan, contrast on one and sampling countries on the other and then mix together the actual results because that would be a fallacy. It would be the same thing as the T tests. We cannot go fish for a P value. We have to commit toe one plant contrasts on that land Contra. I need to make sense for us. This is really important. So in our case, we will stick to repeat it because it's the one that we care about from the repeated. Even if a Nova did not found any differences, we now know that CNN is actually different from LR on from this area by separate. So those are the plan contrasts. Finally, we can go to post out tests. Possible test is another word for posterity. Contrasts. So we go on, we select that we want to do a posterior on all the levels off algorithm, which is our only factor. They're in a bunch off post op tests on We can later talk about them. But for now, let's to took a which is the one that we saw we go down on. This is the result of the post op tests. It goes and it just the comparison of logistic regression again, CNN mattresses, regression against us. Are you a 10 CNN against this? Are you? Let's always remember what this is doing is finding a minimal significant difference for all the comparisons. So it controls the global type one error from here, we can tell that actually CNN is different from Misterio. I'm from logistic progression by separate also from Turkey on Took you found that Also, logistic regression is significantly the same asked s are you? So now we have a clear, clear idea of what is going on. If we need to actually go and select one algorithm, our best shot would be CNN on. That would be a recommendation for wire Brian Coffey. Finally, death recommendation is also backed up because we check the assumptions. We didn't know what we

did. The blank on dressed and everything. Saysthe. Same. We went and mix altogether. What we learned these two mothers on this is cow One analyzes multiple comparisons when we have one or maybe more fix factors in the case. We have more factors. Then there are multiple ways up. And I said that it's really interesting to dig into Andi. I leave it for another course. Thanks so much. And now let's do a wrap up off a nova, which was a great too on one of the hopped up except the course

Summary

[Autogenerated] Andi. This is the end for journey with a Nova and we learned a lot. Another is a great it nick and can be extended in several ways. We can start adding more fix factors and have more complex science. Or also we can start out in time is a variable or even we can combine everything on Also, some of the categories may be random, but going back let's prop up for now and over, he said. Technique to compare means off multiple levels but trying to split the very relative that I said into two categories between treatments on his side treatments. With that, it's able to detect if there are some differences in the levels. It assumes that all the treatments are normal and their parents are the same. This is very important and we have graphical and an analytical ways of checking it to continue our analysis a bit further. Then we started learning about contrasts that are an efficient way of finding insightful comparisons were controlling the global type one error. There are two types of those are pretty an apostate bot base if they come from before the experiment or as a result of Chicken and Oba the most common a burst of the earlier contrast is Tok, which is based on calculating a single MST toe. Every comparison. There are times that took a finds, a group that is equal to the others, even if those groups are themselves. Diferent, however, has to get out and find differences Where are never fail. We're relying on plan contrasts toe. Get some valuable information in those cases. These are a few on having different structures that try to compare the groups against the base land or aggravations of categories. They have the nice features that even if I know what fails, they can bring some insight. That was a big set of values, but we got to the point where we understand everything about comparisons of categories. Let's end this course with an also important concept. How to detect relationships off parables. Let's move on

Predicting Linear Relationships with

Regression

Introducing Demo: The House's Price

[Autogenerated] we are almost at the end. We don't know how to attack the problem off a continuous parable being explained my quality barrel, which we call explanatory variable. But what happens if both barrels are contented here? In that case, we can no more. We can actually try to understand the function of the relationship between the variables. How one explains the other on dhe that can be a powerful torch on bread. It new and seen values that is regression. And it's the core off a lot of machine learning nowadays, as always, First, let's contextualize the problem with our case study were engineer, said the consultant films that were contacted by a real state agency. They told that they're worried about their price estimations. I want to make sure that decision is that other woman in that spirit, we were assigned to understand, if possible, how to predict the price of house price on its square meters. Because the agency is convinced that it's one major explication of house prices nowadays for the purpose. The agency gave us a set of house prices with it's a square meters

