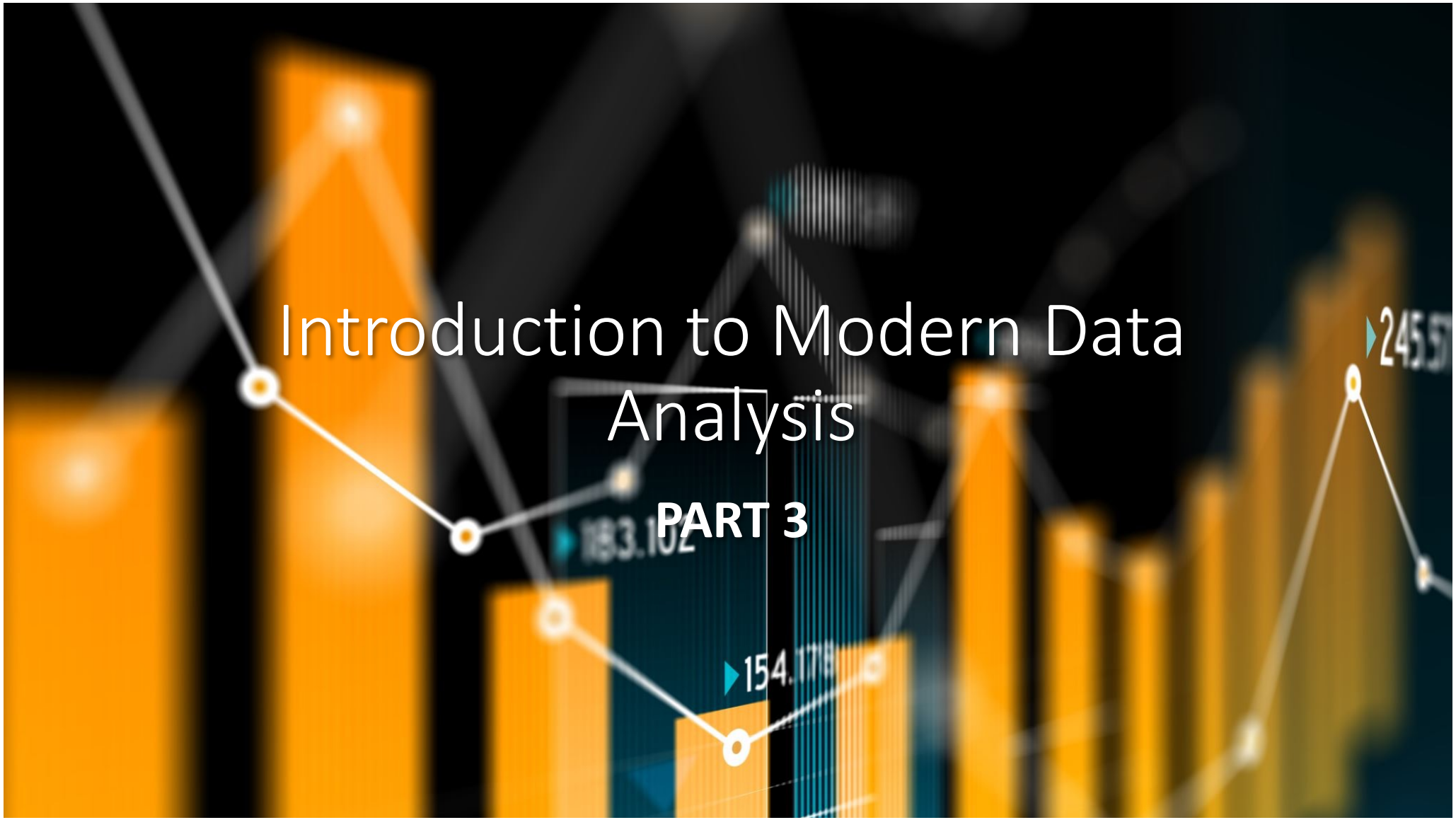
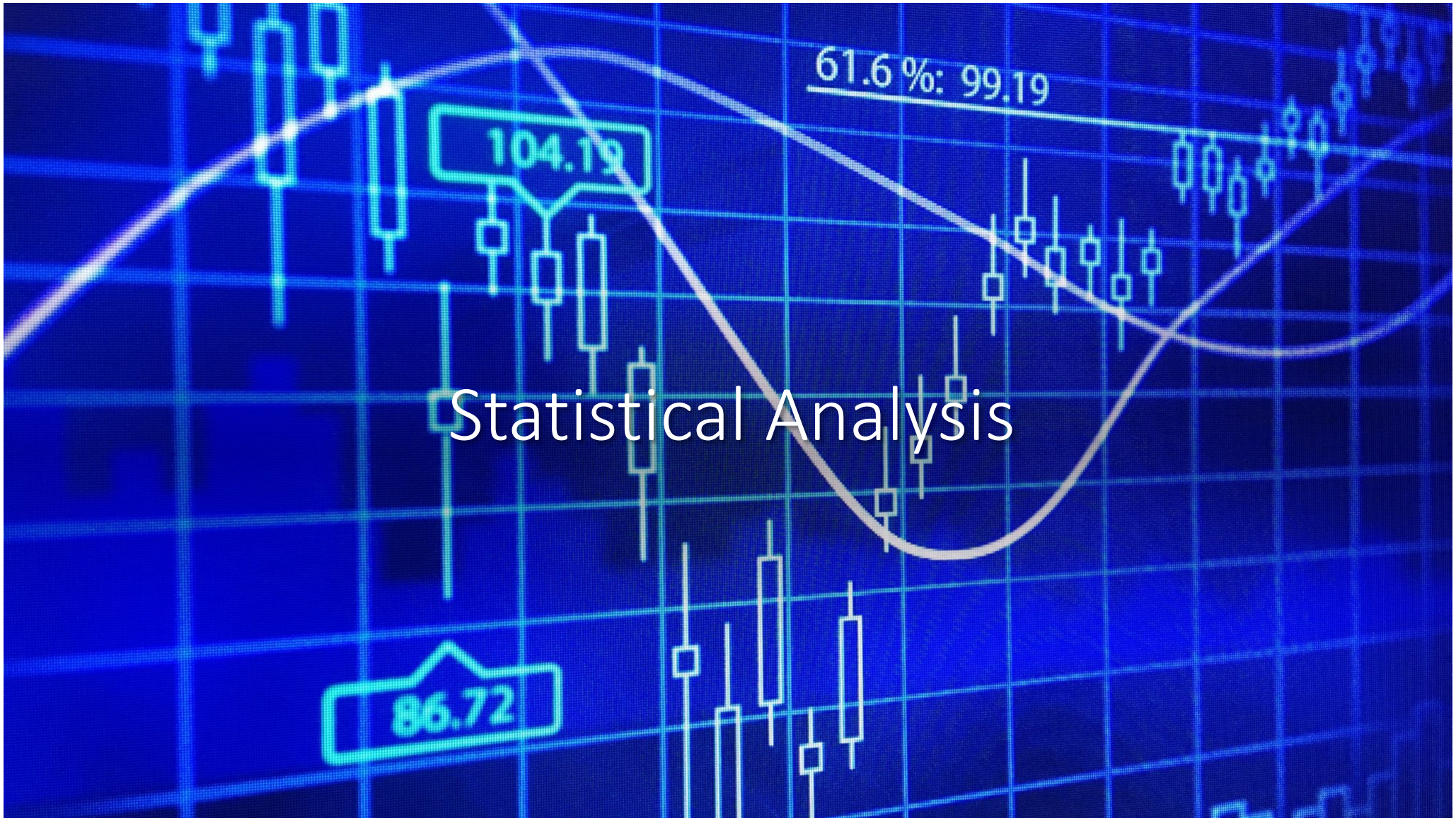


Introduction to Modern Data Analysis

PART 3



Statistical Analysis

A Very Quick Tool Discussion

Things to think about when you select analysis tools

A. Capability: What is their functionality + performance – do they have all the techniques, do they have the processing power
B. Integration: How do they connect to other parts of your pipeline
C. User-Experience: What is the user experience like – what background/level of expertise do you need to operate this tool, how easy is it to use this tool?
D. Cost – short and long term

Tools for statistical analysis (I)

- R packages
- Python modules
- Enterprise (aka $$$$) Specialized Commercial Software (SAS, SPSS)
- Other GUI – Excel, PowerBI*
- Other niche – e.g. Julia

Tools for statistical analysis (II)

RULE OF THUMB 1: A t-test is a t-test is a t-test.

