GABRIEL SANCHEZ-MARTINEZ: Any questions on Homework 1 before we get started?

AUDIENCE: Yeah.

GABRIEL SANCHEZ-MARTINEZ: OK, fire away.

AUDIENCE: I guess, first, do you think we have like this minimum cycle time, like a theoretical minimum cycle time and then what was actually [INAUDIBLE] cycle time?

GABRIEL SANCHEZ-MARTINEZ: So cycle time, just to review-- it's the time that it takes a bus to-- from the time [AUDIO OUT] for a trip. It goes all the way one way, has to wait at the other end to recover the schedule, comes back, waits to recover, and is ready to begin the next round. So that's a cycle.

AUDIENCE: Since you have [INAUDIBLE] going on, if you had 4.1 buses, then you use a cycle time. Then obviously, you can't do that?

[INTERPOSING VOICES]

GABRIEL SANCHEZ-MARTINEZ: So you would need five buses--

AUDIENCE: Yeah.

GABRIEL SANCHEZ-MARTINEZ: --if that's what you've got. Or you would have to do a trade-off with reliability if that were to happen.

AUDIENCE: I think most of my questions were on this very last couple of questions.

GABRIEL SANCHEZ-MARTINEZ: Yeah.

AUDIENCE: We were aggregating a bunch of data for-- [INAUDIBLE] you did it across both directions and then asked, how does it change when you would like to evaluate each direction separately in layover time?

GABRIEL SANCHEZ-MARTINEZ: This is the penultimate question, correct?

AUDIENCE: Yeah.

GABRIEL SANCHEZ-MARTINEZ: So that's the hardest question on the assignment.

AUDIENCE: OK.

GABRIEL SANCHEZ-MARTINEZ: It is a challenge question because there are different cases that you have to analyze. That's maybe the hint, right? There are some cases. And for each case, there is a probability that that case will occur.

AUDIENCE: Yeah.

GABRIEL SANCHEZ-MARTINEZ: And-- let's see if this starts-- there's a probability that it will occur and then a consequence, or something happens in that case. So you have to look at each case and then aggregate the cases together, if that make sense.

AUDIENCE: Yes.

GABRIEL SANCHEZ-MARTINEZ: We're taking questions for Assignment 1, which is due on Thursday. Any other questions?

AUDIENCE: That's it.

AUDIENCE: [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: It is due at 4:00 so at class time essentially, yeah. I actually [AUDIO OUT] if you 4:00. I said 4:05, so you have five minutes.

AUDIENCE: Can you [INAUDIBLE] what assumptions there are [INAUDIBLE]?

GABRIEL SANCHEZ-MARTINEZ: In what question?

AUDIENCE: When you said it seems to be the reasoning or assumption about the schedule [INAUDIBLE]? Which metric do you use? Based on the data, which [INAUDIBLE]?

GABRIEL SANCHEZ-MARTINEZ: Yeah, so that's Question 3, correct?

AUDIENCE: Yeah.

GABRIEL SANCHEZ-MARTINEZ: So I can't really explain. I can't give you the answer to the question. So what I'm looking for there is your intuition and your understanding of why you would pick which statistics from Question 2, where it tells you calculate all these things. Now I'm saying pick from those statistics what you would use for t and for r. And you may want to combine different statistics for the computation of r. Yeah?

AUDIENCE: [INAUDIBLE] multiple valid responses but--

GABRIEL SANCHEZ-MARTINEZ: Yes, some more valid than others, but some that are definitely invalid and some that are almost 100% valid but not 100% valid. So there are several correct answers, and some that are very good answers because you can justify the choice of the statistic conceptually. Yeah. Any other questions on Homework 1? I can take some more questions after class, if that's OK. So we had a snow day if you had a good time, and/or at least, you could use it to catch up. So the schedule is a little different now. I've posted an update about that on Stellar (class site).

There's a new syllabus. And we're going to do some [AUDIO OUT] different [AUDIO OUT]. You may remember that we have three introductory classes on topics of [INAUDIBLE]. And

then, we had model characteristics and roles. And then, [AUDIO OUT]. We're going to shuffle a little bit. [AUDIO OUT] Microphone working? So because the second assignment is on data collection, we're going to cover that today. And we're going to give you that homework today, so that you can get started on the data collection side.

Then, we're going to cover some of the short-range [INAUDIBLE] of planning concepts. Nema is going to do that-- Nema Nassir. You might recall him from the previous lecture. And then, we'll finish with [INAUDIBLE] and costs in March the 2nd, OK? So remember, there's no class on Monday the 21st.

AUDIENCE: You mean Tuesday?

GABRIEL SANCHEZ-MARTINEZ: Sorry, yes, Tuesday. I think, there's no class on Monday. And then, Tuesday there are classes. But it's Monday's schedule. So we don't have class. Thank you for bringing that up. OK. I'll leave Homework 2 for when we finish with the lecture. But I'll distribute it later. So let's just get started on that. So data collection techniques and program design-- that's the topic for today.

Here's the outline. So we're going to cover a summary of current practice quite quickly. Then, we're going to talk about data collection program design process, the needs, the data needs, the techniques for data collection, the sampling. We're going to get into the details of how we get sample slices. And we're going to finish with special considerations for surveys and surveying techniques. so where are we? Where is the transit industry in terms of data collection, and sampling, and these things?

Largely, there's been a transition from manual to automatic data collection. As you might imagine, with the internet of things, and sensors, and the internet, and wireless, it used to be that if you wanted to have statistics on your running times, you had to send people out. We call those people checkers. And those checkers would have notebooks and record running times, and number of people boarding, and these things. Nowadays, with the modern systems, especially the modern systems, we have several sensors and types of sensors that collect some of that data for us. So we're going to cover both approaches.

[INAUDIBLE] data collection to supplement [INAUDIBLE] data collection. And if you happen to be consulting for a developing country that is working with a system that has not yet brought in automatic data collection technologies, it's also useful to know all about the manual design and manual data collection process. [AUDIO OUT] took this class and ended up working in large

consulting firms have gone off to help countries put in new transit systems.

And one of the first things they have to do is back to these slides and see what the plan is going to be, and how many people you need, and how much it's going to cost. So very useful topic. So as I said, there's automatic data collection. There's manual data collection. There's sometimes a mix of data collection techniques. Often, what happens is that we just send people out and collect data. Or we just extract a sample of automatically collected data.

And we don't really think about sampling, and the confidence interval, and how sure are we of that result that we're going to influence policy or make decisions that will affect service. How sure are we of those? So statistical validity. Often, there's an efficient use of data. And ADCS, which is Automatic Data Collection Systems-- we'll use that abbreviation throughout the course- presents a major opportunity for strengthening data to support decision making. We'll talk about how that happens. Let's first compare manual and automatic data collection.

So what happens with manual data collection? You hire people, as I said. You hired checkers. So initially, there's no setup cost. There's a low capital cost to that. But there's a high marginal cost because if you want to collect more data, you have to hire more people. Does that make sense? If you want to bring in an automatic data collection system, you might have to retrofit all your buses with AVL sensors. And that's going to cost you initially. So that's a high capital cost relatively. But low marginal cost-- once you have those systems in place, they keep collecting data for you. And it's almost free.

You do need some maintenance on these equipments. But comparing to manual data collection, you have low marginal cost. Because of that marginal cost difference, it tends to happen that when you have manual data collection, you only pay checkers for small sample sizes-- just what you need. Whereas, once you put in automatic data collection systems, they keep collecting data. So you get much bigger data. Bless you. OK, in both cases, we can collect data and analyze it for aggregate analysis and disaggregate analysis.

So you might want passenger-specific data on things. Or you might want things like just averages and aggregate things, total number of passengers using the system. And when you're doing manual data collection, you can look at quantitative things, things you can measure and count. Or you can also observe things qualitatively. One example that I saw in a recent paper was considering the [? therivation ?] by student in some country. And they didn't ask people if they were students. They were looking at people's-- more or less, are they

young? Are they carrying a backpack? And that would be the labeling for your student.