The Power of Regression

[Autogenerated] all these is contextualized within statistical Mother that has this form independent. Bearable is a function off one or more experience, very burials plus an error. There are two main forms off the statistical Mother in the case of Splendor of Arable, our qualitative. Then we have an over. But if experience their barrel is quantitative, then we have regression. Regression is a technique to understand how it changed in the explanatory Barry will affects the dependent parable. For example, how changes square meters off the house affects its price. A regression. They have one or more explanatory variables. For example, if we want to estimate price by square meters on number of bathrooms, it can be then simple. In the case it has warm. Maribel are multiple if it has more than wearables like the case with bathroom. Finally, one drink. Classify regression based on the model it come in linear. In the case, we say that there's a leaner relationship between eye on X. In the case, the moral looks like I is a constant blast. Be another constant Times X plus an error. We're eight on be our perimeters or nonlinear when this relationship is not linear for example, why is a constant plus Ito there be time, ___, pleasant error on this model, we will attack simple linear regression. An example of this problem could be the following picture, where it appears that the relationship between square meter some price is linear on our problem to solve would be who would be the estimated price for a house off 212

square meters if we check that value? Wasn't Ben shirt. It's a prediction for attacking that problem as always. First, we need a new hypothesis on an estimator and in this one is an interesting point because we're tend to say that the North hypothesis is that the model is Lena. But what would be the alternative hypothesis? In that case, this is not what we're trying to prove that is, our more. But what we tried to check is the linear relationship between I and eggs in math. The linear relationship will be detected by beta. Therefore, the correctional help offices is that better equals zero. That means there is no linear relationship for lower could be an hour then. The alternative hypothesis is that Peter isn't zero because then in this case. If we reject, we can say that there's a leaner competent. It may not be the only one that's really crucial. And if we don't reject, we can say that after the mother with stated the best estimate of four Price would tell me it's average, which is what we've been saying after now with the confidence intervals. So we're good in that spirit than our estimator needs. Toby be estimator for Beata the slope. So all we need right now is how we calculate those estimators on how we can make a test to check their Selina relationship.

Minimum Squares Estimation: Find the Linear Part

[Autogenerated] estimators off the question of the line A M B. R found Be a minimum squares estimation. Let's see how that line is found from the picture. What the software thus is changed am be to find one that actually minimizes the green lines. This is called minimum squares estimation, because those green lines are squares off the distance between the point on the actual line, which will be its prediction on how we test if there is an actual relationship, there is a way of splitting the variability of each burnt into explained by the model. This is the variability between the predicted value by our line on the general mean remember, the general Means is our best Muller under the new hypotheses on DDE not explained by tomorrow the variability of the actual point to the predicted one. Therefore, with this, we can do a number on example output off the movie for linear regression, then looked like this where a nova is for better. So in this case, as the P value is significantly, that means that the model explains a good portion of the variability. Therefore, there is at least a significant linear relationship between House prices and square meters. However, does that not mean the relationship is poorly linear? We are only testing the linear component. For example, let's check this picture. In this case, the relationship is more quadratic. However, if we ran in another over the linear model, we will find significance. Why? Because the linear component exists, although also there may exist some quadratic component. Finally, we come on our component in the number they are square. What is that is called a square or coefficient of determination, and it's widely using statistics and non statistic worlds. What is it?

It is the proportion of total wearability explained by the mother. Um, it is denoted by an R Square, therefore, is a battle between zero and one on. Being closer to one means that the model explains more part off of our ability in the sense off the picture that when we explained that over, so what he does and it doesn't mean they are square. What is it? It is a non statistical measure off how good the model feeds. It measures the proportion of variability explained by the model on. It's a measure of how predictive our model can be in the range of the planetary variable. However, it is not a statistical way for in the moralist linear, it is not conclusive to say there are not more possible explanatory variables on being zero does not mean there is no relationship between the parables. Let's look at some examples the following pictures. We will have opposing examples off good linear and nonlinear relationships. On this one. There is a clear, no linear relationship between the barrels. Therefore, we can expect on our square close to Syria on this one. There appears to be a linear competent, although it may not be the best model we can expect on our square close to open five or kind in the middle of serum one. Finally, this last one is more linear on a linear model appears to explain a lot off the bar ability. Therefore, Esquire will be close to one. It is important to understand that underscore close to one doesn't mean the moral is exactly Nina

Prediction and Extrapolation

[Autogenerated] once we have our model on, there were significant differences. We can use that to actually predict the house prices in new values. In statistics. To predict about you is to use the model that was evaluated within the range of the use of planetary variable. When that does not happen, we call it extrapolation. As an example, suppose the mile for house prices wants that the house price is 2000 and six times the square meters. My No. 6689. Then if we want to know the house price for 212 square meters, that would be a prediction on its value actually is 2000 and six times to country and 12 minor, 6689 which is close to $420,000. But if we want to know the house price for 415 square meters, which I will tell you it was outside the range of square meters that were on the moral, it would be an extrapolation because in our case, the biggest parley with measure waas 375. Therefore the value here would be close toe. $820,000. But this is an extrapolation on Dwight difference in name because for predictions which have a confidence in their all for the slope, the intercept on the scientific value. So actually cantina regression and stop looking like this. We're in gray. We see the confidence. Interrupt the prediction on dhe as we move further from the center, the uncertainty speaker. There are formulas to calculate those intervals, but we don't really care about them right now because there is always suffer. The software will give us the confidence in there. Well, we won't.