RULE OF THUMB 2: Do NOT implement any statistical analysis technique by hand (unless for fun/better understanding).

R will 99.99% guaranteed have any statistical technique you want for free

Python probably will have most as well

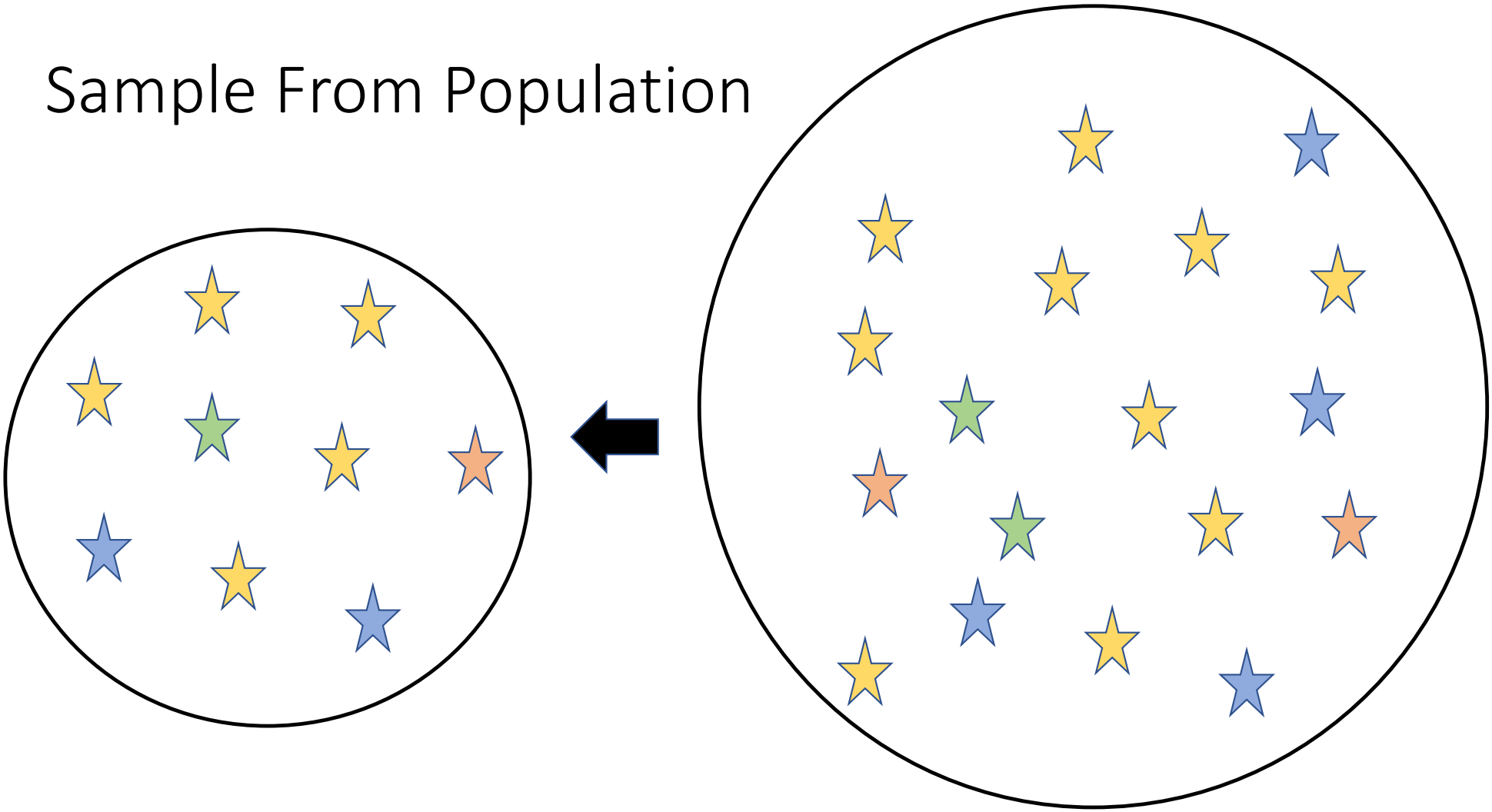
So tool choice basically comes down to how you want to prioritize/optimize A, B, C and D

PowerBI and Statistical Analysis

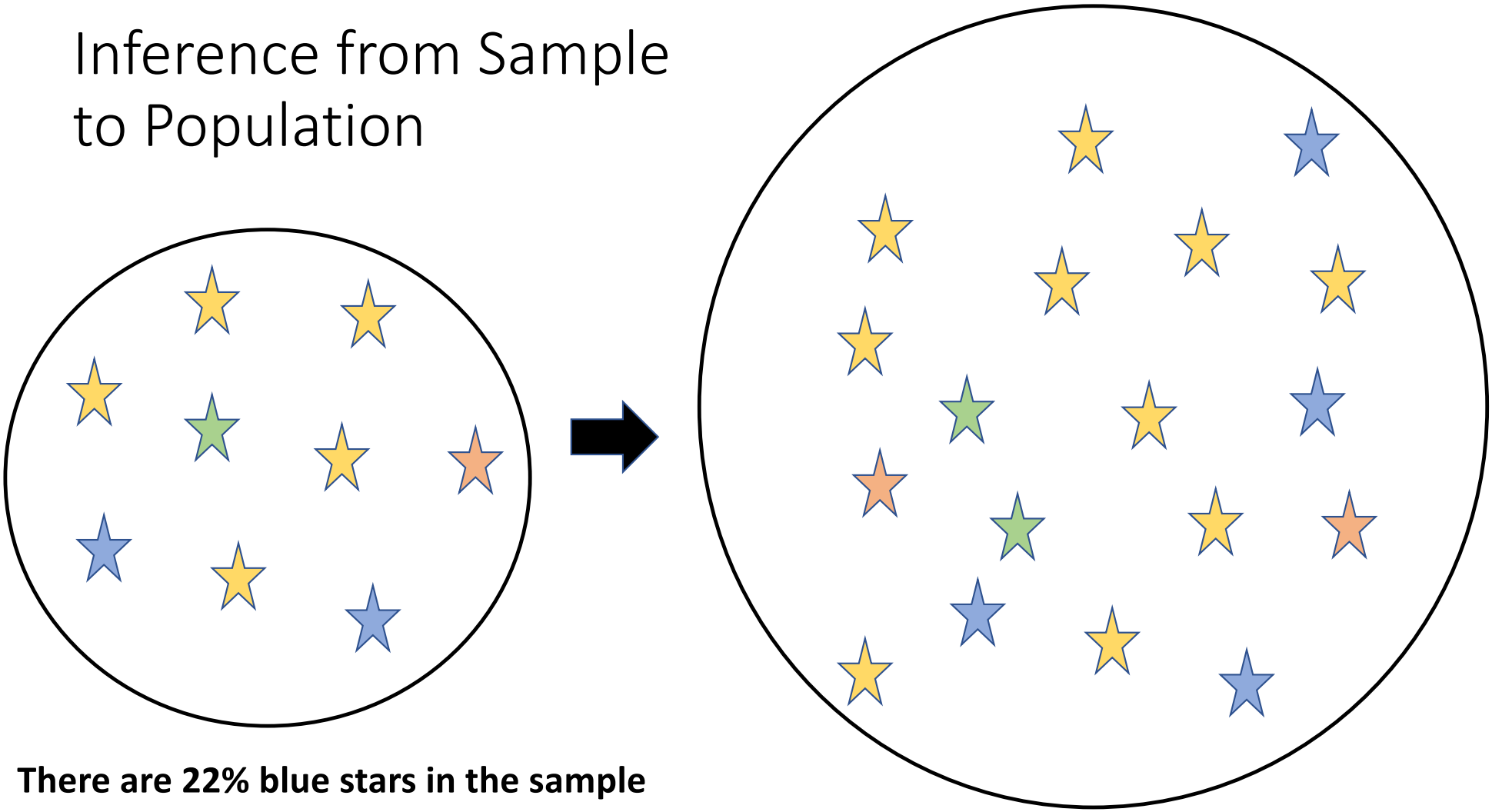
- PowerBI can be used for basic descriptive statistics (e.g. mean, max/min)
- PowerBI has **some** DAX functions that can be used to carry out **some** of the calculations required for some types of inferential statistics (e.g. confidence interval for a mean)
- However, this DAX functionality is quite limited! Even Excel is arguably better in terms of its statistical analysis functionality.
- Better option – use another tool to generate statistical results, then import and visualize in PowerBI
- Even better option – embed R code right into PowerBI