So that's something that a sensor might not do so well. Although now with machine learning, who knows? But we haven't seen that so. So you can do qualitative observations when you're doing manual data collection. Manual data collection tends to be unreliable, especially when people aren't very well trained and when you have a group of different people collecting data. So each person might have different biases. It's hard to reproduce the exact bias across persons. With automatic data collection, you do the errors. And often, they are not corrected.

But if you do correct them, and you estimate those biases just for them, you can end up with a better result. Because of the small sample sizes in manual data collection, you tend to have to have limited spatial and temporal coverage of data. So for example, if you're interested in ridership in the system, it's unlikely that you will cover ridership in holidays for [INAUDIBLE] system because there are only a few holidays. And usually, you're not mostly interested in holidays. So chances are, you won't have data collection for holidays.

Whereas once you install automatic data collection systems, they keep collecting data. So you get data at midnight on President's Day. So they're always on. They're always collecting data. Manual data needs to be checked, cleaned, analyzed, coded, and sometimes put into systems before they can be analyzed. That could take a while. You need to hire people to do that. Whereas automatic data collection systems often send their data to databases in real-time or very close to real-time. [INAUDIBLE] you can start analyzing things the next day.

So you arrive in the morning to your desk at a transit agency, and you have performance metrics for yesterday. So you wouldn't be able to do that unless you have people working very hard if you're using manual data collection system. When we talk about automatic data collection systems, there are many. But there are three types that we refer to very, very often. And so the first one in AFC, Automatic Fare Collection Systems. This is your fare box or your fare gates in your smart card, your Charlie Card. You're in Boston. You tap to enter the bus. And you tap to enter the subway system.

Increasingly, it's based on contactless smart cards. And those contactless smart cards have some sort of RFID technology with a unique identifier. When you tap that card to the sensor, the sensor will read that identifier. And it'll do things like fare calculation for you. But that record gets sent to a database. And it's there for people like us to analyze and make good use of it for planning. So it tends to provide entry information almost always. In some systems, like

the Washington, DC metro or the TFL subway, you tap in to enter and exit. So you have both origin and destinations.

And if you always have the systems on, then you have full spatial and temporal coverage of all of the use of the system at an individual passenger level. So very disaggregate-- sorry about that. Traditionally, these systems are not real-time. So it might take a while for those transactions to make it to the data warehouse, where they're available for planners to analyze it. The calculation of how much fare in some systems is in real-time. In other systems like the Charlie Card, the stored value that you have is stored on your card.

So it may take a while if you tap at a bus for that bus to go to a garage and get probed-- and for the data that has been stored in that bus to be extracted from that bus to the central server. There is a move-- and we'll talk more about this when we get to fare policy and technology-- towards using mobile phone payments and using contactless bank card payment systems. And those systems often do the full transaction over the air in real-time. So we're starting to look at the possibility of having all this data in real-time or almost in real-time. But it's not there yet.

AUDIENCE: [INAUDIBLE] can I ask a question about that?

GABRIEL SANCHEZ-MARTINEZ: Yeah, of course.

AUDIENCE: In terms of smart card, where this balance is stored on the card--

GABRIEL SANCHEZ-MARTINEZ: Yeah.

AUDIENCE: --if one can figure out how to hack that card--

GABRIEL SANCHEZ-MARTINEZ: Yeah.

AUDIENCE: --then what can [INAUDIBLE] fares through an elaborate technology that I couldn't do and most people couldn't do. But maybe some could.

GABRIEL SANCHEZ-MARTINEZ: Yeah, definitely. So the Charlie Card system is an example about-- actually, MIT students were the first to hack it.

AUDIENCE: I'm not surprised.

GABRIEL SANCHEZ-MARTINEZ: So it's older technology. It used a low-bit encryption key. That's a symmetric encryption key. And they just brute forced it. They figured what the key was. They happened to use the same key for every card. So once you broke that key, you could take any card. And with the right hardware, you could add however much value you want to that card. And--

AUDIENCE: [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: Yeah, yeah, exactly. We don't think it's been a major problem.

AUDIENCE: But it happens.

GABRIEL SANCHEZ-MARTINEZ: I haven't seen MIT students selling special MIT cards. But that would be criminal, of course. Yeah, so newer systems have much stronger encryption. And they have different encryption keys for each card. And certainly, when we're moving towards contactless bank cards, we're talking about a much more secure encryption. It's your credit card that you're using to tap or your Android or Apple Pay.

AUDIENCE: Account based [INAUDIBLE].

GABRIEL SANCHEZ-MARTINEZ: Account based-- and essentially, what you have is a token with an ID. And then, the balance is not even stored on your card. The account server is handling the balance and those things. So much more difficult to break. Yup. OK, AVL systems, or Automatic Vehicle Location systems-- so these are systems that track vehicle movement. So for bus, they tend to be based on GPS. You have GPS on a bus, on the top of the bus, a little hub. And it collects data every five seconds or every 10 seconds.

And these positions might get sent either in real-time, or maybe they get stored on the onboard computer and then are extracted when the bus reaches the garage. So just GPS-- sophisticated AVL systems for bus also have gyroscopes to do inertial navigation and dead reckoning, especially when the GPS precision drops. And that happens especially with the

urban canyon effect. If you have tall buildings, GPS signal bounces around. The dilution of precision messes up the position of the bus.

Or maybe you're entering a tunnel, and you want to continue to get updates of positions inside the tunnel. So this is a temporary system that kicks in and interpolates positions and figures out how the bus is moving. For a train, it's usually based on track circuits. So we're going to talk more about track circuits. But essentially, a track knows if a train is occupying that segment or not occupying that segment. And there are often some sensors that read with RFID technology the ID number of a car. And sometimes, you have a sensor in the front of each car and [AUDIO OUT] each car.

And so a computer will look up the sequence of readings and follow track circuits as they are being occupied and unoccupied-- and in that manner, track trains throughout the system. These systems were put in place mostly for safety to prevent train crashes. And because of that, you would need it to know buses or where a train was. They are available in real-time. They were designed from the beginning to track vehicles in real-time. So that's what we have. I guess what's newer is that now, we're collecting them and keeping them in a data warehouse so that we can analyze running times.

AUDIENCE: [INAUDIBLE] these systems have benefit to the consumer?

GABRIEL SANCHEZ-MARTINEZ: They do. And that's the newest thing that has happened-- that nobody thought about consumers when they were put in place. So yeah, we are talking about tracking, knowing how many minutes I have to wait for my bus, for example. And those things are pushed through a public API, so that if I'm a smartphone app developer, I can go ahead and pull data from this next bus app and make an app. And so people can download it, and they know how many minutes they have to wait. Yeah, so definitely. So we have seen a lot of AVL being pushed in that manner. We have not seen so much AFC data or APC data being pushed.

Obviously, you wouldn't want all the details of AFC being pushed. But you might want to know how crowded is my next bus, or how crowded is my next train. And you might actually alter your decision whether to wait for a crowded train or walk a longer time based on that information. So that's coming. I think, in the next few years, that's going to start happening. So passenger counting-- many different technologies exist. For bus, we tend to have these optical sensors in the back. You might see them if you pay attention-- broken beam sensors. They look like two little eyes with two little mirrors on each door.

And so when you cross the beams, if you press one beam first and then the other, that sensor will know-- is a person coming into the bus? Or is a person exiting the bus? And you have that at each door. And it counts those beams going in and going out. And often, this is slightly inaccurate. So you might get more boardings and lightings for a given trip. So at the end of a trip, whatever remains in terms of imbalance between boardings and lightings gets zeroed out. And the area is distributed throughout that trip that was just run.

And often, you still have to do some error correction after that. But it's a way of counting people getting on and off. And that's useful to get how many people are riding the system and also the passenger miles-- the passengers multiplied by distance, which is often a required reporting element in things like the NTB, the National Transit Database. So for rail systems, we have gates that count how many times they open and how many times they close. So you might have that kind of counting in rail.

