



Data Science for Smart Cities

CE88

Prof: Alexei Pozdnukhov

Today



Variability of samples, confidence intervals

Variability of predictions

Minilab 8

Midterm planning

Statistics terminology



Inference: Making conclusions from random samples

Population: The entire set that is the subject of interest

Parameter: A quantity computed for the entire population

Sample: A subset of the population

In a **Random Sample**, we know the chance that any subset of the population will enter the sample, in advance

Statistic: A quantity computed for a particular sample

Parameters and intervals



A reasonable way to estimate a parameter such as the population average, max, or median is to compute the corresponding statistic for a sample.

Different samples will lead to different estimates.

Goal: Infer the variability of a statistic, using only a sample.

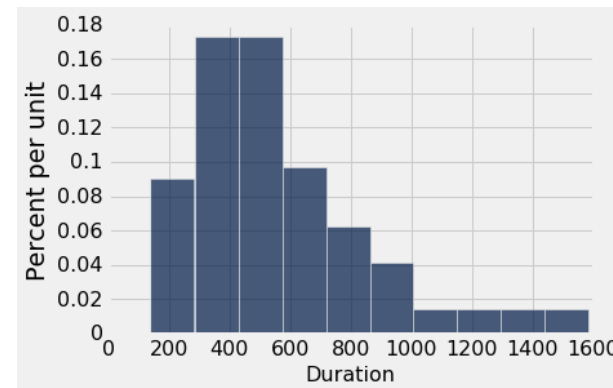
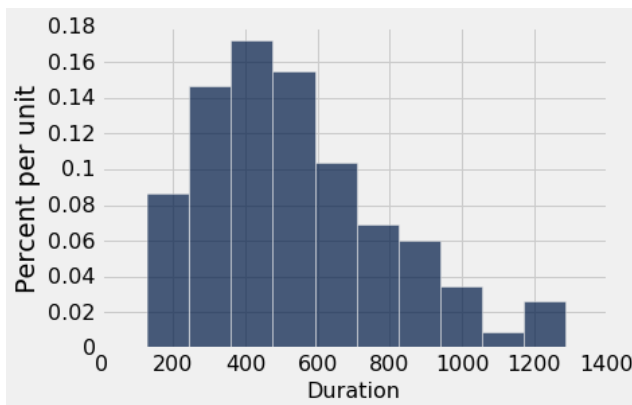
A possible solution: apply bootstrap resampling, as variability of the sample represents that of a population.

Confidence intervals



Estimation is a process with a random outcome

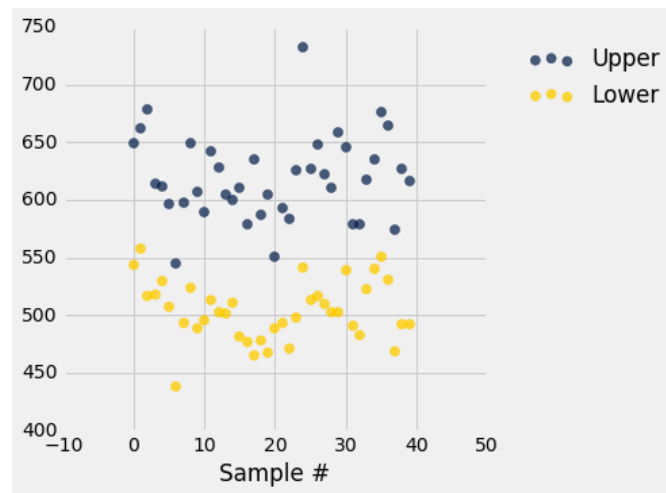
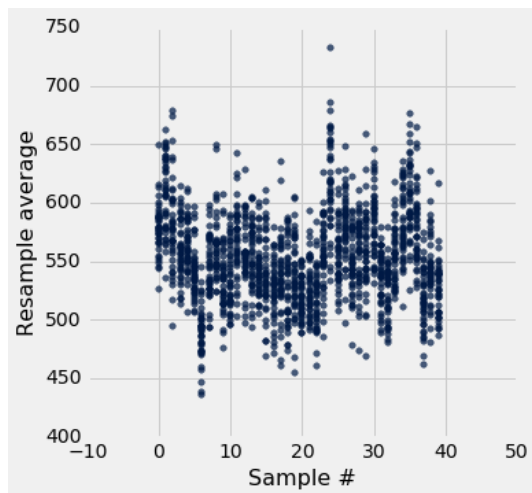
Population (fixed) → Sample (random) → Statistic (random)



Instead of picking a single estimate of the parameter, we can pick a whole interval: lower bound to upper bound

A 95% **Confidence Interval** is an interval that will contain the parameter for (at least) 95% of samples

Confidence intervals



BTW, for a particular sample, it's right or wrong & you don't know 😊

It's impossible to verify empirically whether an interval is correct when all you have is a sample.