Statistics in the Modern Age

Sample From Population



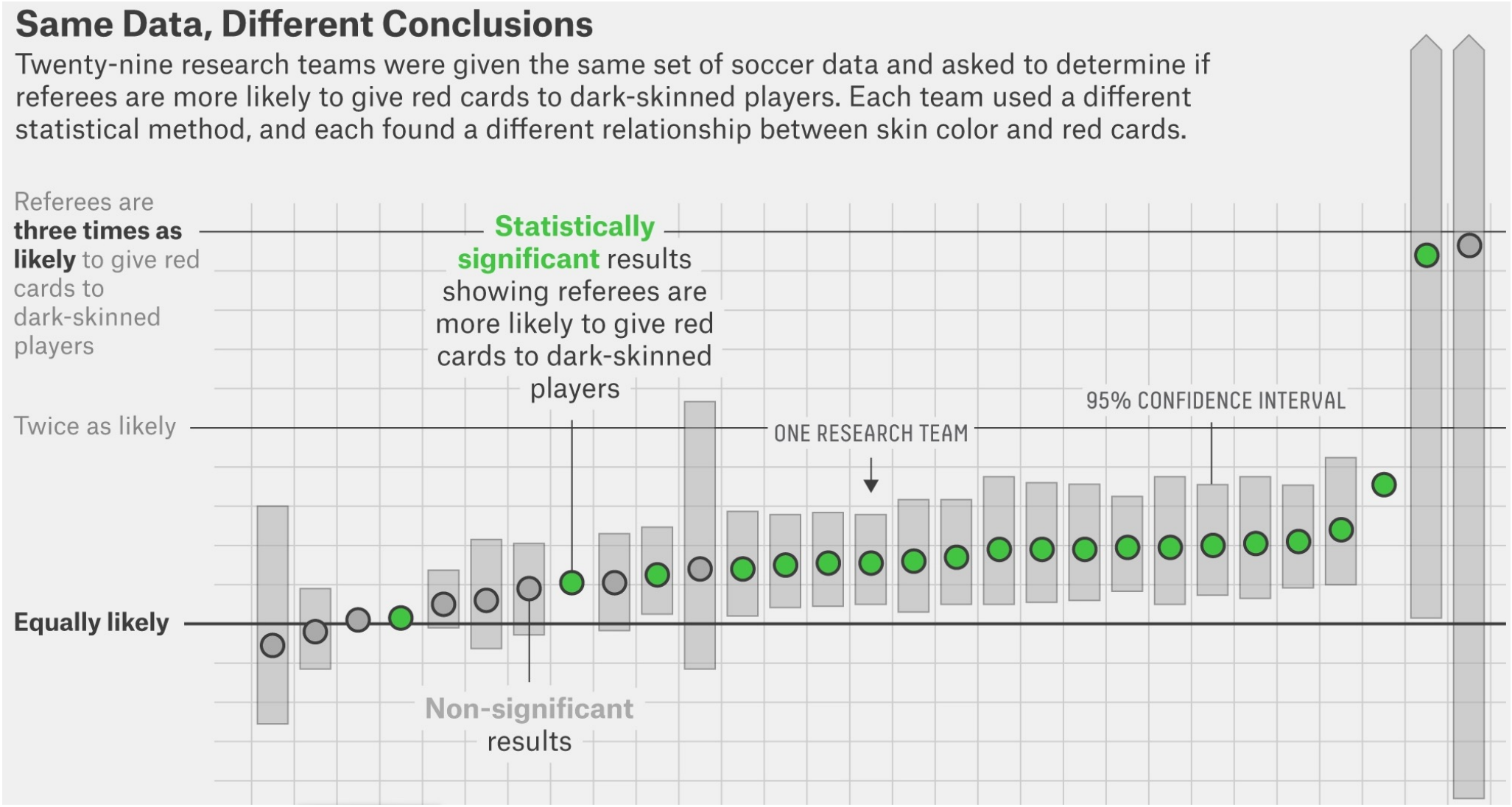
Inference from Sample to Population



There are 22% blue stars in the sample

Present-Day Statistics

- At the moment, statistics is having a bit of a tough time!
- Could this be a growth opportunity for the discipline?
- Perhaps a great time for the democratization of statistics. Desktop data science!



From: Science Isn't Broken - It's just a hell of a lot harder than we give it credit for. (Christie Aschwanden, 2015)

Does this mean we throw out statistics?

No!

- Statistics provides three very specific types of insight or knowledge that we often want, (that other techniques don't do as well):
  - Value (point or interval) estimates
  - Comparison of values, checking conditions (via hypothesis testing)
  - Understanding of associations (A and B are connected in some way)

Physical Sciences

Relatively straight-forward measurements (weight)

Relatively objective measurements

Measure the same object multiple times: Any difference is error

Independence of objects and measurements

Measurement and environment conditions strictly controlled

(sometimes) simple variable relationships

Social Sciences

Measurements can be subjective (e.g. happiness?)

High variability across objects of the same type (e.g. humans are different)

Variability is not (just) due to measurement error

Easier for dependencies to arise

Less control over environmental/experimental conditions

Complex relationships

Applied Data Analysis

All the problems associated with social sciences, plus…

Data is not originally collected with analysis in mind

Typical no control over data collection conditions

Data is typically observational

Independent/dependent variables not easily controlled

Data collection does not necessarily occur with high quality control

Alternate Analysis Techniques

The emergence of alternate analysis techniques means that conventional statistics is ‘under attack’ from many different directions:

Bayesian Statistics

Causal Analysis (Bayesian Networks)

Machine Learning

As a result – much of what you learned about statistics in school may need a bit of an upgrade