You also have video-based counting-- so camera feeds that can be hooked up to a system that will essentially track nodes moving inside that frame. And you can count things that cross a certain line, for example. And you could do that to count flows. And then for train, we also have the weight systems. So this is only in trains. The braking systems in trains apply braking force in proportion to the load on each car. So if you have a very heavy car, you need to apply stronger braking force than in a car that is almost empty.

If you don't do that, then you apply a lot more force per weight on the lighter car. That car is going to be the one pushing the other cars or pulling the other cars through the coupling. And that will eventually break the [INAUDIBLE] at a faster rate. So what you want is, each car to slow down at the same rate by itself as much as possible. And for that, you need to brake in proportion to the weight. And therefore, you have these weight systems. They used to just do that.

And more recently, we hooked them up to a little storage device that keeps track of the weight and maybe Wi-Fi, so that each time it reaches a station or the terminal, it sends the data off. And we might have a rather somewhat [? unprecise ?] idea of how many people are in the car just based on an average weight of a person. And these are traditionally not available in real real-time. [INAUDIBLE] you have questions? Yeah?

AUDIENCE: You could also just reconcile it with the other system, right?

GABRIEL SANCHEZ-MARTINEZ: Of course, yeah.

AUDIENCE: So if you have--

[INTERPOSING VOICES]

GABRIEL SANCHEZ-MARTINEZ: Yeah.

AUDIENCE: --people early can transport to get on to.

GABRIEL SANCHEZ-MARTINEZ: Yeah.

AUDIENCE: [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: Yeah, definitely. Yeah. And that's cutting edge research that's happening right now. How do you do data fiction and merge different systems? They all have errors. And how do you detect when one is more erroneous than the other? And how do you mix these data sources to get the most precise, not just loads, but paths within a network and things like that. Yeah. So any questions on these three very important automatic data collection systems?

AUDIENCE: [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: Yup.

AUDIENCE: So if there [INAUDIBLE] AVL, what kind of reason can be [INAUDIBLE]?

GABRIEL SANCHEZ-MARTINEZ: So the question is, why might some of these technologies produce errors? And in particular, you're asking about AVL. So each of these has a different behavior. And within each of these categories of technologies, each vendor's system might have specific things that happen. With AVL, the most common thing is end of root problems-- detecting when a trip actually begins and ends. So AVL systems, you have this GPS coming in every five seconds. Depending on

your chip set, you might get it more frequently than that. But you also actually sometimes hook it to the doors.

So if the door is opening, you say, well, I must be at a stop. And therefore, let me find which one is closest. So there are ways to correct it. But when you get to the end of the route, it's not clear always-- have you finished your trip? Or rather, are you starting your trip already? So maybe if the terminal is at the same place on the trip-- the previous trip ends at the same place that the next trip begins, there might be a time where the doors open and close various times. And the trip isn't ready to leave yet. And so you really have to wait to see the bus leaving that terminal and moving.

Sometimes, there are false starts. So maybe another bus comes along, and it needs that space. So the driver moves the bus a few meters forward. And the system thinks my trip has started. And then, when you're looking at aggregate data, you're looking at, say, running times at the trip level. You see these outliers with very long times. And if you were to plot them by stop, you see that the link between the first stop and the second step is sometimes very high, 15 minutes.

And so you can throw those out. Or you can do some interpolation or imputation of data. Some systems that care very much about that will purposely place the terminal stops sufficiently far apart to prevent that from happening because it is a problem. And this data is crucial to planning service and figuring out how much resource you're going to put into each route. So yup.

AUDIENCE: For tap cards, [INAUDIBLE] and metros, some of them we have to tap out to exit. It is because of variable [INAUDIBLE].

GABRIEL SANCHEZ-MARTINEZ: Yes.

AUDIENCE: But in some systems, it's still a flat fare. You still have to tap out. Is the reason behind that mostly data collection? Or is there anything [INAUDIBLE] you're going to still have to tap out [INAUDIBLE]?

GABRIEL SANCHEZ- So yeah, no examples of it come to mind. You might know one.

MARTINEZ:

AUDIENCE: MARTA?

GABRIEL SANCHEZ-MARTINEZ: OK, I haven't visited. So yeah, data collection might be a reason to do that. But I'll have to get back to you on why MARTA did that. But yeah, most systems that have controls in and out are for fare policy reasons and not for data collection reasons. We're starting to see more interest in data collection and in investing on these technologies just for data collection. So maybe-- but I'll have to check and get back to you.

AUDIENCE: You mentioned some systems separate their depots to not confuse the end [? from the start point. ?]

[INTERPOSING VOICES]

GABRIEL SANCHEZ-MARTINEZ: Their terminal stops, yeah.

AUDIENCE: What are some examples of those?

GABRIEL SANCHEZ-MARTINEZ: TFL will do that in London, yeah. Yeah, so they'll monitor this. And if they see that this is occurring often, they will separate the stops a bit. And the reason they do that is because they have people whose job it is to impute data when it's incorrect. So if they don't do that, and the system is consistently producing bad data, then that means they're going to have to spend human resources on correcting that data. So at some point, it's just easier to move the stop a little bit. It doesn't have to be a long distance.

AUDIENCE: Got it.

GABRIEL SANCHEZ-MARTINEZ: It does not make the same and make it far enough apart that the geo fences can be told apart from each other. Alright?

AUDIENCE: Really small scale data of the EZRide who I work for, actually you could see real-time bus loads [INAUDIBLE]--

GABRIEL SANCHEZ- Oh, interesting.

MARTINEZ:

AUDIENCE: --which was actually helpful if you're dispatching, and you know a bus is getting through people on it. [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: Yeah, for real-time control.

[INTERPOSING VOICES]

AUDIENCE: But the terminal at our station had a drop-off point and a pick-up point. The drop-off point was before layover [INAUDIBLE] was after for this exact reason to make sure that it will go through the drop-off point, reset, until people get off of it.

GABRIEL SANCHEZ-MARTINEZ: Yeah. Yeah, so it happens.

[INTERPOSING VOICES]

AUDIENCE: Definitely. [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: That sounds about right. OK, if there are no more questions on the three very important categories of automated data collection systems, let's talk a little bit about the data collection program design process. So this comes from before automatic data collection. And nowadays, we think a little bit less about this. But it's still important. So if you do need to collect some data, there's a structure that you can follow to do it properly and to make sure that you collect data efficiently, so that you don't spend too much resources on data collection and that you can answer your policy or your planning questions.

So based on your needs and the properties of your agency, I say here, determine property characteristics. That's a North American term. A property is an agency. So if you see that, that's an agency. So based on the characteristics of the service you're running and your data needs, you can select some data collection technique. We'll get into what some of these are. Then, you can develop route-by-route sampling plans based on how variable the data is in each case.

And you can determine how many checkers do I need. A checker is a person who goes out and collects data. And then from that, the cost-- so human resources. It's a planning exercise. And what we do usually is that we conduct a baseline phase. So that's the first time you go out and collect data. You don't know much about what you're wanting to collect data on. So it might be only matrices, or loads, the people getting on and off. So you have to go out and do a bigger effort. And that's called the baseline phase effort. Once you've done that and you've established some tendencies, you might want to monitor that to see if it changes.

So then, you do a lighter weight data collection effort, where you go out and less frequently, using fewer resources, you collect sometimes the same thing. Or sometimes, you observe something else that is related or can be correlated with what you really want. And then based on a relationship between the two, you can estimate what you really want. So you can monitor what you collected. And then, if you detect that there's been a trend or a change, and you need to investigate it further, you might go ahead and repeat the baseline phase to increase your accuracy.

So one of the catches of this is that to determine sampling plans, to determine required sample sizes to achieve some confidence interval, you need to know how variable your data is. And if you haven't collected it yet, you don't know. So you might have some default values that you resort to. And we'll get to that later in this lecture. But you might also do a pre-test, where you send some people out, and you collect some data to really start to get a sense of how variable is it, and how big will my sample requirements be, and how much will it cost for me to do this. So this is the process that you might follow.

