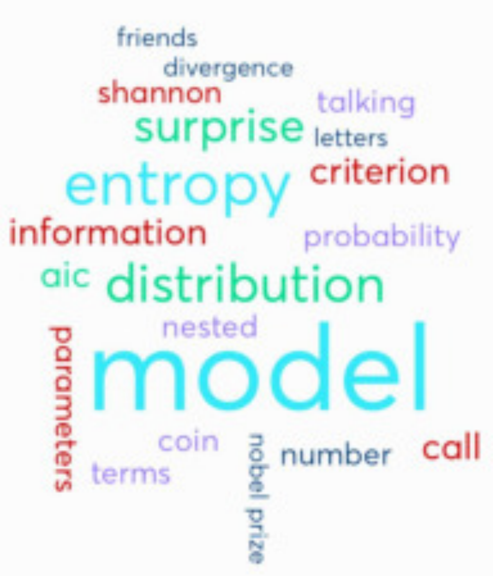
The Podcast *Quantitude*

with Greg Hancock & Patrick Curran

Season 3, Episode 26:

The 411 on Information Theory

Published Tuesday April 12, 2022 • 1:03:37



SUMMARY KEYWORDS
model, entropy, distribution, surprise, AIC, criterion, information, call, Shannon, parameters, nested, coin, probability, terms, talking, number, letters, divergence, friends, Nobel prize

Greg 00:00
All right. Hi everybody. My name is Greg Hancock and along with my always surprising friend Patrick Curran we make up quantity food. We are a podcast dedicated to all things quantitative ranging from the relevant to the completely irrelevant. In today's episode we talk about information theory, what it is, where it comes from, how it works, and how it can be used to make comparative model inferences along the way. We also mentioned Pennsylvania six 5000, the Time Lady, the Nobel Prize for awesomeness juggling and unicycles Enigma, imaginary friends, lemon juice code, red giants and white dwarfs. Bits. A level 11 Paladin, the Hungarian Forrest Gump, Snake Eyes and boxcar Willies, the pay the Reaper divergence criterion, and getting inspirations on a train. We hope you enjoyed today's episode. I have a question for you. Do you remember your phone number when you were a kid?

Patrick 00:56
This is gonna be a little revealing to the audience in a somewhat embarrassing way. Are you ready? 798579279857927985792 that is Patrick's phone number. Wow,

Greg 01:16
did you come up with that? Or do your parents so

Patrick 01:17
no, that was mom, I was maybe five. And we all sang the phone number song. Wow. Back in the day it was in case I got abducted and then dropped off at a rest area somewhere. Yep.

Greg 01:30
I'm pretty sure if someone abducted you, you would be dropped off at the next rest area. You could just walk home. Yep. I didn't have a song. But my phone number was sunset. 39655.

Patrick 01:47
is that a lyric Adam Glenn Miller,

Greg 01:49
right? It's so dated. Doesn't it sound like a 40s movie or something? Tricks did get me on set 3965. We used to also call the Time Lady. So if you wanted to know the exact time there was a phone number that you could call? I don't know if you had that. For us. It was ti for 1000.

Insert 02:12
Good afternoon. At the tone the Daylight Time will be 1238 and 50 seconds.

Patrick 02:19
No, again, I was a later generation than that.

Greg 02:23
Wow. All right. What about if you needed to know someone's phone number? Like in another city? Did you have some number that you called?

Patrick 02:30
Not that I remembered we had the good old phonebook. But that was it.

Greg 02:34
So did you not grow up with 411? Being the go to information number? No. Are you familiar with the expression 411 To represent information?

Patrick 02:44
No.

Greg 02:45
Come on. Hey, give me a 411 on that you that's not an expression that you've heard? No, I have to think that you're the problem in this scenario. 411 is a thing. I hope it is.

Patrick 02:58
There. We'll go with that. Let me revise my statement. Oh, four. Yeah, everybody knows that. I got to 17 year olds, I hear that three times

Greg 03:06
a day. There you go. That's what I needed to hear.

Annie & Kristie 03:09
All right. This is any Christie. We actually do know what the 411 means was that, uh, all the time. Sorry, Dad, you seem to be the weird one here. Again, again,

Greg 03:20
this particular episode is going to be about information. There's this particular area of statistics that sometimes is really prominent and things that you and I do and sometimes it kind of lurks in the background. That's something called information theory, a lot of what we do is built on some of the principles of information theory. And what I thought I would do is try to give people the 411 on information theory, in this particular episode, help people understand what it is what it does, how it touches on a lot of things that we do

Patrick 03:50
good. I asked a legit question. Yeah, you're the one who picked this topic. This is going to drop APR 12. Tell me you did not pick the entire topic because it's close to April 11, which is 411.

Greg 04:05
Moving on, we all right. I'm going to start this story whether or not you're with me,

Patrick 04:13
like I'm a Joyce at this point. Our story starts

Greg 04:15
out with someone that I think hands down should have won a Nobel Prize. Claude Shannon, I don't know why he didn't win the Nobel Prize. It might be because he was in a field that didn't align particularly well with the Nobel categories. Like I know, for example, there's no Nobel Prize specifically for mathematics. There's the bell prize which Norway offers. So who's someone who should have won a Nobel Prize in your mind? Oh,

Patrick 04:41
well, there's an obvious one. Okay, Cal Ripken, Jr. Sure, the Ironman 21 seasons, I think with your Baltimore Orioles, the same Baltimore Orioles where you pretended to be an architect to get me down to the field. We walked in the front gates So people as long as you roll papers up and point to things everybody thinks you're an architect. Yep. But he had 2632 straight game starts. He said you just show up and do your job and good things happen. The Ironman of baseball he is my hero Cal Ripken should have won the Nobel Prize for awesomeness. Is there a Nobel Prize for awesomeness and

Greg 05:32
scrolling through the Swedish categories that they I'm not seeing awesomeness. Specifically, there was this guy named Claude Shannon and Claude Shannon worked in the Bell Labs in the 1940s. And later, but he was an electrical engineer. And by all accounts, obviously a very, very bright guy, as you'll come to understand, but also kind of a quirky guy like the guy who could juggle the guy who could ride a unicycle down the hall. So he was this eccentric, but really, really smart guy. And one of his passions was cryptography. So how we encode things, how we break codes, and it, it turns out, that was a particularly important thing to know how to do during the Second World War.