The Assumptions of Regression

[Autogenerated] on. That's always there is no statistic beauty from me without its assumptions because they are critical for a moral to be valid. Assumptions off union regression are the values off X need to be measured without error. That means that they were pressed on the ship, the values off, why I need to be independent. This is always normality of receivers with means hero on almost that's t c D off the rights of use. It is no surprise that they like an Oba since what we're actually doing is an anathema as always for normality. We have sharp reawaken analytical test on the kill you plots where we want the people to be over the significance and the rest of us align with the diagonal in the killer's blood. However, there was an issue with Thomas elasticity in this case, because we may not have enough samples per volume X to have an actual Innova, we may have only one value off. Why provide of eggs? Therefore, we have to analyze the rescue deployed qualitatively on the following example. We have four different things we can see in a picture. A. We can safely assume there is a most elasticity sees the residues district randomly in E. We can say that there is a most domesticity because for each value of X, the various east similar. However, we can also say that the linear model does not explain everything because we can see a pattern in the rest of us in B. On the hand is a clear there Asiatic case where there is more variance as we have bigger values affects in GI. Finally, we can safely assume there is a most elasticity, but we can also see their son abnormal value. We call those out liars and they're really important when we move for the new statistical models on why we care about those are normal values. Let's check the following examples on the first row. The persons affected I'm normal value made them all much worse than what it actually waas. We verify that because removing their value their square in Brooke significantly Therefore abnormal values can make our model look worse. And it's not always that easy s here toe Find out there was an abnormal value on the other hand, in the second row on abnormal value, may the mother look a lot better, so it goes both ways. Actually, there are no clear linear relationships on that. Sierra 0.67 in their square was flying. It is really important to look after of liars because they can brings from possibilities of common interests elasticity. They can also alter the normally assumptions, and finally they influence the mothers to not be what the true that is.

Demo: The House's Price

[Autogenerated] and we got to the demo. However, this model was full of examples to understand what was going on. Now we are ready to help our real estate agency and make a linear model that will predict with statistical significance, the house prices in particular. We will check the scatter plot on a license. Linear model is the correct one. We will run the linear

regression to verify the significance of the model and finally analyze all assumptions undiagnosed metrics of our model. Let's stack elite on we're back again in Giamatti Hear already loaded That set for these module where we have what the real estate agency sent us, the square meters on the price, they said for different houses. So let's go and tackle this data. The first thing we ought to do it to get a plot to understand what is the functional relationship between these two barrels? We expect that square meters explain some off price. Therefore, when we go to analyses, we want to do a scatter plot. However, that doesn't come right out of the box. Which a mommy. So what we need to do is go toe class models Tamami library and here If we go down, we will find that there's a library that's called Scatter on. Actually help us to do these scatter plots that are really useful in regression already installed mine. But if not, you can click on soul. Let's hide this form and then goto exploration on. We have already open the possibility scatter plot and this other bloke us. We want to explain price by square meters. Let's go and select price as the Y axis has square meters as the X axis. And here we have or scatter plot. Right now, we won't care about the regression line and merging us because those are our things that right now it's not part of the problem, so we can see that either appear self, although from slide striations we have a really good Lena relationship between square meters on price. So that's a good way off saying OK, that's right, linear regression. For that, let's click on regression on linear regression for linear regression. We will have they sing type of form that we have for another, because then again, remember, linear regression is just a way of estimating the perimeters and then doing in a number they're dependent variable here will be a price. We woke up factors. What we have is a co bury it. That means that it's unbearable that tries to go with the dependent variable and we want to explain by it. That means this is the barrel that actually goes with the slope to be. When we did that, we already have our I know what for the linear regression. So if we check on the different results we had the model has on Arc Square score off syrup on 99. So it was a really good explanation of the variability by the linear model on. We also have the moral conditions. But we want more than that. We want the validation than over. For that we go to model feet any more feet we can ask for the F test. We don't know what this means for the over and model. Let's not take care on the other off. It measures like a c r B i C, because those are are really good when we want to compare different models. And we may looked at them at different courses with the mother feet. We already have that know what we want. So we know that the F score playing by the model was 1370. The degrees of freedom off the model is one. This will be a constant always on. The P value is under the significance of 5%. Therefore, we know that there's a significant linear component in the relationship between price and square meters off course. As we know, this doesn't mean that this is the only relationship. But we are testing the linear component. And what is that slope? Well, it's actually 2006 on. We know that it's different from Ciro because off the

