Data Science for Smart Cities

CE88

Prof: Alexei Pozdnukhov
GSI: Madeleine Sheehan

115 McLaughlin Hall

alexeip@berkeley.edu
m.sheehan@berkeley.edu
CE88 in title

Today

Midterm Q&A

Variability of samples,
confidence intervals

Minilab / Midterm Q&A

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.

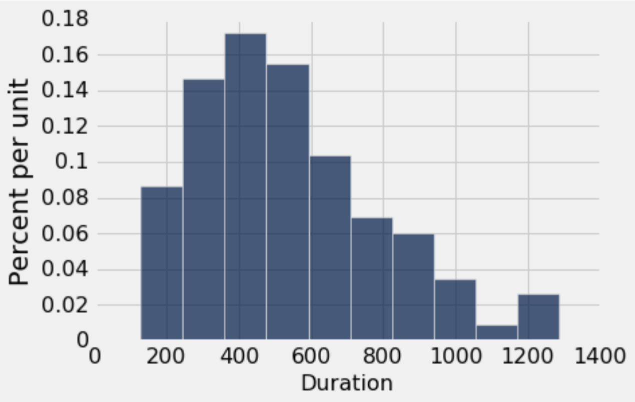
Goal: Infer the variability of a statistic, using only a sample.

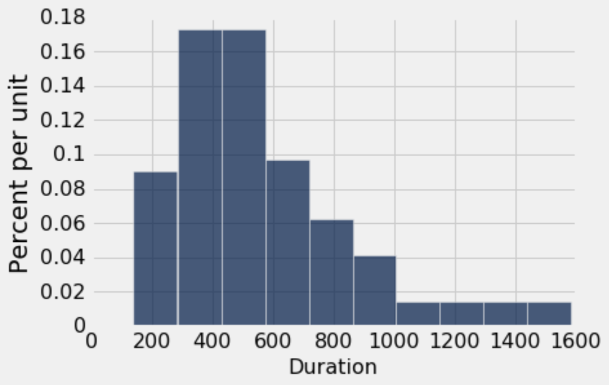
A possible solution: apply bootstrap resampling, as variability of the sample represents that of a population.

Confidence intervals

Estimation is a process with a random outcome

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

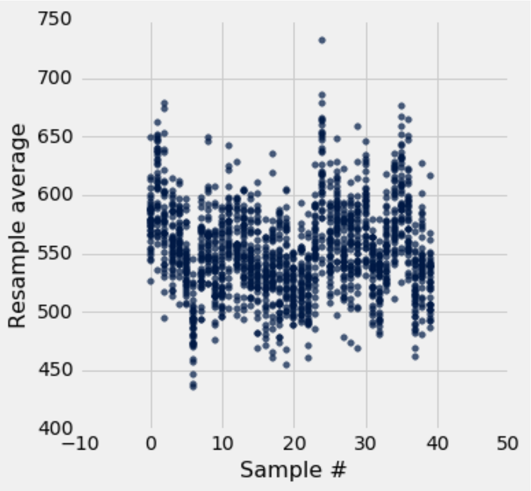


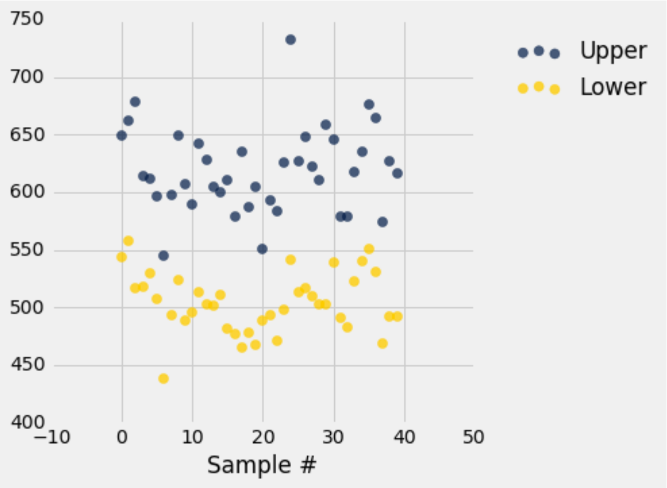


Instead of picking a single estimate of the parameter, we can pick a whole interval: lower bound to upper bound

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

Confidence intervals

Minilab 8 – study the variability of the sample mean

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

Decision making under uncertainty

In the Midterm, we study the reduction of VMT we expect to achieve.

Our intervention measure is not free, the proposed approach costs $1/citizen.