Patrick 06:11
Oh, now you got my attention, Enigma!

Greg 06:14
You go, a neuron fired. For God's sake, that's amazing. Did you used to write like coded letters to friends or anything as a kid,

Patrick 06:21
my imaginary friends, I never got coded letters back, it was kind of strange.

Greg 06:27
We had our own symbol system. But we also wrote letters in lemon juice. So if you take a piece of paper and you write in lemon juice, it's invisible. But then if you hold it over, like a really hot bulb, 100 watt bulb, then the letters actually start to turn brown, and you can read the secret message. That's how

Patrick 06:45
they cracked the Soviet code back in the Cold War.

Greg 06:50
Anyway, so this is Claude Shannon guy had some amazing friends also in the 1940s. One of them who would sort of sneak over was Alan Turing, the Alan Turing would come over from London hanging out at Bell Labs. And apparently, they would sit there and have tea and not talk about all the things that do make during the Second World War,

Patrick 07:10
like, what are you up to? Nothing? What are you up to? Nothing? All right, we're gonna go ride unicycles. Yeah.

Greg 07:17
Another friend of his was John von Neumann, one of the leading computer scientists of the time, between Alan Turing and John von Neumann. I mean, this guy was really hooked up in terms of people who were, I believe, the tired term for me as thought leaders, but they actually were. So Shannon had an incredibly ambitious goal. And that was to try to figure out how he could quantify any type of communication. And in his view, pretty much everything we do is communication, whether it's talking, or pictures, or movies, or art, or music, the spread of disease, all of these things are variations on communication. And he wanted to come up with a mathematical model that could capture the information that was contained in communication, whatever information means in each of those contexts, that was his goal. And that sounds so unbelievably ambitious, that it sounds made up.

Patrick 08:19
I actually have two thoughts on that is one is, it must have been an amazing place to be in Bell Labs during the 40s and 50s, as people forget the incredible achievements that came out of their research division. And ostensibly, it was a business endeavor. But they had some of the best scientists in the world working there at various times. That's absolutely remarkable. But the other one is exactly in that same period of time of not only coming out of World War Two, but then moving into the Cold War.

Yeah. And that notion of cryptography of imposing structure on chaos through numerical analysis is just mind boggling.

Greg 09:08
Yeah. And Claude Shannon wrote a paper that is like the paper, the landmark paper called A Mathematical Theory of Communication in 1948. And it laid the foundation for trying to encode what he termed information, and it lays the foundation for what we now refer to as information theory, how do we quantify information? How do we communicate that information, how much information is contained in messages? And this caught on in a huge way, engineers found ways to use it physicist psychologists, we obviously use information right the word information is something that's certainly on our radar, and it's just got its tentacles into so many things. So I want to set the stage for this. I'm going to think of a number between one and 32 and I want you to try to guess that number between one and 32 Any number let me make the is easier you can ask a series of yes, no questions to try and hone in on that if that's what you like.

Patrick 10:05
All right, I know you reasonably well. It's Friday, you have these goofball superstition beliefs. It's 13.

Greg 10:15
Screw you.

Patrick 10:17
Seriously? I'm not talking about a one in 32nd chance on that.

Greg 10:25
All right. All right, we're gonna start that again. I'm thinking of a number between one and 32. You can ask a series of yes. No questions and stop using your psychology. Neuroscience? I'm not sure which it is. Um,

Patrick 10:38
okay. Is it even? No? Is it less than 16? Yes. Is it a prime number?

Greg 10:45
Yes.

Patrick 10:46
Is it less than 13? Yes. Is it three?

Greg 10:51
No. Is it five? No. Is it seven? Yes, it is. Right. So I wouldn't exactly stand up and take a bow. If I were you. It's entirely possible, you could have been more systematic. So I did this

Patrick 11:09
with Quinn, right? Your 17 year old son,

Greg 11:12
my son, actually 16. So the way Quinn approached it is he said, Is your number 16? or below, which is what I asked. Sure it was. And then he asked, is it eight or below and so he homed in on that. And he said, In the end, I could get any number in five questions or less. Now it took you I think, seven or eight questions to do it. But he unlike you is very logical in his thinking. And I said, so how do you get that you can get it in five questions. And he said, Well, you can only split it five times, right? You split 32 ones to 16, twice to eight, three times to four, another time to two, and then you're down to the last two. And he's exactly right. If you think about it, as this tree branching out, it's gonna branch up to five times before you hit the number. That's not the only way to do it. You could have another series of questions like, Is it five. But that might be slightly less efficient than using a more principled approach.

Patrick 12:11
I would like you over dinner tonight to remind Quinn that I can pick them up and throw them down a flight of stairs, if I feel like it. Just point that out if you

Greg 12:20
mentioned that. So in that case, you know, when you're guessing a number, each question that you ask if you say is your number 16? or below, I say yes or no, there's the same amount of surprise. If I say yes, or I say no, you're like, okay, and then you say, is it below eight? There's not a lot of surprise each time you sort of plod along until you hit it. What if I said to you, the sun will rise tomorrow? How surprising is that statement to you?

Patrick 12:46
That would be very low surprise,

Greg 12:48
very low surprise. Exactly. What if I told you that the sun is going to go supernova tomorrow and become a supermassive black hole?

Patrick 13:00
Our son is not big enough to be a black hole, it will go red giant and then be your white dwarf. Okay, so the surprise would be very, very high, it would

Greg 13:09
be very high. That's right. And your case, it would be infinitely high. Because you just said it couldn't even possibly happen. Right?