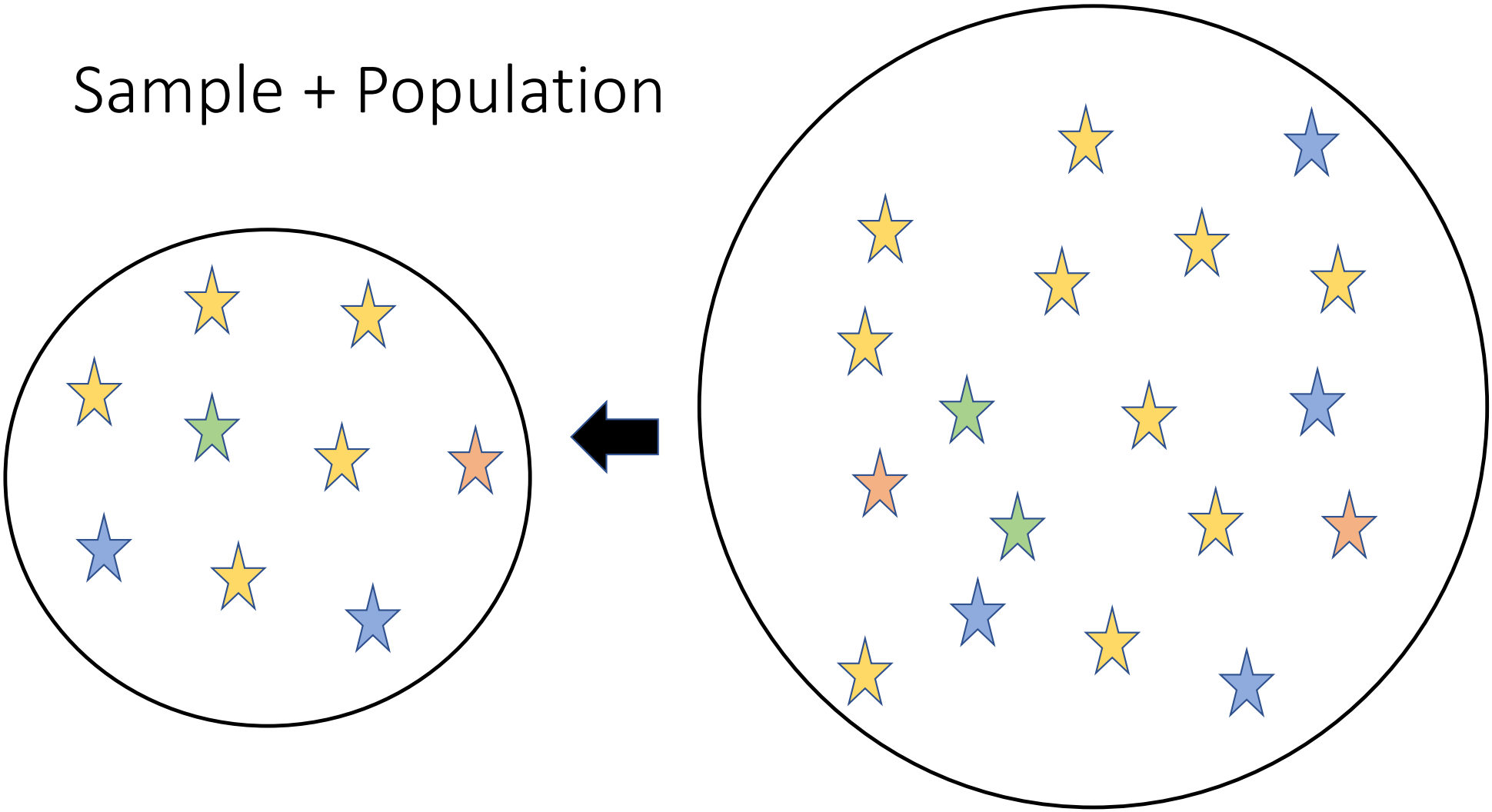
Strategies

How to deal with all of this?

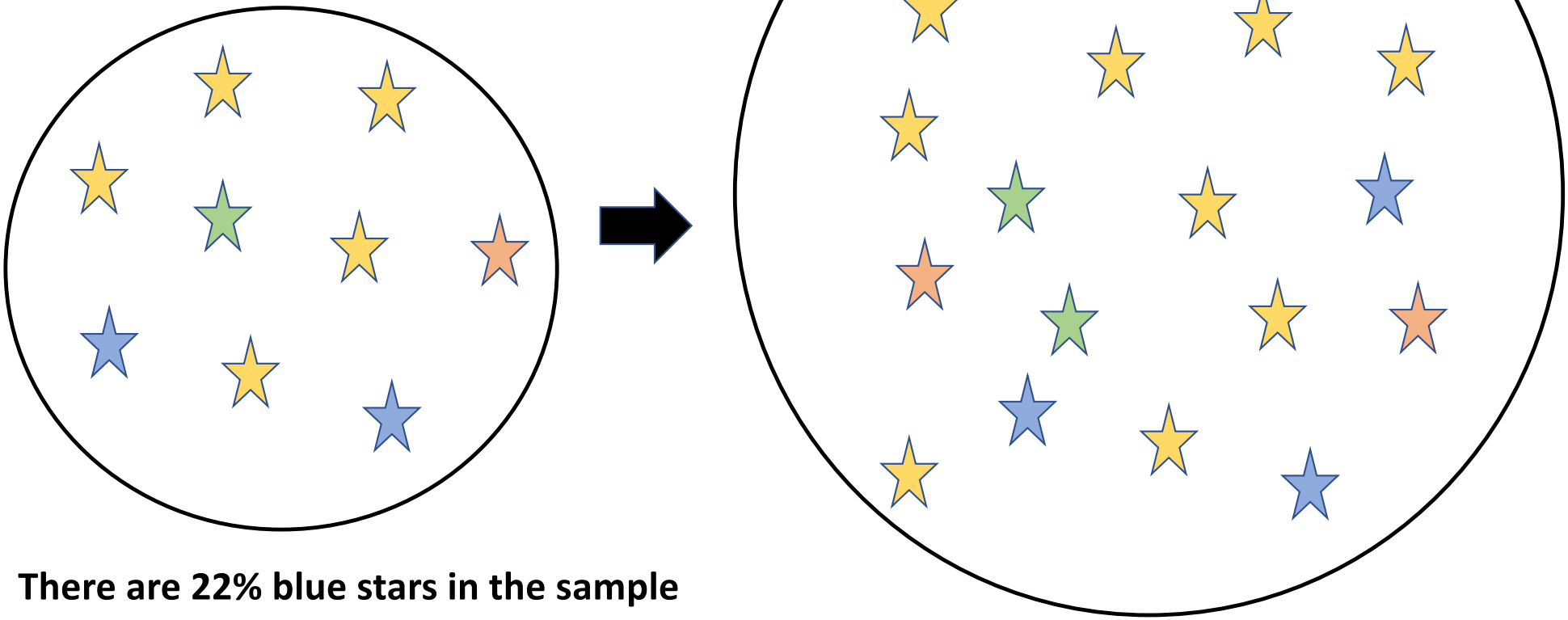
- Tools are getting more intelligent
  - Example – with categorical data, if you have small counts you often need to switch techniques - e.g. if you have less than a certain number of values in a chi squared test, you need to do things slightly differently
  - Good tools will automatically make this adjustment for you
- Have an expert statistician on staff or on call BUT – one versed in modern methods and who is not afraid to get their hands dirty!


Some Useful Statistical Concepts

Sample + Population



Inference From A Sample



There are 22% blue stars in the sample

Is It a Sample or Population?

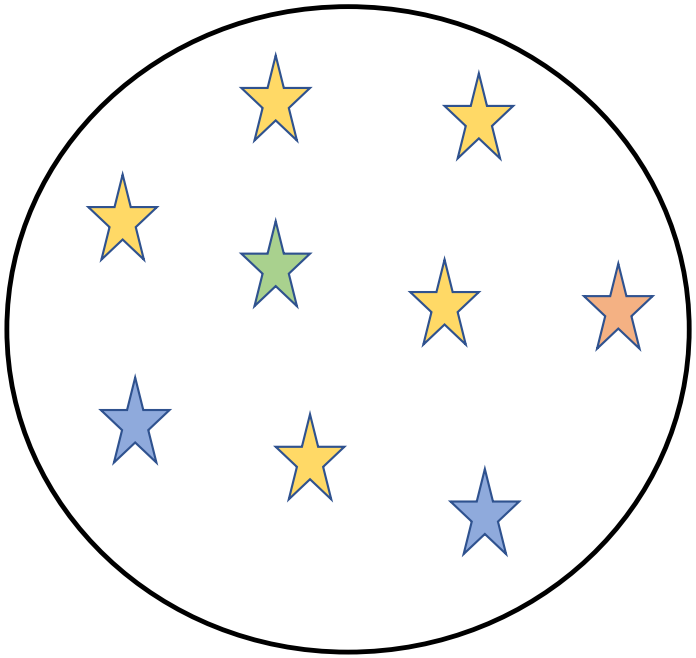
Are we dealing with a sample or a population?