And there are different data needs by the question that you're trying to answer. So one way of looking at that is, are you collecting things that are for specific routes, or for specific route segments, or at the stop level? Or are you using more aggregate system level data collection? Are your questions more system level? So system-level things are more about reporting, and they might be tied to things like federal funding. Whereas route-level things and stop-level things are more important for planning.

So when we talk about route and route segment level, we're looking at things like loads at the peak load points or at some other key points. How many people are in the bus? The running time is by the segment to do schedule that has time points or maybe end-to-end to your operations plan. Schedule adherence-- are these buses running on time? Or are my schedules not realistic? Total boardings or revenue, two things that are highly correlated-- so

number of passenger trips.

Boardings by fare category-- so you might say, well, I want boardings, but I want to know how many seniors are using this, and how many students are using this, and how many people are using monthly passes, and how many people are using pay-per-ride. So you have different fare categories. And you might want to segregate the data by that. You might want passenger boarding and lighting by stop. So that's what APC would give you if you have an automated system. But you might also use a write checker, who sits on the bus and counts people boarding in a lighting.

Transfer rates between routes-- to see you maybe you're looking at changing service so that people don't have to transfer. Passenger characteristics and attitudes-- this usually requires some degree of survey, where you ask people things, passenger travel patterns. At the system level, we have things like unlinked passenger trips, passenger miles, linked passenger trips. This had the whole system level. So sometimes, you do route level or route segment level analysis, and then, you aggregate to get the system-level things. That's usually how you proceed.

But the requirements in terms of how many of these you have to sample might be different. So if you want to achieve a certain accuracy at the system level, you don't need to achieve the accuracy for each of the routes that are in that system because you might have-- so if you want to say 90% confidence in some system-level data element, you might only need 80% or 70% of the element level. And once you bring those altogether, you achieve the 90% that you need.

So data inference, I talked about how sometimes we can infer items if we don't observe them directly. So from AFC with AFC is a low-fare collection system, we have boardings because people are tapping into the bus or tapping into the subway system. And if we have APC, we count people getting on. So we can look at total number of boardings that way, if that makes sense.

That's pretty direct. Sometimes, you want to correct for errors in the APC system, or you might have things like variation affecting that number-- like it goes from AFC to how many people were actually in that bus. How many people actually boarded? So you might do a little bit of manual surveys to check what that relationship is and apply some correction.

For passenger miles, we need to know how many people are at the bus between each stop

here. So AFC gives you boardings and only boardings. APC gives you ons and offs. If every bus had APC, then you could calculate passenger miles directly. But often, you have systems where only a portion of the fleet has APC. So maybe 15% of your fleet is equipped with APC. And from that, you get the sample OD matrix. And you can use that OD matrix to convert from boardings only to the distribution and the ons and offs at all bus routes. And from that, you can get passenger miles.

Or you might just use your buses that have APC, if that suffices for your data collection unit. Same thing with peak point load-- similar idea. The AFC only measures boardings. So it doesn't give you the peak point load automatically. But from APC, you could get it. And it you can establish a relationship between boardings and the peak load point, then you can use that model to infer the peak load point from just boardings. So this is a key thing to be efficient about data collection. Any questions on this idea? Yup.

AUDIENCE: So to get passenger miles, you're also going to have a GPS system as well to know the distance? Or are we just basically [INAUDIBLE] this is the routing [INAUDIBLE]?

GABRIEL SANCHEZ-MARTINEZ: Both.

AUDIENCE: [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: Yeah, both.

AUDIENCE: [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: What tends to happen is that the APC, it'll come in. And it'll say, at this stop, this many people boarded. This many people are lighted. So you have other layers in your database that say where the buses and what the distance is between stops and the stop pair level. So you then essentially know how many people are riding on each link and how long that link is, and you multiply the two. So yeah, passenger miles. Yeah, more questions.

AUDIENCE: Yeah, for these checks that are going on like the more manual checks--

GABRIEL Yeah.

SANCHEZ-MARTINEZ:

AUDIENCE: --I know often, there's derivation checkers who are coming into a check.

GABRIEL SANCHEZ-MARTINEZ: That's right, yeah.

AUDIENCE: Do they also use that data to cross-reference the passenger counts? As in, [? this ?] person gets on, and they check everyone's voice to [INAUDIBLE] DFL.

GABRIEL SANCHEZ-MARTINEZ: Yeah.

AUDIENCE: They then know exactly how they go on the bus.

GABRIEL SANCHEZ-MARTINEZ: Yes. Yeah.

AUDIENCE: Do they use that data?

GABRIEL SANCHEZ-MARTINEZ: Yeah, they can. In the APC, sometimes there's reliability problems, especially when vehicles are very full because sometimes, people will block the sensor by the door. Actually, people like to stand by the door all the time, even when the bus isn't full. And that kind of affects APC. You might notice this on the one. If you take the one-- so yeah, you sometimes have a little bit of a manual effort to figure out. Just learn about your APC system, and what are the errors, and when do you see them. It often happens that you have more variation when you have very high loads. And that's when APC is least accurate. So it all comes together. Yeah. Questions on the back? I think I saw a question. No?

AUDIENCE: Yeah, I noticed that in Chicago, when the bus would be crowded, then people get off the bus. They let people off--

GABRIEL SANCHEZ- That's right.

MARTINEZ:

AUDIENCE: --and then back on.

GABRIEL SANCHEZ-MARTINEZ: Yeah. Yeah. These double things. But somebody might be by the door just blocking the two little sensors--

[INTERPOSING VOICES]

GABRIEL SANCHEZ-MARTINEZ: --the two little eyes. And that's it, no records of people getting on or off. So if you're doing a little data collection, as I said, we use checkers. And actually, your second assignment, you will be checkers of some kind. The typical checkers which you won't be in this assignment are ride checkers and point checkers. So a ride checker sits in the vehicle and rides with the vehicle. And the typical thing that these ride checkers are looking at is, how long did it take to cover some distance? So what was the running time for that trip? And also, people getting on and off-- so they act as APC essentially. And they act as AVL.

So AVL and APC together might replace most of the functionality of a ride checker. Although a ride checker often can conduct an onboard survey, asking passengers about where are they going, or their trip purpose, or things related to social demographics, which are qualitative and cannot be collected with the sensors. Point checkers stand outside of the vehicle. They stay at a specific place, and they can look at headways between buses-- so how long did it take between each bus to come by, and how loaded were these buses?

So if you're interested in the peak load point, and you know where the peak load point is, and you just want to observe, measure what are the loads of the peak load point, then you can just station a point checker at the peak load point. And if that person is strained, we'll be able to more or less say how many people are in the vehicle from looking at the vehicle.

With automated data collection systems-- yeah, with a fair system, we have passenger accounts. We have transaction data, which is very rich. It will tell you not only that somebody is entering or exiting, but also how much they're paying, sometimes information about the fare product type, which might help you infer if this person is a senior, or a student, or a frequent user, an infrequent user-- so many things that are very useful for planning. And we'll get to play with some of these later in the course. And then, there's Automatic Passenger Counters, APC.

So as more and motor systems switch to automatic data collection, we still use some manual data collection, but not in the traditional sense. Now, we reserve those resources for things like surveys about social demographics and other things. And we also carry out web-based surveys, which would have some biases. But if people registered their cards, and you have email accounts, you can maybe send a mass email to everyone and carry out surveys. The MBTA does that. Maybe some of you are in the panel of people who are e-mailed every now and then. Is anybody in that panel? No hands. I'm in that panel.

But I know somebody must be. So yeah, they send an email, and they ask about your last ride. And they say, where did you start from? What were you doing this trip for? How long did you have to walk? Are you happy with the system? Was your bus on time? Yeah, things like that-- how satisfied are you? It's a survey with qualitative questions that you couldn't collect automatically. It's [INAUDIBLE] seeing things about your experience outside of the bus, which there are no sensors for.