value and we have, it's a confidence in terrible. It's so far. Forensics class miners 54.2. All right, with all these, we already have our linear model. We know that they house prize them will be 2000 and six times the square meters minus 6689 and we have the confidence interest for all those values. But what about the suction checks? This is really important, as we've always seen. Therefore, let's check it. If we go toe assumption checks, we can do a lot of chicken, but what we care about is the key part of the receivers on their A serial plot. From there. If we go to the Q plot in the receiver plot on the kitchen, plus we can see that the receivers appear to be really close to the diagonal. Therefore, we can say that all the population off receivables is a normal distribution. More specifically, we cannot say it isn't right. And if we check their civil bloods, it looks like other issue Iwas This reeled uniformly and randomly across the zero. Therefore, we can say that we don't have strong Indians to reject the null hypothesis or off almost elasticity off the receivables. Therefore, we are good to go. The X was measured without error, thereby use off. Why are independent? This is an assumption. We're supposing that the real estate agency didn't make an influence on one price against the other. We know that they're received. Waas are actually normal on we know that they received was our most fantastic. Therefore, the linear model we found out it's valid and we have a good clear explanation of the linear competent off the price by square meters

Predicting Non-linear Relationships with Regression

Introducing Demo: The Housing Problem

[Autogenerated] on. We got to the end this quick. Miller will extend the previous swim by analyzing how to tackle some non linear Katie's with what we know. This is super useful in real life, really used, and the most fun part is that now we will know why it works. We're having a chill day as consultants again. When the real estate agency called angry, they say the more than ___. Apparently they tried to use it on super extra large houses, and the mother was returned in really low house prices. Then he's up to us to save the day by providing a new mother that can take that into account.

When the World Is Not Linear, You Just Linearize!

[Autogenerated] suppose we go to the data The real estate agency. Send us now and we get the following plots. This is clearly nonlinear. What can we do? This is a normal case where we have a non linear relationship. However, there are ways to handle these. That process is Lena Realization. You know, the station is the process where we transform a hypothetical, non linear relationship for leaner one. Then we can do a linear regression to validate the original mother. Furthermore, the regional perimeters then are recoverable by transformations. Let's check an example. This appears to be an exponential. Therefore, we can propose the following model. House price is a constant tans. Ito, there be a constant times the square meters. That seems like a good model. Right? Then what can we do? We can take logarithms if we take leverage. Um, then the hovering off the house price. This is a natural longer in the locker room with base e. It will be the following formula here. We use that the luxury off the product is the some off the logarithms on the luxury, um off the exponent is actually goes down and it goes off one times the exponent. Therefore, if we said why to be the hovering off the house price, we end up with a linear relationship. Where were renamed Lowering off the house prices a new arable on delivering off a, which was a constant, is another concert which we call size A. Now let's suppose your mommy return that these final why it's actually three plus two times the square meters. How we record the regional relationships Well, we know that the lumbering off a waas free on B was soon. Therefore, we go back to the original. Relationships can say is e to the free and then as the equal sue house prices each other three times e toe there to times square meters and we got through the original relationship and this is what we will do in Germany and what other relationships were coming in your eyes? Well, we have a power relationship. Then we're good to go because taking logarithms, I'm having a new variable called loge Arena. Why, on another variable called lover in law affects we're good to go that cell in a relationship also, how an inverse relationship were also good to go because we can rename one Over X to be a new barrel exterior on with Have a Linear relationship became Finally, if we can't hyperbolic relationship were also good to go, renaming both miracles as their embassies. So now we can actually use linear regression in a lot more cases where we have nonlinear proposals. But there is a final case. We are not talking. What if my barrel is polynomial? What if there was a term X Square? Well, in the spirit of unionization, we may propose that why equals a Plus B explosive cc where C is equal to X square. However, if we do that, it's leaner. But it's not simple regression because we have two variables, not just one. So we have two possible ways of tracking this. The 1st 1 would be we can go and study multiple regressions and take a decision that isn't normal continuation to this course. But if we don't want the coefficients, but we're interested in knowing that type of relationship, then there is something