Given a particular dataset, the answer to this question depends entirely on your inference goal

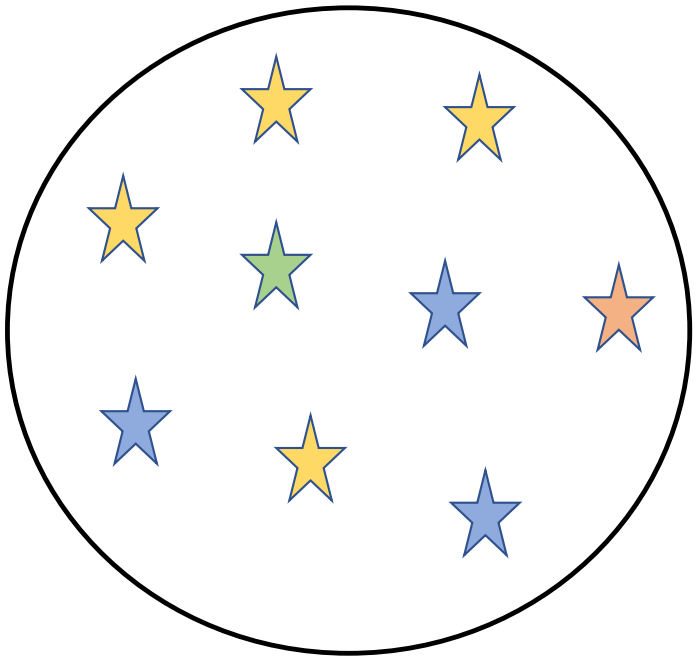
Examples:

- I want to understand this year's finances, and I currently have full data on this year's finances – no inference required – you have a population
- I want to compare this year's finances with last year's finances – no inference required – you have a population
- I want to use this year's data to derive fundamental laws about financial transactions– definitely inference, definitely a sample

We know that different samples will give us different values

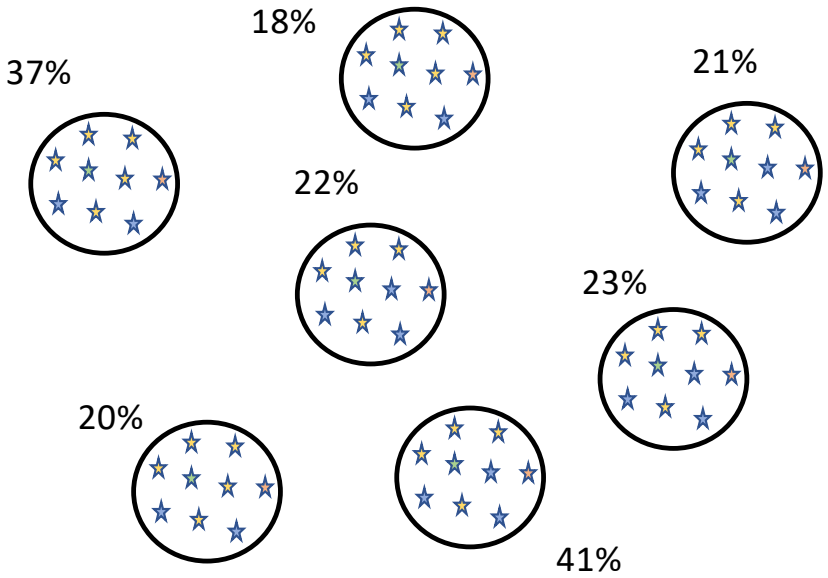


There are 22% blue stars in the sample



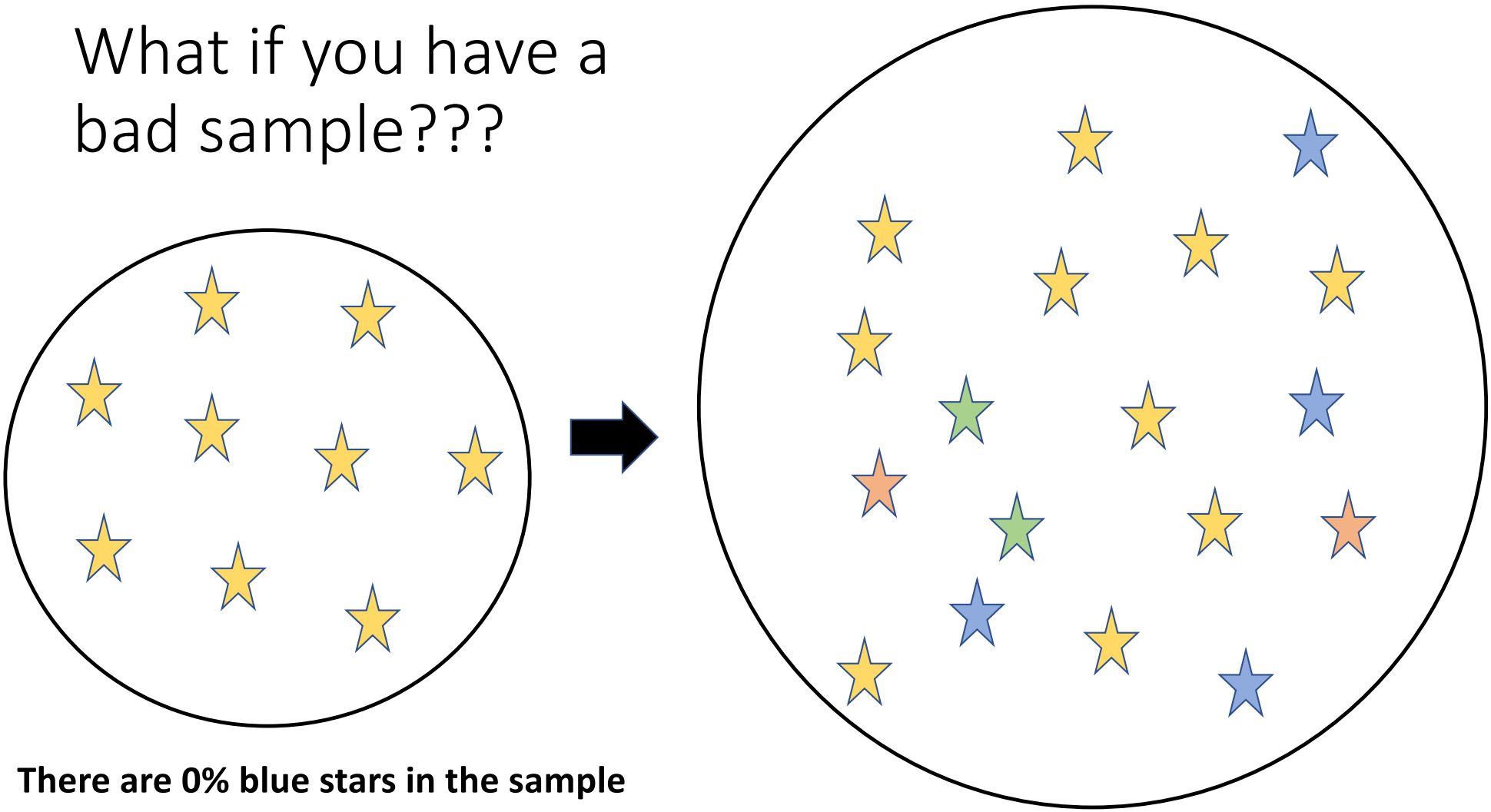
There are 33% blue stars in the sample

A population of samples? Very meta!