Patrick 13:15
Well, then also we would all be atomized when that happened. So I wouldn't have a real recollection of being surprised. Hey, I

Greg 13:23
didn't expect. So when you go through that process of guessing my number at each one of those question points know, the level of surprise is more or less the same irrespective of what answer I give, when we go through day to day and I tell you, hey, the sun is gonna rise to Hey, the sun is gonna rise to hate. And yes, I know the sun doesn't rise, the Earth rotates. Overall, there's not a lot of surprise, unless I tell you, the sun is not going to rise on a given day and find red giant white dwarf, whatever. That would contain some amount of surprise. But on average, there's not a whole lot of surprise that goes on from day to day. Well, one of the key concepts in information theory is this notion of surprise, and how we quantify surprise. And it's going to relate to probability, which is not in and of itself, very surprising. So I'm going to give you three scenarios, you're gonna have to keep track of a little bit of stuff in your head right now. There we go. Imagine you have a coin, and the probability of it coming up heads is one in eight. Okay, I'm writing this down, actually. So a measure of surprise, not the measure, we're ultimately going to get to but a measure that would capture some degree of surprise, if I told you, hey, I flipped the heads would be one over that probability. So if you've got a one and eight shot, we could say one over that probability. Well, it just gets you right back to eight. We could say like, I'm eight surprised whatever that means. That's one way to operationalize how surprised you are right. So if I tell you I flipped ahead you go, wow, there was only one in a chat. I'm somewhat surprised that you did not off the charts surprised, right, but somewhat surprised. Scenario two. I have a coin that has a 5050 shot have coming up heads. So if I tell you hey, Patrick, I just flipped ahead. So how would you characterize your surprise?

Patrick 15:06
I'd say, Yeah, that's what you'd expect half the time. That's right.

Greg 15:09
Much more nonplussed, much less plus to what did we decide on that one? But

Patrick 15:14
it was equally likely. Can we just go with that? Okay.

Greg 15:17
And if we think about this in terms of the probability and its reciprocal, like we did before, we could say that whereas before, you would have been eight surprised. Now you are too surprised one over point five. That's one way to characterize surprise. So now, if I tell you I have a coin that will only come up heads, and I tell you, I flipped the heads. How surprised? Are you? None surprised? None surprised? Well, if we think about this, following the pattern that we've been doing one over the probability, and that case, the probability of the heads is one. So one over one gives us a number one. And while that's the smallest of all the numbers we've had so far, the idea of one characterizing the absence of surprise is a little bit funky. And so historically, the decision was made to instead of use one over the probability as a measure of how surprising a particular event is, the log of that was taken, and it was originally a base two log,

Patrick 16:13
remind listeners, what a base two log is.

Greg 16:17
Do you want to do that? Patrick?

Patrick 16:18
No, no, no, you're on a roll. I don't want to interrupt. Why don't you go ahead.

Greg 16:22
I appreciate your commitment to the flow of pedagogy. Yeah, so the logarithm base to have a particular number is what power would you have to raise the number two to two equal that particular number? So the log base two of eight, what power? Would you have to raise to two to get to the number eight? The answer is three. The log base two of eight is three, the log base two of the number two, do you want to do this one, what power do you have to raise to two to get to go ahead, you're on a roll. It would be one. And then if we have the number one, and we wanted to take the log base two of one, what power would we have to raise to two to make it one and the answer is zero, something to the zero power gives us one. So a measure surprise that was used in the development of the concept of information was a log base two of one over the probability. So for that coin that had a one in eight shot of turning up heads, we would say that there is three units of surprise. And for the coin that had a 5050 shot, we would say there is one unit of surprise. And for the coin that had only heads to come up, we would say there is zero units of surprise, zero bits of surprise. In fact, the term bit comes from this, I don't mean the English term bit, but the computer term bit binary digit, because when you have a 5050 coin, one binary digit can help you to differentiate between a heads and a tails. But if you have a coin that only comes up heads one in eight times, that's essentially like having three coins worth of information, right? The probability half a half a half. And if you have a coin that only comes up heads, there's no point in having the coin at all, because it has zero surprise, it has zero bits of information to be

Patrick 18:08
able to contribute. So that nicely bounds it at zero, right? Yes, that's exactly

Greg 18:12
right. So if we go back to when I asked you to guess my number, let's go back to the second time, I asked you to guess my number out of 32, we could say well, the chance of getting that right on the first shot without your psychobabble would be a one out of 32. The inverse of that is 32. And the log base two of that is five meaning that being able to get that number on a single guest is like five bits of information. That's pretty incredible. And this quantity that we're talking about right now for a particular event, which is the log base two of the reciprocal of the probability, this measure of surprise has the name entropy. I don't know if that's a term that has come up in your life at all.

Patrick 18:56
I've encountered that more in the physical sciences. But it certainly exists in statistical literature as well, particularly if you talk about things like time series analysis, that's a term that occasionally comes up

Greg 19:08
Absolutely. And right now I'm talking about it in the context of a single event like flipping a heads, but usually it is talked about and for our purposes, it's going to be talked about in terms of the surprise associated with every possible event in a probability distribution. So if we have a coin that comes up heads one and a times, or P of point 125, then the probability of tails is point 875. So we could figure out how much of this surprise or entropy is associated with each of those outcomes. And then we could average them so that we have a sense of how much of this entropy exists across the entire distribution. So for the coin that we have, that is point 125 heads and point 875 Tails, there's a surprise associated with the heads of three because we would have to raise two to the third power, the surprise associated with getting a tails is going to be much smaller because the tails is much, much more likely. And if we went ahead and took the reciprocal of that probability, and the log base to the surprise associated with flipping a tails would be point 193. And even if we don't necessarily have some internal metric for that, that is low surprise, certainly relative to the heads that we have. Now, if we take an average of those two numbers, a three, which is the surprise associated with the heads for that coin, and a point 193, which is the surprise associated with the tails, and we average those two, we don't just want to add them up and divide by two, because one of those events is far more likely one of those events is far less likely. So we're going to weight them in this average according to the probability of their occurrence. So if you got one in eight shot of getting a heads and seven eighths shot of getting tails, we're going to wait those two measures of surprise. And overall, we get an average entropy of point 544. That's some measure of inflammation associated with the system, some measure of chaos, some measure of surprise, there's so many different terms that are going to be attached to this. But one way to think about it, and this is going to be really, really weird. But this notion of entropy of point 544 is if we want to put it in terms that we might better understand, it would be like having a coin, that if we took two and raised it to that power, that coin would have about 1.46 sides. So that's less than having two sides. And that's because this coin has less what we might feel is surprise overall associated with it, because most of the time, it's gonna give us tails. So this idea of point five for entropy sometimes is characterized back in terms of well, how many sides would that coin have to have to correspond to this? That is for a coin that has a one in eight shot of a heads versus a seven and eight shot of tails.