All right, sampling strategies-- a bunch of different ones and the simplest one is called simple random sampling-- very, very simple. So when you have sample random sampling, what happens is that every trip, if you're looking at surveying trips, for things like how many people boarded this trip-- let's take that as an example. Then, if you're using simple random sampling, every trip has equal likelihood of being picked and being surveyed. So if you go through your process, and you determine that you need to observe 100 trips to get an average reliably. And you're going to use that to plan something, then you need to look at 100 trips.

So if you use simple random sampling, you take your schedule, and you randomly pick 100 trips. And that's your sample. Those are the ones that you send people out to collect data. Now, there's a little bit of a problem with that. It's not the most efficient method because if you're going to send someone out, and that person is going to be active, and require some time to get to the site and some time to return, then once they're out there, you want them to collect as much as they can. So that's not simple random sampling. That's cluster sampling.

Before we get to that systematic sampling-- so typically, instead of picking randomly, we say, OK, we need to get 10% of the trips. So let's just make it such that we count. And maybe it's every five trips, we have to survey it. So now, it's evenly spaced. And this is useful for some things. One example is weekday, picking the weekday that you're going to survey on. So the technique that is often used is sample every six days. Why would that be? Yeah. So if you do it every seven, then you always have a Monday. And that's going to get some bias if Mondays

happen to be low ridership days or high ridership days.

So if do every sixth day over a year, you have a good sample of every week day. So that's an example of systematic sampling. But you still have that issue of it might not be the most efficient. Cluster sampling, sometimes it's more efficient once you send out a person to collect data to do as much as possible. And you survey a cluster. So one example is, if you're distributing surveys to passengers, and you need to distribute 100 surveys. If you do 100 simple random sample, then those people might be in different parts of the system. And one might be the first person you see getting off at South Station.

And then another one by me might be the first person you see getting off at the Kendall station. So that's very inefficient. So a cluster might be everybody on board a bus, and that will get a bunch of people together. However, it's not as efficient statistically to do that. So you can't just add up to 100, and you're done because there might be some correlation within the people riding that vehicle that they will tend to answer in a similar way. So you might need to increase your sample size when you use this technique. But still, you might have a more efficient sampling plan.

Then, there is the ratio estimation and conversion factors. We gave examples of this already. This is in the context of baseline phase and then monitoring phase. So you start out with a baseline phase. And in the baseline phase, you collect the thing you really want and something that is very easily collected with lower resources. And you make a model of the thing you really want as a function of the thing that is cheap and easy to collect. And then, on the monitoring phase, you only measure the thing that is cheap, and easy, and quick. And you then use the model to estimate what you really want.

So converting AFC boarding to passenger miles, we give an example of that. We're converting loads at checkpoints to load somewhere else. So maybe only measure loads with a point checker at the peak load point. And you have some relationship to convert those loads to loads at other key transfer stations as an example. And then, the stratified sampling-- so one of the things that determines how big of a sample you need is the variability in the data that you're collecting.

So correlation, when you're looking at a whole system with multiple routes or multiple segments-- maybe when you look at one route, there's some variability of running times. But they have a central tendency as well. And when you've got a second route, you have also

some variability and a different central tendency. So you bunch all the data together, some of the variability across data points in our data set are going to be the inherent variability of each route. And some of it will be systematic-- the differences between both routes.

So if you do a simple random sample, and you don't separate the systematic variability from the inherent variability, then you're going to get a wider variability. And you will require a bigger sample size. Stratified sampling is an approach where you determine sample sizes for each of these separately. And it's more efficient if you do it well because you eliminate the need, or you at least reduce the need, to collect data for the sake of the systematic differences between different parts of the system. Any questions on these methods? Yes.

AUDIENCE: [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: Yeah, so let's maybe pick another example. Let's say that you're looking at the proportion of passengers in a bus who are students. And you're distributing a survey. And they tell you whether they're students or not. And you want this for the whole system or for at least a group of routes. And it tends to be that some routes don't serve universities and don't serve schools. So they have a lower proportion of people. And then, some routes that do go through universities, and they have a higher proportion of students.

So if you just want the system-wide proportion of people who are students, and you join all these data points together, there's going to be a lot of variability in what proportion that is across every trip that you survey, correct? So in some sense, it will indicate that because of that variability, you're going to need a higher sampling size. You're going to have to survey more trips to get at your desired accuracy level and tolerance. But now, if you say no, I'm going to split routes in two, into two stratas. One is the routes that serve the universities. And these tend to have around 50% proportion.

And then, there's the routes that don't serve universities. And these tend to have proportions near 0. So if you're in your 0, you might require a lower sample size to cover those. And you can just very efficiently cover most of your bus routes that way. And then, focus your efforts on just the ones that have higher proportion. And you achieved your system-level tolerance requirements with much fewer, with by far fewer resources required to collect the data. Does that answer your question? Yeah.

AUDIENCE: [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: So what he meant by inherent is that within each bus route or within each strata, there will be some variability. Even within the trips that are serving universities, every trip might have a different proportion. So there's going to be a little bit of variability in that. But if you mix that with trips that are not serving students, then you pull all that data together. Then, it's going to look like the variance of that data set is much higher. All right, so we've tossed these terms around-- tolerance, confidence, level accuracy. So let's define them more precisely.

Accuracy-- when we talk about accuracy, that has two dimensions. So somebody might say, the average boardings per trip is 33.1. And then, the question that follows is, do you mean exactly 33.1? How certain are you of that? And how accurate is that? So when we talk about tolerance, there's relative tolerance, and there's absolute tolerance. Relative tolerance is expressed in terms of a percent of the amount you were collecting or a fraction. So you might say mean boardings per trip is 33.1, plus or minus 10%. And that's the 10% of 33.1. That's why it's relative tolerance.

Then, there's absolute tolerance. So mean boarding per trip is 33.1, plus or minus 3.3. Now, in this case, these two are equivalent. 3.3 in absolute terms is 10% of 33.1. But this was expressed in absolute terms, and the previous one was expressed in relative terms. So don't always assume that if you see a percent, it's relative because if what you're measuring is in itself a percent, unless you're using a percent of a percent, then it's absolute. So here's an example. Mean percentage of students is 23%, plus or minus 5%. That's absolute because it's 5%, not 5% of 23%.

First, we talked about, is that exactly 33.1? Or is it something different from 33.1? Then, the second question is, how sure are you, how confident are you that the number you give, plus or minus the tolerance you give, is the right answer? So now, you say I'm 95% confident that the mean boardings per trip is 33.1, plus or minus 10%. So now, you combine the tolerance with the confidence level. And that's the full expression of your accuracy. And that's what you need when we look at the data collection.

So you have two different things that you could play with. And what happens typically is that you choose a high confidence level-- 90%, 95 percent are typical. And then, you hold that fixed. And you calculate what level of accuracy you need. Or rather, you decide what level of accuracy you need, depending on the question you want to answer, and the impact it could have on the system. So if you're looking to [INAUDIBLE] something that will have very

significant effects on the service plan or maybe on investment in the system, then you might need a higher accuracy.

But if you're collecting data just for reporting, maybe it doesn't matter as much. And you don't need to spend as much money on data collection. So as an example here, the National Transit Database-- NTD, we call it NTD-- for annual boardings and passenger miles, it says, you should collect data to achieve an accuracy of 10%, relative tolerance at 95% confidence level. You need both. So take home message about this.

The other thing, the t distribution-- so this is a probability distribution that is bell-shaped. It kind of looks like the normal distribution. And it approaches the normal distribution as the sample size gets very large. This is the distribution that arises naturally when you're estimating the mean of a population that is normally distributed with unknown mean and variance and some known sample size. So to the right here, we have your equations that I'm sure you've seen before for sample mean, sample variance.

And I guess, what's important to think about is that the distribution of what you're collecting-- for example, you might be collecting data on a number of people boarding route 1. So that might have some distribution. As you collect more and more data, so as you survey more and more trips, the distribution of how many people board each trip does not necessarily have to be normal.