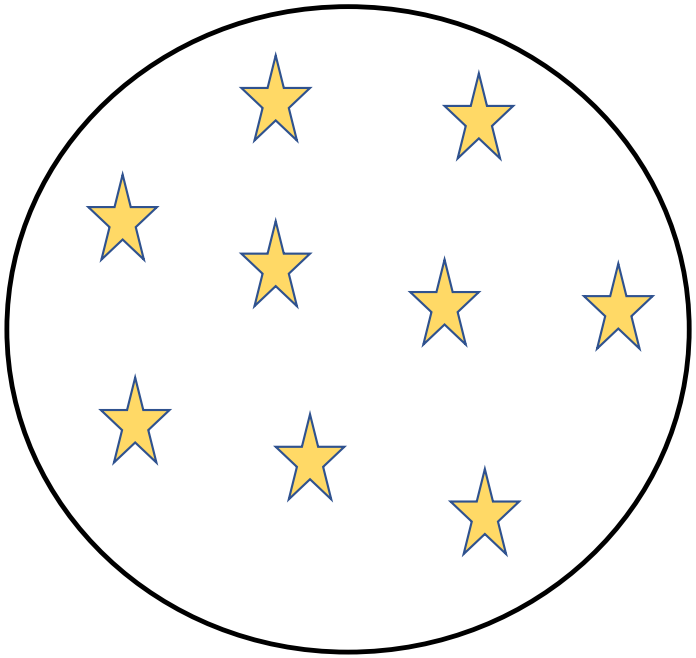
Each sample in the <u>population of samples</u> has a particular percentage of blue stars

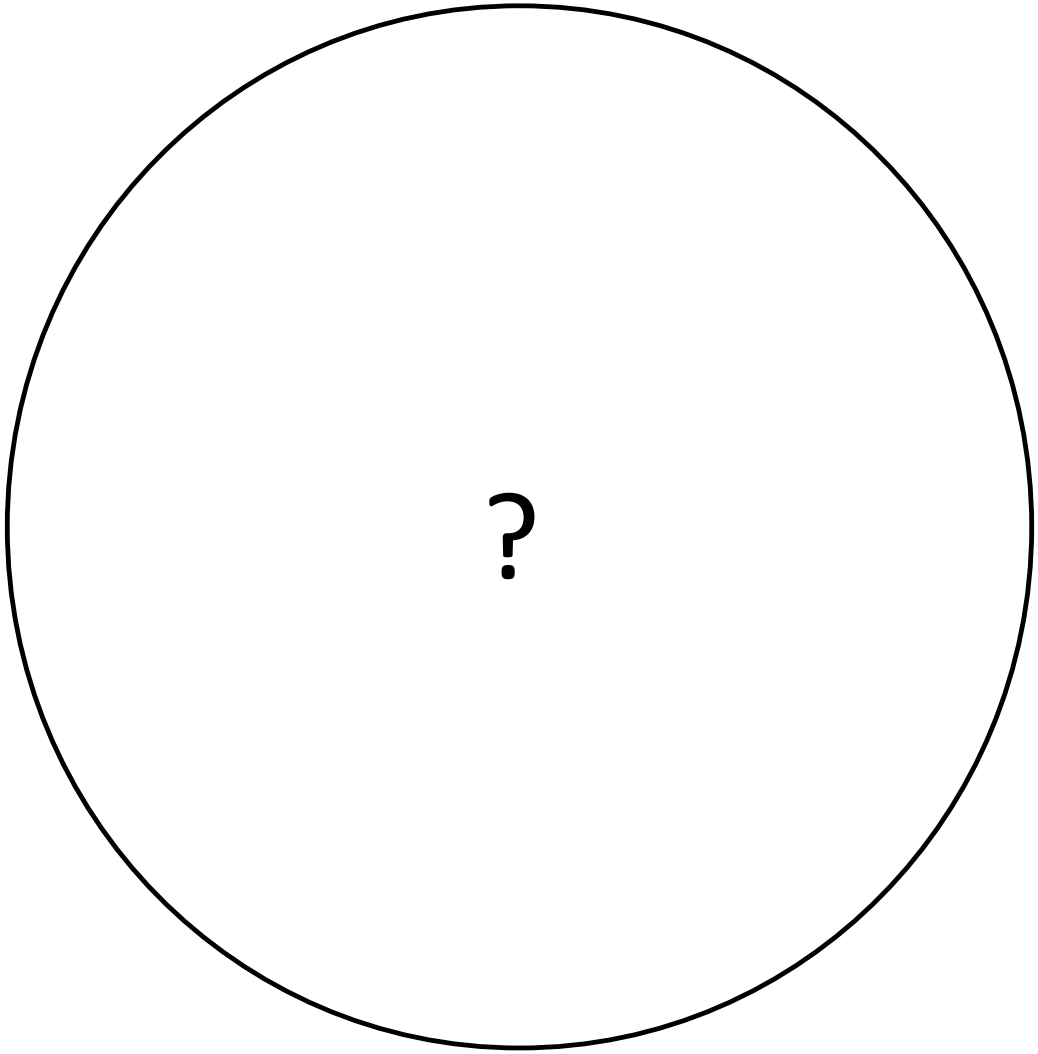
What if you have a bad sample???



There are 0% blue stars in the sample

How would you know?

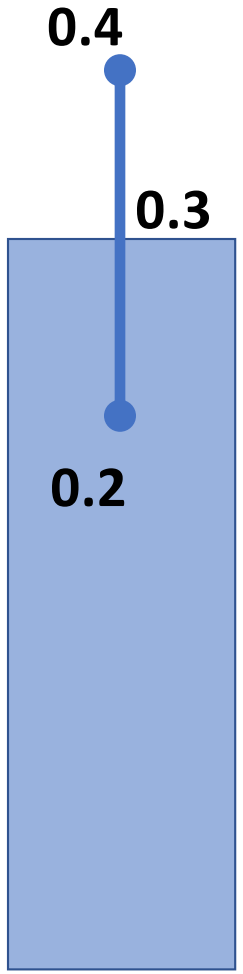




There are 0% blue stars in the sample

Confidence Intervals

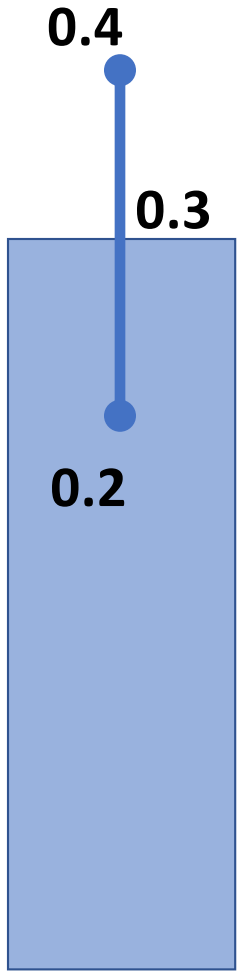
- A **confidence interval** calculation tells you that the true value (e.g. true population mean) falls within the range calculated.
- It is based off the value calculated from the sample, some additional measures taken from the sample, the size of the sample and some additional assumptions.
- The biggest one of these assumptions **is that we have a good sample**!
- Which is to say, a representative sample.



(https://towardsdatascience.com/statistics-are-you-bayesian-or-frequentist-4943f953f21b)

Confidence Levels

- We can further ask: how surprised would we be if it turned out this is a bad sample?
- OR We can flip this around and say – how confident are we that this is a good sample?
- This depends on a number of things– like what?
- We could say we are 95% confident (or 80%, or 90%) depending on what we picked for our confidence level when calculating the interval.
- The confidence interval is connected to the confidence level



(https://towardsdatascience.com/statistics-are-you-bayesian-or-frequentist-4943f953f21b)

Statistical Tests

- What is a statistical test? A method for drawing rigorous inferences from data.
- Usually involves a comparison or check (e.g. is mean A different from mean B, is this distribution normal, is proportion C less than 0.2)

Hypotheses and Significance

- We come up with a null hypothesis and an alternate hypothesis:
  - Null hypothesis – these two traits, A and B are evenly split in the population
  - Alternate hypothesis – there is more of trait A than B in the population
- BUT we know that any given sample will almost certainly not be a perfect reflection of the actual population!
- How do we deal with this?
- **Significance** is another strategy to deal with this issue.

Significance: Definition and Interpretation

If the sample is big (and the difference is big) and we have reason to believe the sample is representative… we would be pretty surprised if the difference was due to a bad luck sample.

Significance quantifies this intuition.

Technical Definition of Significance Level: the probability of getting results *at least as* extreme as the ones you observed, given that the null hypothesis is correct.

Example: There's an 20% chance that we would get a difference in proportions this big or larger if the difference between the populations was actually 50%.

Interpretation: How surprised will you be if the null hypothesis turns out to be true, under these circumstances.

Significant is not the same as substantial.