Patrick 21:55
I really like that notion of the sidedness of the coin, one of my kids is way into dungeons and dragons and the number of die that they have that are multifaceted.

Greg 22:09
I liked actually that you brought up Dungeons and Dragons, it's like having a dye that has not even two sides, it has 1.46 sides. So if you roll that die, there's not a lot of surprises associated with it. But if we had, I'm still using the coin examples, we had a coin that had a point five chance of heads and a point five chance of tails. And we computed that thing that we're calling surprise, the log base two of the reciprocal of probability, there is one unit of surprise associated with heads, there's one unit of surprise associated with tails. And if we average each of those one units of surprise, overall, we have one unit of entropy associated with that. And not surprisingly, that would be like having a coin with two arrays to that one power, like having a coin with two sides. And that really is sort of like our go to metric, right? A coin with two evenly probable sides has one bit of information, the coin that has a greater chance of tails has a smaller amount of information. And if we extend that all the way to that weird coin that I said

that only flips heads, the amount of surprise associated with heads with it's got a probability of one so you take the reciprocal of that probability, you still get one log base to the surprise associated with a coin coming up heads is zero. The surprise associated with the coin coming up tails is technically undefined, but it doesn't matter, because there's no probability associated with it. So if you take the average of the surprise of the two outcomes, all you got is the one outcome the heads coming up. And so it is like having a coin with one side. And that's not so surprising. We're in your case, a Dungeons and Dragons die that has only one side. Oh, I wrote a senator. I don't know what your people do.

Patrick 23:49
I played Dungeons and Dragons. once in my life. It was high school. It was all paper and pencil, I got some kind of weapon on a roll that was fairly rare and messing around. This guy was hassling me. And so I said, I pointed at him and unleashed the fury of whatever. Well, evidently that there's no joking around, and dungeons and dragons, and they took that as an attack on my friend and I killed him, which was a character he had been developing for months and months. And they all quit and kicked me out of the group. That's a true story.

Greg 24:32
You do not mess with a level 11 Paladin, or whatever. I don't know. So for these three coins that we were talking about one that has a much greater chance of getting tails. It had this entropy of point 544 And the coin that was a 5050 shot had one unit of entropy across heads and tail possibilities. And then the coin that had no shot of tails only a shot of heads had zero entropy associated with, in fact, Shannon derived this particular quantity, which is like that weighted average amount of surprise across all of the outcomes in a particular space, but he didn't even know what to call it. And so the story goes that John von Neumann was there at Bell Labs and Shannon was having a conversation with this guy who is like huge in terms of math and physics and computer science complete Rockstar. And the story goes that John von Neumann told him to call it entropy, because what he was talking about was related to things in physics from Boltzmann and physical chemistry. And then the story goes that von Neumann also added, besides, no one really knows what entropy is. So just call it that and no one will mess with you. I will tell you just a little side note, that for me is really cool. Von Neumann was actually a good friend of my grandmother's that is really, really cool. So my grandmother, I've mentioned my grandmother a whole lot of times. This is just so weird. She trained with Maria Montessori. She was friends with Anna Freud and met Sigmund Freud at their home in Vienna. And she was friends with John von Neumann when she emigrated to the United States after the Second World War, because there was this very tight knit Hungarian community in New York City. And she and Yano, she, as she called him, became very good friends, which is just crazy. In fact, one of my favorite things was she and von Neumann went on a car trip to Princeton, so that my grandmother could have a long lunch with Albert Einstein just the two of them a four hour lunch with Einstein, how crazy is that? That is quite remarkable. She's like this little Hungarian Forrest Gump, who just traveled through all of these circles. So I am two degrees of separation from von Neumann three to Claude Shannon, which is kind of cool. But she did not, to my knowledge ever hang out with Kevin Bacon. Alright, but I digress. Alright, so I'm going to take these ideas and move them up into slightly more complex distributions. Right now we were just doing with a distribution of heads and tails, right? Now I'm going to imagine that we have a distribution where I have the numbers two through 12, not one through 11, two through 12. So there are

11 possibilities there. Imagine it's like an 11 sided die. My understanding is, is that it would be called a hen deck a hedron, or an uneca hedron, or an N deck a hedron.

Greg 27:35
It ends in DECA hedron. Right? So just put whatever prefix you want. And congratulations, you're like a level one dwarf in Dungeons and Dragons. And yes, I know there are people out there who are like what I can't believe he just said that. I'm sorry, you know who you are. So let's imagine for simplicity that each side of this 11 sided die had an even chance of being flipped. So one out of 11 shot. If we figured out the surprise associated with each of those numbers two through 12, and then average that amount of surprise. It turns out that this Shannon entropy would be 3.46 We would say that this die has 3.46 bits of information associated with it overall. Now using that as a frame of reference, where the numbers two through 12 are evenly likely. Let's think about rolling two six sided dice. If I tell you snake Guys you know what that is?

Patrick 28:29
Yep, Snake guys. And boxcar Willie's snake guys are two ones. And boxcar Willie's are

Greg 28:34
two sixes Excellent. What's the chance of getting Snake Eyes? One in 36 1236 and boxcar Willie's I don't know about the Willie. But boxcar Willie's

Patrick 28:42
Come on dude. No boxcars. I knew somebody who had a Pennsylvania six 5000 number I thought it was your you know what boxcar Willies are? It is also one in 36.

Greg 28:54
There you go. So do your people have some fancy name for rolling a seven? It's just

Patrick 28:57
throwing a seven or rolling a seven. So somewhat less creative? Oh, cool name like that. All right.