But it turns out from the Central Limit Theorem and other laws and properties of statistics and probability that the distribution of the estimator-- so the distribution of the mean that you calculate based on that sample that you collected-- is normally distributed as the sample size increases. So if you have a lower sample size, instead of using the normal distribution, use t distribution. Sometimes, we call that a student, the t student distribution. And this distribution gets wider as the variability increases and as the sample size gets smaller. It has a property called degrees of freedom, which is sample size minus 1.

And you can see from this chart right here when you have degrees of freedom equals 1, which means you collected two data points, it's wider than when V approaches infinity. And what you have in black here, the thinnest and least variable of these, is essentially a normal distribution. And this is the distribution not of what you collected. It's not the distribution of the number of people who boarded route 1. It's the distribution of the mean that you estimate.

AUDIENCE: [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: Exactly, it's a sampling distribution of the mean. And if you were to repeat that experiment with the same number of trips but different number of trips, you might get a slightly different mean. So if you were to repeat that many, many times, the distribution of those means would be shaped in this manner.

AUDIENCE: [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: Yeah, well, student t distributed. And as sample size increases to infinity, normally distributed. Harry.

AUDIENCE: So just for V equals 5, I think you [INAUDIBLE].

GABRIEL SANCHEZ-MARTINEZ: 4.

AUDIENCE: 4.

GABRIEL SANCHEZ-MARTINEZ: Sorry, 6. 6.

AUDIENCE: Approximately 5 [INAUDIBLE].

GABRIEL SANCHEZ-MARTINEZ: Yes, 6. Yeah. I mispoke. [INAUDIBLE]

AUDIENCE: When there's a sample variance, sigma x squared equals roughly. Is that not supposed to be an equals? Is that not the way the sample variances define? Because I thought it's the--

GABRIEL SANCHEZ-MARTINEZ: So-- --it's below the variance of distribution. But that's roughly [INAUDIBLE].

AUDIENCE: Yeah, I guess the issue is that you don't know the true mean. So you're using an estimate to calculate the sample variance. And therefore, it's almost, almost the sample variance.

GABRIEL SANCHEZ-MARTINEZ: Right. But I thought--

AUDIENCE: You're using an estimator to do the-- that's what you have to do.

[INTERPOSING VOICES]

AUDIENCE: He's incorporating the fact we're dividing by n minus 1 rather dividing by [INAUDIBLE].

GABRIEL SANCHEZ-MARTINEZ: No, so n minus 1, that has to do with the degrees of freedom issue. And that's to go from population variance to sample variance. But the other thing that happens is that if you're doing the population, then you know exactly what your mean is. It's exact, right?

AUDIENCE: Yeah.

GABRIEL SANCHEZ-MARTINEZ: And then in that case, you would know what the exact variances is as well. Yeah. So the n minus 1 is just to remove a bias that would arise from collecting only a sample.

AUDIENCE: But here for example, you can say this is equals to [INAUDIBLE].

GABRIEL SANCHEZ-MARTINEZ: Yeah, yeah, yeah, yeah.

AUDIENCE: You're working with the sample to know it would be an approximate [INAUDIBLE].

GABRIEL SANCHEZ-MARTINEZ: Yeah, in practice equal 2.

AUDIENCE: As your sample distribution increases, then obviously, your sample increases--

[INTERPOSING VOICES]

GABRIEL SANCHEZ-MARTINEZ: And therefore, this becomes more and more accurate.

AUDIENCE: [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: Exactly.

AUDIENCE: It should be approaching more [INAUDIBLE].

GABRIEL SANCHEZ-MARTINEZ: Yeah, so I guess what's important to realize is that this is an estimate of the population variance, which in itself uses another estimate. And I guess, that's why that's there. But it's a very small detail. I didn't mean to distract you.

AUDIENCE: So for the n, is it the sum of all the different samples of [INAUDIBLE] or is it just--

[INTERPOSING VOICES]

GABRIEL SANCHEZ-MARTINEZ: So you don't ever repeat the experiment like this. This is more of a theoretical explanation to why there is a distribution to the mean, even though you only have one. You only have one mean, right? Because you're going to collect data. And once you finish collecting data, you're going to calculate the mean of all that data. So you only have one mean. If you were hypothetically to repeat that experiment, and you calculated separate means for each one, then you would get a distribution that would look like this. In practice, you would just increase your sample size and still compute one mean, which would be more accurate. Yeah.

OK, let's move on. So tolerance and confidence level-- so we have these distributions. These are the distributions of the statistics, of the mean in this case. They are bell-shaped. As your sample size increases, the degrees of freedom goes up. And your accuracy goes up. And the variance of that statistic distribution decreases. So it gets thinner. So here in red, you have a distribution with a smaller sample, and therefore, less accuracy or less confidence would look like. And then as you increase your sample size, you see that it becomes more peaky.

So when we talk about tolerance, and let's come back to the concept of absolute tolerance in particular, we're talking about the distance between the center of that distribution, which is a symmetrical distribution, and some limit. So we're saying, if you have a tolerance of plus/minus 10. Then, you're going to measure 10, say 10 boardings, from the center to the right and from the center to the left. And that's your absolute tolerance. So when you calculate absolute tolerance, you can express that tolerance as a function of the variance and/or the standard

deviation, rather of your mean.

So instead of saying 10, you could say 2 times the standard deviation of that distribution using the equation that we just calculated. And that's very convenient. Why would we do that? Why would I want to complicate things that way?

AUDIENCE: [? Outside ?] [? of ?] a cumulative

GABRIEL SANCHEZ-MARTINEZ: No, I mean, there's a mathematical convenience here. What is this a function of? It's a function of the standard deviation of the thing you were collecting and your sample size, right? And what do we want to do? We want to determine how many things we need to collect, right? So here we go-- we have n. And now we can solve for n, we have the sample size that we require for a given tolerance. So we're going to decide what the tolerance is and calculate sample size, a minimum required sample size. You can always collect more data.

All right. So again, to review, this is the same equation I had in the last slide. You have absolutely tolerance. You can express that as a multiplier times the standard deviation of the mean. And then you solve for n, and you get this equation right here. t is your tolerance and you can-- oh, sorry. t is the number of standard deviations from the mean. d is your tolerance, which you choose. And this is something that you know, or collect, or approximate.

So these are all given. Where does t come from? Well, we said that we're going to use the t distribution, right? So the t distribution has a table-- or it has a certain shape, rather. And using Excel or looking up at some table, you can figure out what t is for two times the standard deviation from the center.

So you can just plug it in from Excel or from-- it's a property of the distribution, essentially. Once you pick a confidence interval, you know t. If you want to go to 95, it's a certain value. If you want to go to 90, it's a different value.

OK. When we look at relative tolerance, relative tolerance is just absolute tolerance divided by the mean that you are collecting, correct? Because instead of saying plus or minus 10 boardings, we're saying plus or minus 5% of the mean. So we just take absolute tolerance and divide by x bar, the sampling mean, the sample mean. And we solve for n again.

So what we have now, it looks very similar as to the question right here. But now we have the mean and the denominator. OK, this quantity, standard deviation divided by mean, sample

standard deviation divided by sampling mean, is called the coefficient of variation. And there's a convenience to this. And there's actually a reason why sometimes relative tolerance is preferred to absolute tolerance. It's because of this, because there's a mathematically convenient characteristic of property coming out of this-- that you don't need to know the standard deviation of what you're collecting to figure out your sample size.

We're kind of running in circles here, right? We're saying that to determine sample size, you need to know the standard deviation. Well, I haven't collected data. So I don't know how variable the data is. So that's an issue. Now I have to estimate what that is.

It tends to happen that the coefficient of variation is a more stable property than the variation in itself, than the variance or the standard deviation itself. So you're more likely to get away with using default values for the coefficient of variation than you are with assuming a specific standard deviation.

AUDIENCE: It should be noted that it's unitless, coefficient of variation.

GABRIEL SANCHEZ-MARTINEZ: Yes, it is unitless. Thank you. OK. So what happens is that relative tolerances are typically used for averages. So here's an example-- you measured 5720 boardings plus minus 5%.