else we can do here. If you remember from Germany, there were some plum contrast that were called Paulino Mia, and we did not look a lot into back then. But let's check it right now. If we actually have month measurements provider of eggs and we care about the type of emotional relationship that means the relationship is quite erratic. Andi, I don't care about the coefficients. Then the paranormal contrast all I was to do that, considering each value of X a category for another. The only issue we're supposing is that also, they the values, our experiment or a barrel actually are increasing values. Here what we have, it's a barrel. That explains a cubicle relationship with seeds on yourself. Treating this problem as their aggression, he could. We just took the measurements for 10 2030 and 40 seats separately. Therefore, when we run the plenum, in contrast, we can see that a Nova found that there are significant cubic relationships. I told you know I was great. The only thing we couldn't have was actually coefficients. But if you want that, just go on down on multiple regression on What if there isn't a giving relationship? Let's check the following example. In this example, the relationship is quadratic on check. The contrast for Q week was none. Significantly another rocks when there is security relationship. It finds it, and whether it's not, it finds it. Also, I think that is the main conclusion of this course canola rocks and help us all the way.

Demo: The Housing Problem

[Autogenerated] After all these superliner station knowledge, let's use it to help our real state agency so they can still use Our service is we'll revisit them. Example for Lena rise the new moral and finally to Bali. Eight assumptions to Lena Riz ation and beyond. On. We are finally in Germany for our last demo. In this demo, we will have already loaded that that isn't that the new Riel state Agency sent us where we cover more square meters and we have to analyze what's the model? Why the linear model did not work on which one will work for that. The first thing we will do is actually getting scattered blocks. So we know some war is going on. If we go to exploration, scattered blood on we plod price against square meters, we can see that actually that appears that we don't have a real in a relationship, but we don't know if that is. Maybe because the final battle for 900 square meters isn't a normal value. Or maybe it's that it's more exponential and it's just a start off exponential curve. Let's find out that right also because that will bring us a good Insight on Fielder's You would go to data. We can see that there's about uncle filters. Let's click on that filter is a nice way off in a function to say which data points to include I'm not. Let's remove the last one to see what could be going on a friendly removing that one. We're good with the linear model. And they show was that just our linear model doesn't extend to 900 meters. So we say that the row should not be Row nine and we enter. We can see that actually, a real cross

shadowed their own. Hang on what happened with the scatter plot That's Heidi's and we're kind of in the same. So the issue wasn't best that there was an abnormal value, but it appeared that I'm normal value because exponential justice issue that the final values are actually very, very big, and then they hide the other ones. But it doesn't appear that issue that this is a linear model with one of normal value. So let's remove that filter on, see what we can do to solve this problem. We double click on the filter on. We will remove it. We fight and we are back again to the initial case. So it appears that this is an exponential. If you remember the slides, that means that we need a new barrel called logarithms off the price on global rhythm. The price with travel in your relationship with the square meters. Let's do that before then. Let's click on Compute and we will have a new completed barrel that we will call lover rhythm off price. If we kill the conventions, we already have the natural logarithms on off. What, yes, off price we play can enter on were then we have our new computed variable. Let's let's get her blood to see if this one behaves more correctly. For that, we call back to analyses, exploration, scatter, plot and now we will scatter plot. The longer the price gets square meters and apparently we're in a really good shape because thes appears a lot like a linear function. So the next step, then would be to do the linear regression there, fight assumptions, get the confessions from their recovered original relationship and see what else we can do. So we will click a regression minion. Regression. We go on, we say, like the dependent barrel last longer in price in the square meters, Pascoe, Barrett and we have all in a regression. Now let's have a little more things because we'll really know. We have to ballet the assumptions and see what the mulligan's tell us. So if we go to assumption checks, we will ask for the keys. You part of the receivers on their civil probe for Serena, most elasticity. We tried this form on. We go to Mother Feet. We will ask for the Anoma off the whole model. Finally, we will go to moral coefficients and we will ask for their confidence in general because this is the really cool thing about the movie. It comes all together and we already know everything about the moral. This is no a test that you can check is actually in a nova on the conditions when you have a multiple regulation. So right now it's the same as the overall mother left test, so we will not take it. And in the end, let's get those in convenience in their wives for the predictions. So we will add a square meters to say this is the experience every variable. We want to get the confident in their US about. We'll select that. We want the plots on tables. With that we get everything we can from the linear regression which hide the form and analyse the results. The first result we can see is that actually the explanation off the linear competent. It's really good for the variability off our new that I said. Actually, there are squares one F score. It's really, really high. The P value is really, really small. Therefore, our linear model for the luxury market price is really significant. If we have balance to the model coefficients, we can see that actually intercept was 9.88 with a really small error on the square meters lobe wa syrup or in Syria to fire. That means that the original moral