Greg 29:04
So do you know what the chance of throwing a seven is? 636. Wow, that was really impressive. So rolling a two or rolling a 12 would be a lot more surprising, right? That would have more surprised than rolling a seven, because there's a greater chance of rolling a seven. And we can actually quantify rolling a two in terms of its surprise if we took the inverse of the probability and took the log base two, we would get a 5.17 which in and of itself might not hold a ton of meaning to you. But relative to rolling a seven, the surprise associated with rolling a seven would only be a 2.58. So there's a lot more surprise associated with rolling a two or rolling a 12 than there would be rolling a seven. And if we average the surprise associated with a 234 all the way up to 12. The overall average surprise would be a 3.27. And if we think about that, in terms of coins, that would be like having a two raise to the 3.27 power a coin with 9.65 sides or like having a die with 9.65 sides that were equally probable. So whereas the other die where all the numbers were equally likely was like having 11 sided die having a die where some of the numbers are a lot more probable than others, overall has less information that

would be like having a die between 9.65 sides. So what we're doing is we're quantifying the amount of information associated with the entire distribution through this measure called entropy. And one way you can think about it in terms of distributions, and this is going to become increasingly important to us because we think about distributions all the time is you can think about entropy as almost being how flat a measure of how flat the distribution is, the flatter the distribution, the more information there is, the more peaked the distribution overall, the less information because we are less surprised by those events that are closer to the peak and more surprised by those events that are away from the peak. So entropy can be tied to this idea of a distribution. And that's going to be really, really important for us.

Patrick 31:01
Could I throw you a legit question? Yeah, I'm just trying to get my head around. What is the motivation for doing this one, we have a binomial probability mass function, we have a moment generating function, we can derive probabilities, what's the value added to moving to this entropy like expression? Well,

Greg 31:23
first of all, the idea of entropy is not just for these kinds of discrete distributions that I've been talking about. It also exists for continuous distributions. And that's important because you and I, although we deal sometimes with these discrete distributions, we deal a lot more with distributions that are either continuous, or we lie to ourselves and say that they're continuous, right? So the idea here applies to distributions that have continuity. And that might not be enough cell for you yet, but I'm getting there. Okay, with regard to Claude Shannon, and the idea that he was developing this kind of mathematics for communication systems. Let me give you an example to just close that loop, and then bring it into the world that you and I tend to live in. Have you ever read one of those passages that several of the letters in each word have been replaced with almost seemingly random letters? And you can still read the passage perfectly? Do you know I'm talking about oh, yeah,

Patrick 32:16
there's some classic cognitive psychology experiments, where how many? Can you punch out and still understand the word through the context? Yeah, there's some really cool studies where you read paragraphs, and there's some words that you can omit almost entirely, but still read the word. Yeah. Based on the information around it.

Greg 32:39
Yeah, that was a perfect description. Exactly. Right. And what that signifies is that not all of the letters in a message, carry the same amount of information. There are some that, you know, if you eliminated them, you would go I have no idea what this is saying. But there are others that are really superfluous. And you can put blanks there. You can throw in random letters, and you read that paragraph as if there was nothing wrong with it. It's weird, right? You read it and you go, Hey, how does my brain know all of this? If I tell you, I'm thinking of a five letter word, and I asked you what the first letter of that word is? What do you think the first letter of that word is?

Patrick 33:15
Yes. Okay.

Greg 33:16
You're wrong. Sorry. But you had 26 letters in our alphabet. Anyway, you had 26 letters to choose from? I'm going to tell you, it's a q. Now I'm going to say what's the second letter? You exactly right. So that first letter contains a lot more information, second letter, basically no information, because in our language anyway, the letter you almost always follows Q. And if we know that you is that second letter, then the third letter is almost certainly going to be a vowel. And what Shannon was trying to help us to understand is that in our communications, not all letters contain the same amount of value in communicating a particular message. And we do this all the time, maybe not you. But our kids do this all the time, when they're texting, right, they have distilled entire conversations down to just a few letters,

Patrick 34:04
or an eggplant emoji.

Greg 34:08
So the idea here is that there's a lot more communication that can happen. But all the stuff that fills in the middle, we understand whether it's via convention, or just our understanding of how language works, that not everything in there is worthwhile. And so we can distill down WTF from a somewhat longer phrase or, and Shannon was interested in developing a mathematical model to understand how much information is there really in communication? And how can we make it a more efficient and then be in cryptical? Now, obviously, there's a lot more to that. But this isn't in the end about communication theory. I want to bring it back to information theory and distributions. You and I deal with distributions all the time. And every time we use a model. One way to think about what we're doing with that model is trying to represent the distribution of data through some number of moving parts. Whether you think about them as knobs, or sliders or whatever, we call those in our models the parameters. And when you and I do structural equation models, for example, we have parameters associated with that model. But at the end of the day, what those are, are governors of a particular distribution, a particular model implied distribution. And if we think about the data that we have as containing a certain amount of information and a certain amount of noise, what we're trying to do is come up with a model that has really good fidelity to the underlying process the underlying signal while cutting through the noise. So the idea of information associated with models and randomness and noise and systems is something that Shannon's work helped us to try to think about when we get into the modeling that we're going to do, in fact, something that you and I have talked about a lot, both computationally, I would say and also philosophically has to do with this idea of underfitting versus overfitting of our models. So we want to have this model that finds that sweet spot that has enough parameters to be able to explain what's going on satisfactorily, right to have fidelity of the signal, but not so many parameters that it's fitting more than just the process that's operating underlying the data. But it also starts fitting the noise, right? If someone said, Hey, we want you to come up with a model that completely explains what's going on, I would just add all the parameters under the sun that I needed to until I perfectly nailed every single contour of the distribution. And so we try to find that sweet spot, which is really tricky to do, where Sir William of Ockham tells us it is vain to do with more what can be done with fewer and so we try to find that balance where we are not underfitting. Right, not where we are failing to capture the process, but not overfitting, where we're starting to pick up some of the noise.