So if you were to get the absolute equivalent of the absolute tolerance of that. That would be 5% of 5720. That would be 286 passengers. That's a weird thing to report. 5% is more understandable, right? And it kind of makes more sense. So that's what we want naturally, anyway. So as I said, the coefficient variation is typically easier to guess than the mean and the variance separately. So we use that.

Here's an example using the t distribution, where the sample is not large enough to assume a normal distribution. So we say, let's have a relative tolerance of plus minus 5%, a confidence level of 95%, and a coefficient of variation of 0.3. So we start out assuming large sample, and therefore degrees of freedom is infinity. We can use the normal distribution.

If we look at the normal distribution, with plus minus 5%, confidence level 95%, the t is 1.96. So we look that up on a table, or we use Excel norm dist, or-- yeah. t dist for t and norm dist for normal. We got 1.96.

We plug in the relative tolerance, the 0.3-- we get 140. 140 is not quite infinity, right? So if we look at 140 as a sample size, that would imply that all the degrees of freedom is 139. Now we

go back and look at the t dist, and we change 1.96 to the value from the t distribution for that degree of freedoms. And we get 140.73.

So you're sort of seeing that you were almost right. 140 is very large. In practice, you would just round up a little bit and get a nice round number, and you would even play with this once you're looking at planning who you're going to send out and how many hours you're going to collect. You want to get at least 141, but if you're going to have people in units of eight hours, for example, or units of four hours, then you might as well finish the batch for four hours, the last one. Maybe you'll get 150, 160 from that.

Here's an example of that equation with different assumptions of confidence and tolerance. And so we're using 90% confidence, and we're assuming a certain sample size here. So you can see that, as the tolerance decreases, which means that you require a greater accuracy for different coefficients of variation, the sample size can get really large. So if your data is not very variable, then you can sample just a few trips. And you know because they don't vary that much what the mean is. But if there's a lot of variability across strips, then you need more. So that's what you see as you go down the rows on this table.

Here we have tolerance. If you only have to be 50% accurate, plus minus 50%, then you don't have to collect that much data. If you want to be more precise, and you want to say plus minus 5%, then you need a bigger sample size, right? OK.

Proportions-- and the homework, actually, is based on proportions, so this is important. Consider something, a group of passengers, to estimate the proportion of passengers who are students. So from probability, when you are looking at an event that can either be 0 or 1, or black or white-- in this case, students or non-students-- there's a certain probability that that person is a student, right? And what you want to estimate is that probability or, in other words, what percent of the things you observe are students.

So from the properties of the Bernoulli distribution, the variance is p times 1 minus p. So if everybody is a student, or nobody is a student, either way there's no variability, right? So you would have 1 times 1 minus 1, 1 times 0, 0-- no variability. Though at the peak variability, the highest variance of this distribution, is when 50% of your people are students, so 0.5 times 1 minus 0.5, 0.25. That's the highest variance, OK?

So the tolerance is typically specified in absolute terms when you're estimating proportions,

because the proportion is in itself a percent. So you use absolute tolerance. And you just substitute, essentially, this variance. You put in the variance of the Bernoulli distribution, which is p times 1 minus p. And that's how you get the sampling equation, sample size requirement equation.

Here's a problem. We don't know in advance what the proportion will be, right? And we need that to know how many people we need to survey to figure out-- or how many trips we need to survey to figure out-- sorry, how many students we need-- how many riders we need to survey to figure out what the average number of students are. OK, so--

AUDIENCE: And it's also a [INAUDIBLE] p times 1 minus p [INAUDIBLE] is a constrained number.

GABRIEL SANCHEZ-MARTINEZ: It is a constrained number, and that's exactly where we're going. So we use something called absolute equivalent tolerance instead of absolute tolerance. We assume that p is 0.5-- that's the maximum it could be. So let's go ahead with a worst case scenario.

And then what happens with p itself? Well, if your percent is high, then you can tolerate a bigger number, right? So if it's 32%, you're probably OK with plus minus 5%. If your average were 1.2, plus minus 5% is not that good, right? You need a higher-- you need a much stricter, tighter confidence interval for that. So probably not good to do plus minus 5% in that case.

AUDIENCE: [? Well, do ?] [? you mean ?] you have a plus minus 5% absolutely percentage?

GABRIEL SANCHEZ-MARTINEZ: Absolute, yeah.

AUDIENCE: And you'd be going negative [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: Negative, which is possible but difficult to interpret.

AUDIENCE: Sorry, so this isn't actually 32% plus or minus 5% of 32 [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: It is not-- yeah, it's absolute tolerance, not relative tolerance, right. So what's convenient about this is that these two factors work in opposite directions. So as you get bigger, as the proportion gets closer to 50%, the variance increases. So oh, well, we need a bigger sample.

But your tolerance increases as well, so you don't need as big of a sample.

And so it's convenient. And the practical solution is assume p is 0.5 and work in terms of absolute equivalent tolerance. So you pick a tolerance under the assumption that our proportion is 50%.

And here's what happens. Yeah, if the expected proportion is 50%, and you say plus minus 5 percent, what you would get is this 5%, if it turns out that p is 5%. But if it worked more to the extremes, like 5% or 95%, what you would actually achieve from having planned the survey, assuming 50%, is 2.2-- so much better, much more acceptable to say 5% plus minus 2.2%, right? So it works out.

And there's a convenient equation if you assume a very large sample, or large enough sample, and you pick 95%, 0.25, which is the variance times the normal distribution t squared is 0.96, which is almost 1. So then you get this equation. You take 1, you divide it by the tolerance that you want, your equivalent tolerance, and that's your sample size. So it doesn't depend on anything about the data in itself. You just say if I want, on whatever I'm collecting, whatever proportion I'm collecting, a 5% absolute equivalent tolerance, then I need 400 surveys to be answered. Yeah?

AUDIENCE: So this assumes a random--

GABRIEL SANCHEZ-MARTINEZ: Simple random sample.

AUDIENCE: [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: Yes, a simple random sample. So you would increase these numbers if you are using cluster sampling to account for correlation. You would have to increase them if you're giving people a survey, and not all of them answer the survey, because you need 400 surveys answered. So if only half of the people answer the survey, then you need to distribute 800 surveys.

AUDIENCE: Do you recommend calculating also that the standard error after this so that [INAUDIBLE] make sure?

GABRIEL SANCHEZ- Absolutely, yeah. You want to go back and check with the standard error and when your confidence interval is and see if you meet it or if you need to add a few days of data collection.

MARTINEZ:

AUDIENCE: Right.

GABRIEL SANCHEZ-MARTINEZ: Yeah. OK, so with proportions, you need a very large sample size to estimate a proportion if you want accuracy. If you say absolutely equivalent intolerance of 4%, then you need 600. That's a big number, so it just gives you an idea of that. If you get greedy with the tolerance, you have to pay for the surveyors to go out. OK.

So the process is you determine the needed sample size just with the discussion of the equations that we discussed. Then you multiply the sample sizes. If you're using stratified sampling or if you have questions that have multiple variables, you need to then make sure that you achieve that sample size for each combination of things that you're measuring.

So if you're, for example, looking at not just boardings, but proportion of passengers that are car-owning, who are pleased. So you could just independently measure pleased, independently measure passengers who own a car. And you might have the tolerance you need on each one, but if you want the combination of that, now you need a higher sample, because you need that number for the combination of those things.

Then there's a clustering effect, so a typical thing if you're doing the clustering of a whole vehicle of passengers is to multiply by 4. And then for things like OD matrices, the rule of thumb is 20 times the number of cells. What does that mean? That if your OD matrix is quite aggregate, and it's at the segment level-- so say you divide a root into two segments, then your OD matrix has four cells. Four cells times 20, that's how many people you have to survey.

If you do error at the stop level, then you have many more stops and, therefore, many more cells and, therefore, a much higher sample size. If you have a response rate that is not 100%, which is always the case, then you have to expand by 1 minus that in the reciprocal-- sorry, 1 over that in the reciprocal.