But if you have the whole population, you can check if the intervals were correct

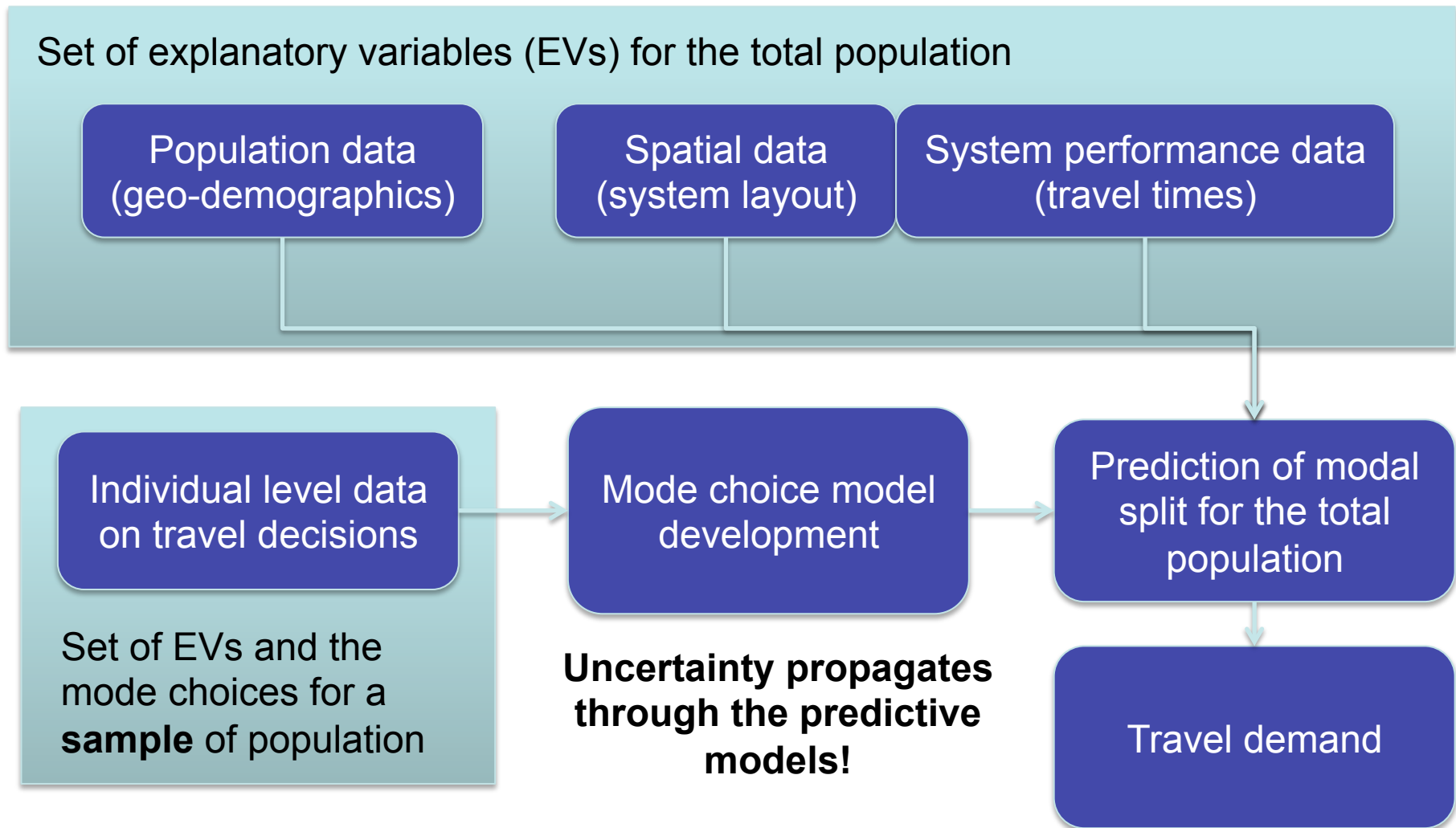


Minilab 8

- **Study the variability of the sample mean**
- **Get a gut feeling on the uncertainty of predictions**

N.B. We can get confidence intervals for any statistic we compute from a sample, not just the mean!

Recall the modelling framework



A proper decision support framework must include uncertainty estimates (for example, confidence intervals) along with any inferred statistic.