Patrick 36:56
So it's the classic crossing courage with close enough for government work. It's exactly what we've talked on prior episodes is courage is numerically represented in your degrees of freedom. So we start the game, let's say in SEM, but it generalizes to a whole lot of modeling frameworks, we start the game knowing we can guarantee a win. And a win is perfectly representing the characteristics of the data that we observed in our sample. And it is an uninteresting game, because no matter what data, you email me, I can perfectly reproduce the covariances variances and means by fitting a model, how do I do that I do it by saturation, I estimate as many parameters as there are observed moments in our sample data. So it's kind of funny, we all play a game that we know we can win. But we don't always want to win, because what we want to do is impose restrictions on the parameter space to say, I'm going to attempt to reproduce the characteristics of my observed sample data, using fewer than all the pieces of information that I observed. And that's where we wander out into the minefield of model fit and goodness of fit and test statistics and modification indices and all of that. So that's what I mean by the courage is how many restrictions can we impose on our parameter space, and the close enough for government work is we are going to get Miss fit because of those restrictions? How much can we live with and still look at ourselves in the mirror in the morning and say, I'm not perfectly reproducing my data. But I'm pretty close. And it's close enough for me,

38:46
because I'm good enough. I'm smart enough. And doggone it, people like me.

Greg 38:52
I love that that is the perfect tension between these two things between fit and parsimony. And I'm going to use that theme and some of the things that I'm going to talk about in just a second. I love that setup. So what that means, though, is that when you fit a model that doesn't have all the moving parts, that would saturate it, in some ways with respect to our sample with respect to the data that we have, we're giving up some information, we're giving up the ability to explain some of that stuff. So in the example that I had earlier, if the truth were that the numbers two through 12 occurred with probabilities that are associated with rolling two dice and adding them up, that would describe a particular probability distribution, if someone said, I think the model is that all the numbers two through 12 are equally probable. Using that to describe what actually would be going on in the data. We would experience this loss of information, right, we would be throwing some stuff out, presumably. And the question that is relevant here, and that relates to something that you were just talking about has to do with a comparison of two distributions of data. distribution that is associated with one model. And a distribution that is associated with could be another model could be a distribution that is associated with truth could be a distribution associated with our data. There's this inherent need to talk about how two distributions relate to each other in order to gauge how well one model is doing with respect to some other frame of reference. And it was back in the mid 1900s. So after when Shannon was working in the 1940s, in the 1950s, colback, and Liebler, came up with a way to quantify the degree of separation between two probability distributions. So we can think about any two probability distributions that we want, like the one that I said, where we have equal probabilities of the numbers two through 12, we could think about the other distribution as the probabilities associated with throwing and summing two six sided dice, we could think about it as the difference between a normal curve and a t distribution with three degrees of

freedom. Whatever two distributions, we have called back, and leibler, came up with a measure that is now called KL divergence. And sometimes people call it a measure of distance, that's really not such a great term for divergence is probably a little bit better. Or sometimes people call it relative entropy. But the idea is to try to get a sense of how different these two distributions are. Or if you think of one of those distributions as some frame of reference. The question is, how much information do you lose when you use a model associated with a second distribution on a process that's actually governed by a first distribution. And the idea of the KL divergence is so simple, if you think about it for a discrete distribution, like for the numbers two through 12, there's an amount of surprise or entropy associated with each number. And in the model that says the numbers are all equally likely, each number has some degree of surprise, or entropy. And in the other model, where the numbers probabilities are those that are associated with rolling in something to six sided dice, each number has a different amount of surprise, or entropy. So number by number two through 12, how different is the entropy? KL divergence literally does the subtraction for each entropy value? And then it aggregates those differences across the whole set of values. But just like the entropy for a single distribution, it adds them up by weighting them according to the probabilities of each number in the distribution. Wait a minute, which distribution does it use to weight it? That is a good question. That means if I treat the equal probability model as some reference distribution, I will get one measure of KL divergence, which is how much the dice summing distribution diverges from the reference equal probability distribution. And if I treat the dice summing probability distribution as the reference distribution, I will get a different measure of KL divergence, which will be how much the equal probability distribution diverges from the reference dice summing distribution. This is why we don't call it a distance, this asymmetry instead, we call it divergence, but it's such a simple idea, right? The probability weighted sum of the differences in entropy for each value in the distribution. And that's why it's called relative entropy. And even though I described it using that discrete two through 12 distribution exists for continuous distributions to it is beautiful. It is simple, it is understandable. And it builds on all the Shannon foundational stuff. And it's directly on point for what we care about when we are doing modeling. how much information do we lose when we choose one model over another? And that's what we do all the time, right? That's exactly what you described, where we have some model that is more parsimonious, that has more courage that has more degrees of freedom and has fewer parameters, we know that we're not going to be able to reproduce the data perfectly. One reasonable question is how much gap then is there between a model that perfectly reproduces those data and the model that we hold in our hand, and the colback leibler Divergence measure is something that in the 1950s, was developed to be able to characterize how different two distributions might be from each other?

Patrick 44:01
And see, I would reframe that as the paid the Reaper divergence criterion, right? You have a model? And you're asking, what is the cost to my model fit of me imposing these restrictions? Dude, how have I not won a Nobel Prize myself?

Greg 44:22
So yeah, I love that paying the repo because that's exactly what happens, right? There's a degree of disparity between these two distributions. And that's what you're willing to pay, because you think that the gap between the two or at least you hope the gap between the two really represents noise and that your model in the end is capturing the signal. Now, if my sole goal were to minimize the gap between

the distribution implied by my model and the distribution that the data actually have, as you said, right, you just saturate that. And then there is no divergence between those two, callback and leibler came up with this measure of divergence as a way to gauge the distinction between two distributions. But then what week could do in theory is if I had, let's say, three candidate models, right model one, Model Two, model three, all of which imply certain distributions. And I had some reference distribution, whether it's another model, whether it's associated with the data, whatever, I can look at the divergence of each of those models with reference to that base model, that comparison model. And I can tell you which of these three is closer. And if that's my criterion for choosing, then I would say, Oh, well, I pick model two, because it's the one that has the least divergence from whatever that base model is that we might be using. So the idea that callback and leibler had in 1950s, was a really powerful one, with the ever so slight problem that it is obviously biased in terms of less courageous models, when someone starts throwing more parameters in there than the model just goes right and start sucking right up to the data. The question is, how do we correct for that? How do we adjust for that lack of courage, and that moves us into the 1970s, where Akaike came in and went through a lot of amazing math, and in the end, sort of United Boltzmann and some of the original entropy stuff done in the 1800s is sure information, maximum likelihood, he pulled all of this into the same place. And ah, it's just so beautiful. What he did is he figured out the degree of bias associated with colback leibler divergence, and it comes out to be this beautiful little equals MC squared formula, where it includes some information about the likelihood associated with the data, some measure of fit, and then a measure that has to do with the number of parameters in the model, which for you is associated with courage. So it came up with this beautiful Akaike information criterion, and information criterion because it was derived from work of Shannon that led into work by colback and Liebler, and came up with the AIC and we use the AIC all the time. I don't know

Patrick 47:01
if this is apocryphal, and someone out there might know this. But in a talk recently, I heard someone tell a story about AKA AKA developing that index sitting on a train, looking out the window, and he had some insight, it's the equivalent of a thinking log.