And then you get a very large number, and you say I don't have the budget for that. And you have to make tradeoffs and figure out what you can do. And maybe you have to-- maybe you can't collect this combination and know that accurately, right? So you revise your expectations.

OK, with response rates, you are concerned with getting the correct answers. You also want to be getting a high response rate. If you don't get a high response rate, there might be a bias. So you have to worry about that.

If you have low response rates, that means you need to distribute more surveys, and that costs money. And there's the bias that I just mentioned, so people who don't respond may not be responding for a reason. And then done that might bias your results. And that might make you decide something in planning that is not the right decision based on what actually happens.

So we call that the non-response bias. OK, so what happens? People who don't respond might be different or might have responded differently to the question had they responded. So here's some examples. If you're surveying people who are standing, they are less comfortable. And maybe it's a crowded bus-- they are less comfortable. Or maybe they're getting off one of those stops that is coming up, so they are less likely to have the time to respond to your survey.

People with low literacy, teenagers, people who don't speak the language, are less likely to respond. And they might have different travel patterns. So if you understand those things, and you get lower samples for them, you might be able to do some sort of correction to those biases. But you have to pay attention.

How do you improve your response rate? Well you can make your questions shorter. You can do a quick oral survey. That's what we're going to do for this homework. You can try to get information from automatic sources whenever possible. So if you have an AFC system, let's not collect boardings, because we know that. And then of course some training, and just being kind, and having supervision helps a lot.

OK, here's some suggested tolerances for different things. So we're looking here at boardings or the peak load. And you see here that the suggested tolerance is 30%, plus minus 30%, when you have a route with one to three buses. And then as you have more and more buses, the tolerance decreases. That means you require a larger sample.

Why is that? Why do you need a bigger sample if you have a route with more buses?

AUDIENCE: You're less likely to sample a different bus.

GABRIEL SANCHEZ-MARTINEZ: Yes, and when you have higher-- when you have more buses, you tend to have higher frequency. There's bunching.

OK, so if you then survey loads, for example, and you only get a few because of the bunching effect and because there are more buses, and you're observing a smaller percentage of them for a given time period, say, you're less likely to have observed the bus that was really crowded, right? So that means that you need to decrease your tolerance. And therefore, it's more expensive to survey that. OK, good.

Trip time-- 10% for routes with less than 20 minutes, 5% with routes of greater than 20 minutes. Similar concept if you have greater than 20 minutes-- there can be just more variability, and you really want to get that right. When you have less than 20 minutes, your decision on cycle times and things like this are not going to have as much impact on the fleet size that you require. As you get bigger running times, a small percentage change in the mean could influence how many buses you need to dedicate to that and the cost of running that service.

On-time performance-- 10% absolute equivalent tolerance. These are typical values-- don't take them as gospel, please. And these are for reporting, not for anything that's very critical for operations. Some of them are. Yeah, 30% at least, I would say, is for reporting. I wouldn't make any critical decisions with 30%.

On-time performance-- we're talking here about whether a trip is on time or not on time-- so Bernoulli trials, right? And there's a proportion of trips that are on time, and what we do is that, we essentially say plus-- if we say plus minus 10%, then we're saying that the sample size should be 1 over 0.1. Yeah.

All right, default coefficient-- these are default values for coefficient of variation of key data items. Ideally, you have your own data that you look at, and you don't resort to this. But if you ever find yourself in a situation where you need to start out with something. Here are some based on studies that previous [AUDIO OUT] They took different routes and looked at loads and running times for different time periods and found what the coefficients of variations were. And here they are on a table for you to use.

In the interest of time, since I want to discuss the homework, I'm going to stop here with slide 25. And I'm going to not cover the whole process, which includes the monitoring phase. And in this slide here, we have how you establish conversion factor. The conversion factor in itself has a variance. So there's some uncertainty about the relationship that you estimate between your baseline data item and your auxiliary data item. So you need to consider that in your

sample size. And here are some tables with some examples of what happens when you require different-- well, when you're variability of or your coefficient of variation of your relationship increases or decreases.

OK, let's look at the homework. I really want to use these last five minutes for that. So please take one and pass. OK, so the MBTA, there's a proposal here in Boston of taking Route 70 and 70A-- they run through Waltham, and they go into around Central Square. And some people are saying those two routes should be extended to Kendall Square, because a lot of people are actually going to MIT, or Kendall Square, or the Kendall Square area-- not just Kendall Square Station, but the whole area around.

So if it's true, A lot of people could benefit from that extension. And we don't know. So what are you going to do? You're going to go to a specific stop where it is very likely that the people who would be going to MIT or those areas of Kendall Square that would benefit from this extension would alight, and you're going to ask people, would you have stayed on your bus if this bus had continued to MIT and Kendall Square? It's a simple oral survey, yes or no question, one question.

You're going to work in teams of four people. The stop that you're going to station yourself in is shown in figure 3. And you're going to collect data for the AM peak, from 7:30 to 9:30. You pick the day. The teams are assigned on Stellar, so please log into Stellar and see what your team is and coordinate with them to pick a day.

And tell me what that day is, because-- actually, right after class, I'm going to set up a shared spreadsheet that you can all access. And just go into that spreadsheet and pick a day. I'm going to put all the days that are available, and you can say team 1, team 2, et cetera. Make sure that two teams don't go on the same day. We want data from different days.

And you're going to all bring that data together in that same spreadsheet, and there are some questions for you to analyze the data that you collected, all of the class collected together. You're measuring the percent of people who would have stayed on the bus, right? So it's a proportion.

And one submission per team in PDF format to Stellar. This is due March 7, but in order to leave you enough time to do the analysis, the data collection efforts should be done by February 28. So please submit your data by the end of Tuesday, February 28 at midnight, say, or sometime before the beginning of March in the morning, where a person would be

trying to analyze your data.

OK, if you have questions, let me know. And if not, have fun. Remember that assignment 1 is due Thursday. Eric?

AUDIENCE: Just the one question: [? is that ?] [? this is ?] going to miss anyone who is transferred to the Red Line to then go to Kendall Square.

GABRIEL SANCHEZ-MARTINEZ: And going back to-- let's see. I forget where I had it. Well, I guess what I-- there was a point I made earlier where we can measure that from automatically collected data, right?

AUDIENCE: OK.

GABRIEL SANCHEZ-MARTINEZ: Does that make sense?

AUDIENCE: Yeah, people who [? car up ?] come from 70.

GABRIEL SANCHEZ-MARTINEZ: So if I see you tapping of the 70 or the 70A, and then I see you tapping at Central Square, I can infer that you were using the service to transfer to Central Square. And then we'll cover ODX, which is an inference model for destinations later in this course.

But looking at the sequence of taps, I can infer-- we can infer-- what the destination of that bus trip was. We can infer that it was the stop that was closest to Central. And later that day, presumably the person who might be going to Kendall Square Station after work taps to Kendall Square. So I might think, oh, he took the Red Line from Central to Kendall. So I don't need to ask those people where they're going. And anyway, they might not care about this extension. So we're going to stand on the bus stop that is after Central Square and see where those people are going and whether they would have stayed on that bus.

AUDIENCE: Is this an actual [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: Some people are proposing it. It is a real proposal. The MBTA is a big organization. So I can't say that the MBTA wants to do this or doesn't want to do this. But some people are interested. And it will get looked into. So it's useful.

AUDIENCE: [? Can ?] [? we ?] [? share ?] [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: Yeah, why not?

AUDIENCE: [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: Yeah. And I guess one other thing that I-- yeah, so we're going to probably make of this like a theme of assignments. So there's going to be another assignment on surface planning, operations planning. So we're going to start looking at this combination of Route 70 and 70A, and we're going to essentially make a thread of this and do some serious planning on some scenarios where the 70 and the 70A could be merged.

And they could maybe be terminated a little-- yeah, we'll make some changes to the service plan under some hypothetical scenarios. And you'll get a chance to do an operations plan on these. And then the last homework will be on policy, so there might be some policy questions that I have in mind about what we could do about service outside, on the outer parts of the 70 and 70A. All right?