Significance: Sample vs Population

As the sample gets larger and larger it gets closer and closer to being equivalent to the full population*

With large sample sizes, even very small differences become significant – we would be very surprised if the difference was just by chance.

For this reason: Significant is not the same as Substantial

Once we get to the full population, any size of difference is significant (aka real!). There's no way the difference is just by chance.

IMPORTANT: no need for inferential statistics in the case of populations – also no need for significance, or tests!

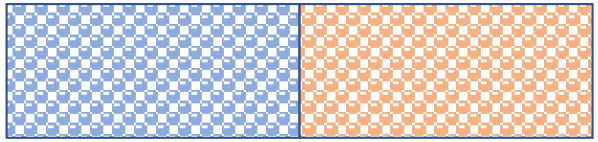
Car Driving Test Analogy*

*(Not a perfect analogy, but hopefully helpful…)

Suppose there are many different organizations that carry out car driving accreditation tests.

If someone came to you and said “I passed my test with flying colours!” you might also want to know something about the place that carried out the test.

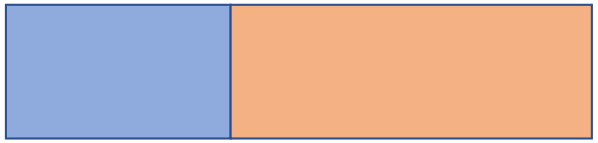
The test itself might be the same across places, but that doesn’t mean that all of the test results are really equal – it depends on how the test is marked



Null Hypothesis

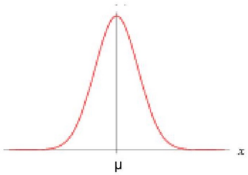
Population Proportions

Possible Samples



Sample 1

Sample 2



Sampling distribution: distribution of sample proportions.

Sample 3

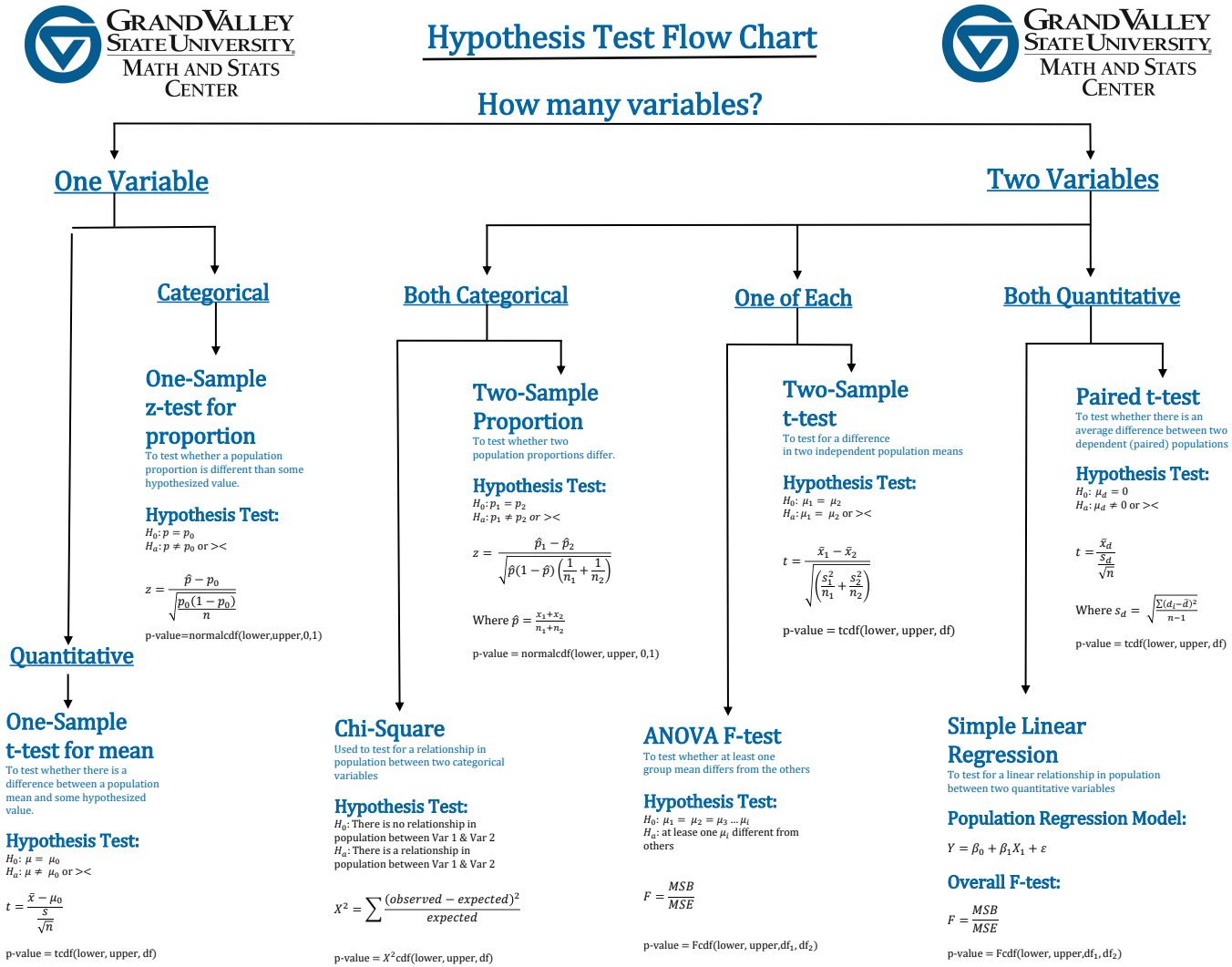
Sample 4

Suppose we end up with Sample 4 – it does show a difference from our null hypothesis. How surprised would we be if it turned out this difference was just due to bad luck (we could have ended up with sample 1, after all)

Tons of Statistical Tests

For a comprehensive table of statistical tests:

Choosing a Statistical Test (Summary and Analysis of Extension Program Evaluation in R, Salvatore S. Mangiafico, 2016 **https://rcompanion.org/handbook/D_03.html**



Statistical Tests in Applied Situations

Problem : Traditional Vanilla Statistical Tests (e.g. T-Test, Z-Test) *really only work well in scientific experiment contexts.* Specifcially, where you have:

Developed hypotheses *in advance* of collecting data

Selected sample size based on power you think you need, again *in advance*

Used a sampling method that will likely provide a representative sample of data

Controlled the conditions of the sample selection as well as other experimental elements (environment, hypotheses being tested) to make it easy to draw strong conclusions

Good reason to believe that the system represented by the data conforms to the other required assumptions of the test (e.g. normal distribution, independence)

Tests Example 1: Chi Square

Count of Companies Relative to Year (In Canada)

Year	Large	Medium	Small
2017	2285	537	418
2018	2274	491	448
2019	2379	570	455
2020	2385	552	435
2021	1416	331	261

But suppose we considered this a sample…

(with the population being…?)

Year	Large	Medium	Small
2017	2285	537	418
2018	2274	491	448
2019	2379	570	455
2020	2385	552	435
2021	1416	331	261

Chi Square Test for Independence

- Null Hypothesis: **The distribution of the outcome is independent of the groups** (no relation between variables)
- Alternative Hypothesis: **there is a difference in the distribution of responses to the outcome variable among the comparison groups** (relationship between variables)
- Let's pick an *alpha value* of 0.05 .
- This tells us that we would be '95% of maximum-surprised" if the null hypothesis was actually true in this situation when our test statistic said it was false.
- we'll reject the null hypothesis if the *p-value** is LESS THAN this value. We want our 'minimum surprise level' to be 95%.
- Remember – small differences can be significant with large sample sizes

*the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct

Code to test for independence

- isize <- read.csv("year_instituttion_size.csv") #LOAD DATA
- isize_sum <- table(isize$Competition.Year,isize$Institution.Size..ENG) #FORMAT DATA
- chisq <- chisq.test(isize_sum) #RUN TEST
- Pearson's Chi-squared test
- data: isize_sum
- X-squared = 4.8042, df = 8, p-value = 0.7783

The p-value is much larger than 0.05, so there is no dependence between variables.