war is that the house price is equal toe e tow the 9.88 48 times e tow the 0.254 times the square meters. That's the original model that we validate him up to the assumptions that is significant. Let's goto assumptions. We go back to the kicker blood and it appears that all the latter bones off their issues are actually language it diagonally except for the 1st 1 But that's okay because all the other ones and the ones in the mirror which are the most importance, are really aligned with it. I am so we can say we're good in normality. Assumption. Let's go back to their receivers. This appears to be a really bad right. It appears that we have a line on the feed it receivers. However, if we check the value on I actually, it's really, really small to e to the minus four. That means that actually this is just were doing and really assume on the recipes. But they are really, really close to zero every rest of you, the issue that the 1st 1 is below Syria and other ones are above it doesn't really matter because the values are really, really small. Therefore, we can say that we are good and I'm saying Stay city off course. It would be awesome to have a live and test, but for that we would need for each square meters more than one measurement down. Let's go below to see what are things we can get from here in this plot, we're getting the straight line. That actually is the model for the local price against the square meters and we don't really see it here because the moral was really good. So their confidence in their eyes are really, really small. But before morning it's not so good there, for the uncertainty was bigger. We would see a gray area where it would say, What's the confidence in the roll for the whole prediction Dynamic rage on. That's what we get below. We get four. They mean they mean minus one standard deviation which is the square root of the bearings on I mean plus one standard deviation. What is the confidence in Terrible? So, for example, for 226 square meters, we know that the local room off the price will be approximately 15.6 on. The confidence in general is so, so small that actually they lower never Limits for the first decimal are the same and we have the same for the other two values. Remember, that doesn't mean that the confidence interval just one value. This means that in this case, the mall was really accurate. So the differences will start to show on the other day, Seamus, that right now they are not so important with all these we actually went on from a non linear model. We need a clean arise it. We did a linear regression We used on overto actually verify its significance. We recover the original relationship, which was exponential. And finally we also got confidence in the rust for the slope for the intercept. So for the whole Mullen on for the predictions on unseen values. So if we have our real estate agency comes and says What's the price for a house off 450 square meters, we can actually go and predict it with the confidence interval. This is a really nice prop up toe everything we've seen in this course because we started on probabilities and then we went to to test which we used every time and we merged Unova linear regression. All the allegations took it. The fortune and relationships off to quantitative barrels with all these, let's do a prop up off the scores. Thanks

Course Wrap Up

[Autogenerated] on. Finally, we got to the end of the last model. Congrats to you. You got to the end. In this model, we discover how to get the best linear predictor of two variables in the case they are considered. You will hear how to assess the validity of that model. Finally, we says how to Bali a those assumptions and how to attack none in your cases by the Paris station. But above all, we learn how to tackle all the different problems that we can have with our that in the case that both wearables are quantitative. We want to know the linear relationship or another type of fortune and relationship with the lunar station. We can get a lot of power that can analyze it in great ways, but this is only the start. From here on, it's a never ending journey off understanding more and more complex models where we can mix regression and a nova, start putting Sam barrows us random and start any more components to have more barrels and find more ways to get the best of our data. It was a pleasure, Toby, on this journey with you and expect to see you again. Thanks a lot. On, Please, please leave me a message off anything that you may have a question on the course. These courses are so difficult to make. They take so much time on power point. It's not really my friend. So it was an honor to be in this journey with you. And I expect to see you again. Bye.

Course author

Axel Sirota

Axel Sirota has a Masters degree in Mathematics with a deep interest in Deep Learning and Development Lifecycle. After researching in Probability, Statistics and Machine Learning optimization, he...

Course info

Level	Beginner
Rating	★★★★★ (25)
My rating	★★★★★
Duration	2h 49m
Released	22 May 2019

Share course

f   🐦   in