Greg 47:19
I love that that's kind of better than the guinea pig thing for school, right, a little more poetic, I think.

Patrick 47:24
But what I like about what you're raising is that allows us to have nested models where we do formal likelihood ratio tests, where we literally estimate Model A and get the chi square and degrees of freedom, we estimate Model B and get the chi square and degrees of freedom. And we subtract the two and the differences distributed as a chi square that only works when the models are nested. And what you're moving into is shrugging and saying, Okay, why this guy? What if they're not nested? Well, that's where we move into the land of information based model evaluation,

Greg 48:02
right, which is not at all a significance test right is a different criterion. And a couple of the key insights that he apparently had on that train, one is that if I were comparing model one, model two and model three to some baseline model, whatever that is, what he showed is, we really can come up with this

particular index for each of the models. And the baseline model kind of cancels out all of these comparisons. So when you compute an AIC and there are different formulations for the AIC, it can be written in a variety of ways. One way is twice the number of model parameters minus twice the log likelihood associated with your particular model. So that is that balance of courage versus fit that comes into play. What he showed is that that's really all you need. And that is, in a way, a measure of generalizability. It's a flawed measure of generalizability. But the idea is that that correction that we put in there, sometimes people will call it a penalty that has to do with how many parameters that are, that correction that we have is something that gives us a sense of how much we would expect these results to replicate. And it's not unlike what we do, just in principle, when we compute an adjusted R squared, right, the R squared that we have is jacked up for our particular sample. But if we want to know something about this model, in subsequent samples, there's this little correction that has to do with degrees of freedom. Well, it's the same idea here that we have some overall measure of fit in the form of some function of the log likelihood. And we want to adjust that a little bit for something that has to do with how parsimonious the model is. So it's all brilliant, elegant comes together so beautifully. And it does exactly what you said. It allows us to make comparisons among models. And what you said is completely right, that when we have models that are nested, we can do a statistical test. Okay, okay, would say who cares? We can use this index to rank models in terms of their balance between fit and parsimony. whether they're nested or not, who cares? Lay out your models, the one with the lowest AIC, that best balance of fit and parsimony is the one that you would choose from among those, even if they are nested, but it doesn't matter, right? It's this great index that cuts across all different types. You know, you and I are very accustomed to comparing models that might be within one particular analytical domain, oh, we're gonna compare some factor models, this particular index can cut across different domains of models, we might have a factor mixture model, we might have a growth model, we could have very different competing models. And the AIC can actually cut across a lot of these domains and really helpful ways. Well,

Patrick 50:39
one of the biggest advantages of nested models is we get a formal LRT with a P value. And one of the biggest limitations of nested models is we get a formal LRT with the P value. And what I mean is all the baggage that comes with test statistics, P value, power effect size is part and parcel of those model comparison tests. And so it's a double edged sword to be sure. And as you're noting, there are many, many times when we might not have nested model comparisons, the big one that comes up, as you alluded to yourself is mixture models. mixture models, for reasons we talked about in a prior episode are not nested. When we compare one latent class to two latent classes to three latent classes, those are not formally nested. So commonly information criteria are used in that setting. Another one that I encountered somewhat recently, myself was looking at level one error structures in a GE analysis. And the details aren't important. But there are a variety of level one residual structures that you might consider, none of which are nested within one another. And also, you kind of don't care about the level one error structure in these kinds of analyses, as you want the one that's optimal for the data, you're not making substantive interpretations. It's kind of a shock absorbers so that you can do other things. Well, you can estimate a dozen of these different error structures, rank order them by information criteria, and pick the one that's optimal for the data.

Greg 52:19

And this particular index doesn't tell you if it's a good model, it just tells you if it is the least sucky among your models. But presumably, if you have some reasonable model in the kind of scenario that you're describing, that might be a very reasonable assumption, then we would pick the model that has the error structure associated with the best balance of fit and parsimony so that we could get to the substantive question of interest with a reasonable shock absorber in place in that example. So we've been talking about the AIC here, specifically as this culmination of Shannon's work into colback leibler into a chi EK. There are other information criteria out there as well. And I don't want to minimize those. There's the Bayesian information criterion from Schwartz, which was also in the 1970s. There's a consistent AIC, which adds a sample size adjustment for the behavior of the AIC, when you have smaller samples, there's a deviance, information criterion, DIC, there are a whole ton of different information criteria that do different things here and there, I don't want to get into all of those, I'm not able to,

Patrick 53:26
say nice shoe, you know, we could talk about that. You know, take time now to do that.

Greg 53:35
But the idea with different criteria here, there are different adjustments here or there. The idea is that you generally want to have some smaller information criterion values when you're making a choice among models. Now, one interesting area that their efforts to try to tug this into, which I think is really, really cool. Some work that Chris preacher has put out there, and a lot of other people as well hasn't quite made its way into our world yet. But it has to do with the idea that the AIC winds up building a correction term that is a function of the number of parameters that you have in your model. But there's a certain assumption that each of those parameters is somehow worth the same amount. And in some work that has been done more recently, like in the last 20 years or so, that idea has been challenged and picked apart a bit. And you and I know that not all parameters are created equal with respect to a particular model, right? So an error covariance in a structural equation model is a parameter and that only has some effect locally between the residuals of two variables. On the other hand, there is a factor covariance, we might have a two factor model and the factor covariance effectively connects work bifurcates these two whole sets of variables, are you going to tell me that those two parameters are worth the same amount? If the error covariance, isn't there the model kinda goes, you know, the parameters are going to be basically the same. But boy, the fit, which would be better if you threw in that error covariance, you leave out that factor covariance. And you could be screwed royally if the variables on one side of your model are related to the variables on another side of the model. So, the idea of correcting for model complexity requires you to properly define model complexity. And some of the work as I said, that preacher and a number of other people have been doing is to start figuring out what the value of every parameter might be in our model. And part of that has to do with figuring out how flexible a given model might be to adapt to changing circumstances. So imagine, just as a quick example, and this might require, I know, you don't have a third eye, but let's try to use your mind's eye. Imagine you have variables one and two, they're completely uncorrelated, and they have paths going into variable three, that's the whole model, that model has five parameters, one degree of freedom. On the other hand, you could have a model that says V one has a path that goes into V three, that has a path that goes into v two, that also has five parameters and one degree of freedom. A really important difference between those two models, though, if you think about all the different possible correlation