In other words, year is not associated with size of company. We cannot use information about year to help us guess proportion of large, small or medium companies in a given year.

Some additional points about statistical tests

Non-Normal Data: If you know you have non-normal data, turn to non-parametric tests. If you don't, you can do a test to see if your data is normally distributed, and then proceed

Non-parametric tests: These make no assumptions about the distribution of the data so you can use them whenever you want. Unfortunately they can be less definitive and informative.

Transforming your data: Another option is to transform you data from not normal to normal, and then use parametric tests.

Alternative To Using Statistical Tests for Hypothesis Testing: Statistical Modelling

Statistical modelling tends to get complicated very fast.

BUT the reality is that real world data is usually complicated.

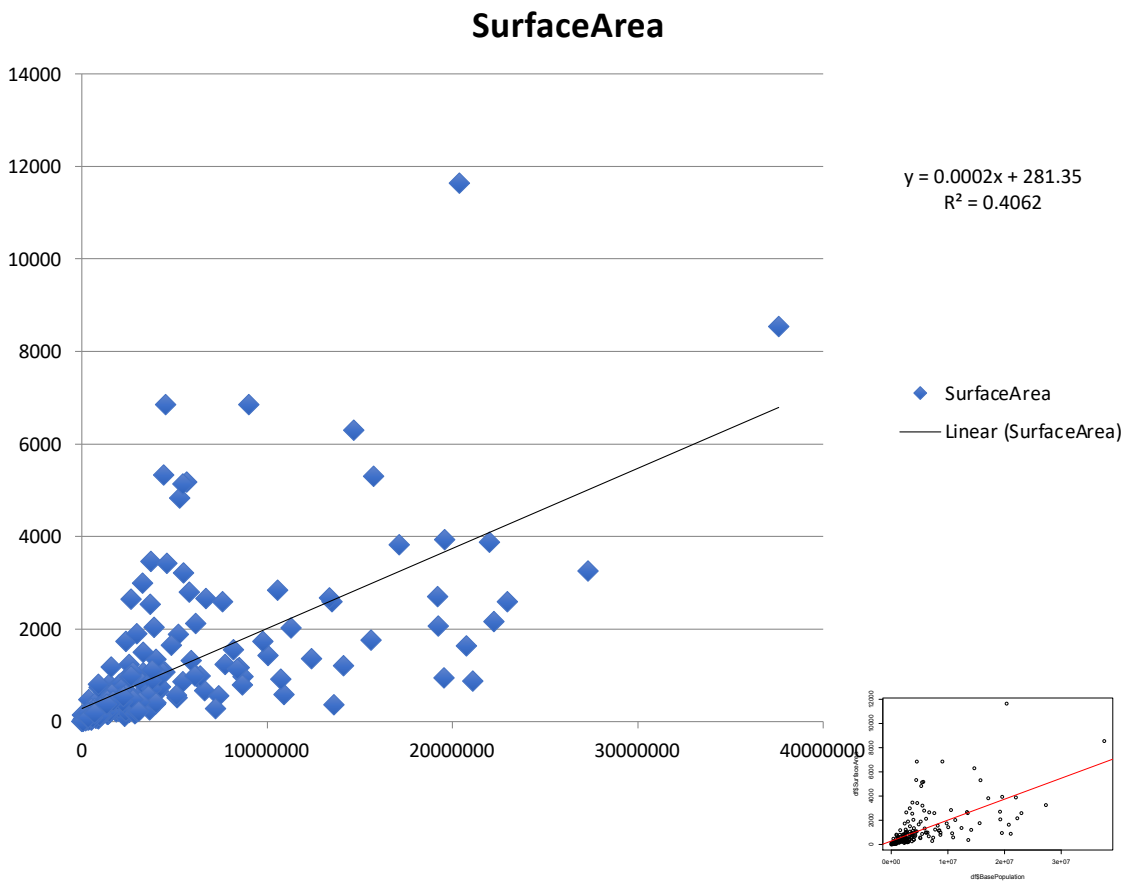
This is particularly the case with OBSERVATIONAL DATA.

What this means is that with observational data you don't really have an easy middle ground – you either:

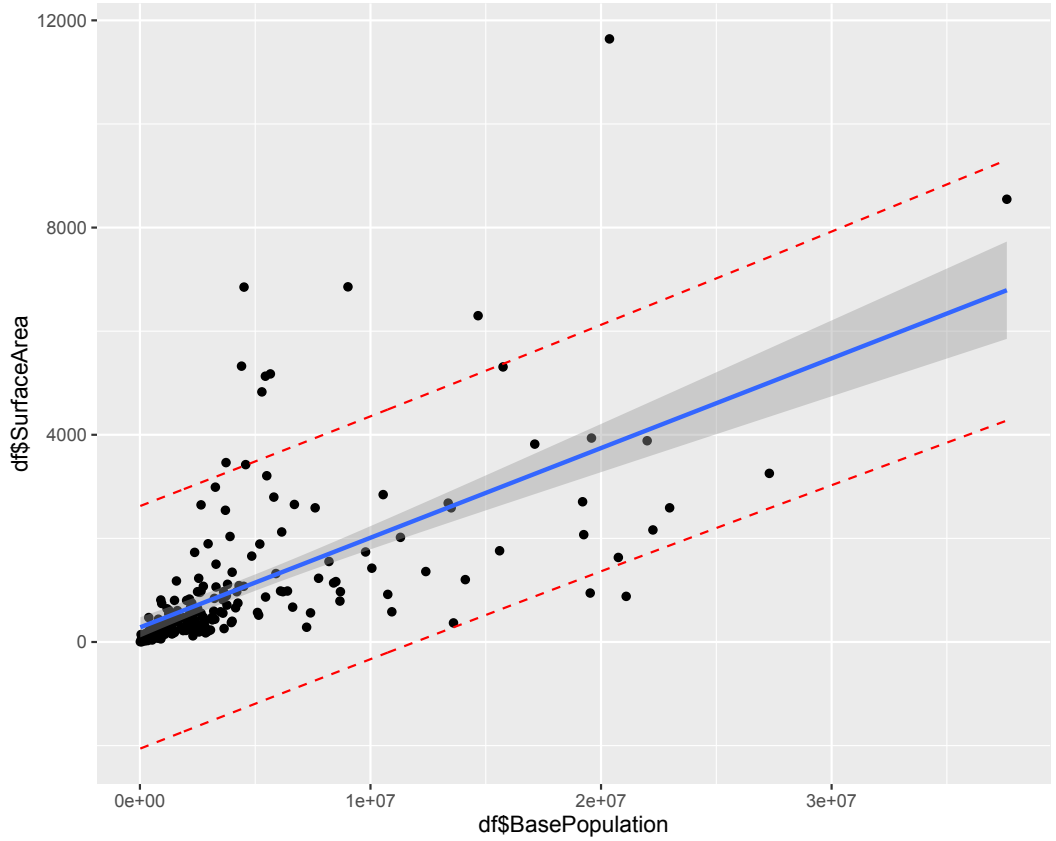
- do descriptive statistics OR
- you jump to consulting with a professional statistician OR
- get used to doing creating complicated statistical models yourselves (preferably with the help of a tool like R)

Two Variable Linear Model

- There are straightforward techniques to let you fit a linear (straight line) model to your data. You can do it in Excel!
- Let your relationship coefficient help you decide if this is a valid (or useful) model



Confidence + Prediction Intervals



blue line = linear model
grey band = 95% confidence interval
red lines = 95% prediction interval

- The confidence and prediction intervals give us a sense of how certain we are in our model.
- Confidence interval tells us what ‘y’ we can expect on average for a given x.
- Prediction interval tells us the range we can expect ‘y’ to fall in, for a specific instance of ‘x’.

Chi Squared Test - Log Linear Models

Chi Squared: Tests to see if there is an interaction between categorical variables

Instead you can use a log linear model to detect and further investigate the associations and interactions between variables

For more stats and ML, check out our Data Science Report Series

- https://www.data-action-lab.com/data-science-report-series/