Now, consider you are a City Mayor. You are allocating a yearly budget towards greener transportation:

Option 1.

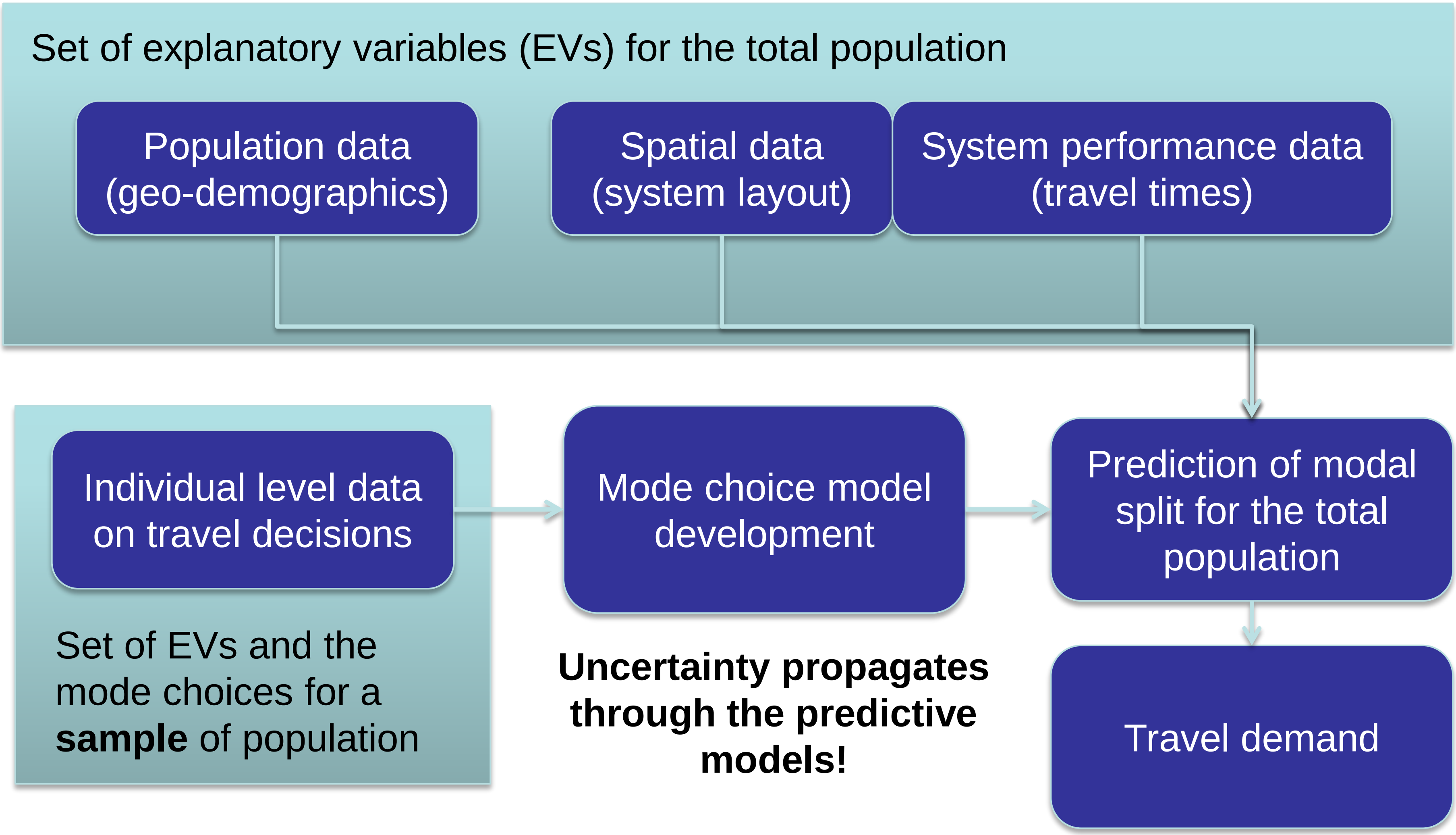
$400
for a reduction of
25%

Option 2.

$600
for a reduction of
22%

Your key point in the program was to achieve a reduction of 20%,
re-election for the next cycle is next year..

Recall the modelling framework



Set of explanatory variables (EVs) for the total population

Population data (geo-demographics)

Spatial data (system layout)

System performance data (travel times)

Individual level data on travel decisions

Set of EVs and the mode choices for a **sample** of population

Mode choice model development

Prediction of modal split for the total population

Travel demand

Uncertainty propagates through the predictive models!

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