matrices that you could have, one of those models is a lot more versatile in terms of the parameter values that you could assign to get it closer to the possible distributions that you could have the possible correlations, you could observe than the other that model that says v one and V two are not connected, and they independently go into V three, there is no universe where v one and V two are correlated, if that's your model. So there's this whole corner of the universe where that model just fails hard anywhere v1 and v2 are actually related in real life. On the other hand, a model that says V one goes into v3 goes into v two allows for some degree of association between v1 and v2 and v3 by virtue of the structure that that model has. And so there's a lot more of this possible data space that that model could adapt to. So it has a greater as we might call it, fitting propensity. So now, if two people come up to me, one of whom does this model that has much less chance of fitting just by virtue of its lack of flexibility, and another model that has a much greater chance of fitting by virtue of its flexibility, who is going to impress me more if they come and say their model fits? Well, not the person with the Super bendy model, the person with the model that is much more rigid, even though they both have only one degree of freedom, that is going to impress me a whole lot more. So some of the interesting work that preacher and others are doing is trying to characterize models complexity in such a way that parameters are not seen as being equal, but rather the structure of the model as a whole. And the role that the different parameters play in that is taken into account such that models that are more flexible to cover more of the possible data space are really in some ways penalized. So the models that become the real tests are the ones that have less flexibility. So it's a very interesting area, where the whole information foundations that have been laid, there's some pull to try to get it to start to be a bit more complex. I think it's really interesting.

Patrick 58:13
I really like Chris's work on that in thinking about model flexibility, then it goes back to the surprise, so kind of squaring the circle as well. How surprised should you be that you got a good model fit? When that lives in a part of the grid, we're darn near anything is going to fit the model reasonably well? Well, not very surprised. But if you get a good model fit in part of the grid, where it's pretty restricted, and there are not a lot of alternative models that would represent the data in that way. Well, then well done. I'm surprised I'm impressed.

Greg 58:48
Absolutely. So there are parts of your model that are contributing more information to your understanding of what's going on parts of your model, that there's less information associated with it. I love that nice tie back, look at you see, I

Patrick 59:00
was paying attention. There are no birds on the feeder, so I was actually following what you were saying.

Greg 59:06
So the purpose of this episode, like a lot of the episodes that we do, is to let you know that things are actually understandable. You know, if you go read an article on some particular state of the art topic, you go, Oh, my gosh, there's just so much there. But everything that we do from the most complex models in this field, that field, that field had to build up from something and all of them were just these

little incremental steps. The stuff that we're talking about here at the end is just a build up from how can we characterize the information in something as simple as the behavior of a coin? In fact, how can we use the metric of the coin this binary thing to characterize the amount of information in everything? Now in the end, we might not be using Bayes to which would be literally quantifying how much information something has in the metric of how many coins it would take to quantify it, but it's the same type of idea where we have information and things that we do a certain amount of stuff. fries with respect to some outcomes less with respect to others, that applies to entire distributions, not just outcomes. And then we can talk about the divergence between those distributions. Which distributions are closer to others? Which ones are farther apart? And how do we make decisions about which distributions might be better? In other words, which models that generated those distributions might be better for us to choose? And what criteria do we use? So I just wanted folks to be able to get a foundation for some of the information kinds of metrics that we use. And I hope this gave a little bit of insight into where some of those things come from.

Patrick 1:00:37
And it gives a mechanism for you to quantify surprise, we're not contribute to the conversation. Right now we have a metric on which Wow, that was a 3.5 surprise that you contributed meaningfully to this conversation.

Greg 1:00:55
On that note, thanks, everybody, for tolerating us on this. I hope that was informational.

Patrick 1:01:02
And remember quantity, dude, your 411 on what you need to know. Look at you that is made up. That's not real. It's real.

1:01:12
411 on the phone one, I gotta go. Gotta go.

Greg 1:01:17
Thanks very much.

Patrick 1:01:18
Take care everybody.

Greg 1:01:19
Bye,.

Patrick 1:01:22
Thank you so much for listening. You can subscribe to Quantitude on Apple podcast, Spotify, or wherever you get the 411 on what's important in your life. And please leave us a review. You can also follow us on Twitter we are at quantity food pod and check out our webpage at quantity food pod.org for past episodes, playlists shownotes transcripts and other cool stuff. Finally, you can probably fly your colors. You can't see it but I spoke colors with a cue with quantity merch at Red bubble.com Where All

proceeds go to Donors Choose to support low income schools. You have been listening to quantity food, a word starting with Q that will remarkably earn you just 20 points in Scrabble corner today has been brought to you by the Nobel Prize executive committee, who was pleased to announce a slight modification of quantum tuned suggestion with the launch of the new category titled The Nobel Prize for non awesomeness. The inaugural 2022 recipient will be the newly implemented MLB designated hitter rule for all national league teams. The conferral of this prestigious award formally brands this decision that travesty that it truly is. Additional funding is obtained by Patrick's Dungeons and Dragons club, who did indeed kick him out after one evening of play on October 22 1981. Reminding Patrick that you may have thought we were all a bunch of uptight dorks but you're the one who ended up a statistician and by Christie and Annie, who will now be walking to school each morning so they have plenty of time to think about the implications of secretly texting audio messages without talking to their father first, how many times do we have to talk to our children about Stranger danger? This is most definitely not NPR.