FUNDAMENTAL STATISTICAL CONCEPTS

Analytics Primer

3 Main Pieces of Statistics

- Statistics and analytics in general boils down to three main pieces:
  1. Data Collection
  2. Data Analysis
  3. Inference

- Together these pieces summarize the data lifecycle from beginning to end.

3 Main Pieces of Statistics

- Statistics and analytics in general boils down to three main pieces:
  1. **Data Collection**
  2. Data Analysis
  3. Inference

- One of the most overlooked pieces, but the most important!
  - Bad data → Bad results!

Populations vs. Samples

- **Population**
  - Set of all objects/individuals of interest
  - Usually too large to obtain information from entire population

- **Sample**
  - Subset of the population that information is actually obtained

Populations vs. Samples

- **Population**
  - Set of all objects/individuals of interest
  - Usually too large to obtain information from entire population

- **Sample**
  - Subset of the population that information is actually obtained
  - **Sampling frame** – actual list from which the sample is taken

Populations vs. Samples

- **Population**
  - Set of all objects/individuals of interest
  - Usually too large to obtain information from entire population

- **Sample**
  - Subset of the population that information is actually obtained
  - **Sampling frame** – actual list from which the sample is taken → MAY NOT EQUAL POPULATION

Parameters vs. Statistics

- **Parameter**
  - Measures computed from a population.

- **Statistic**
  - Measures computed from a sample.
  - Sample statistics is the **point estimate** of the population parameter.

Parameters vs. Statistics

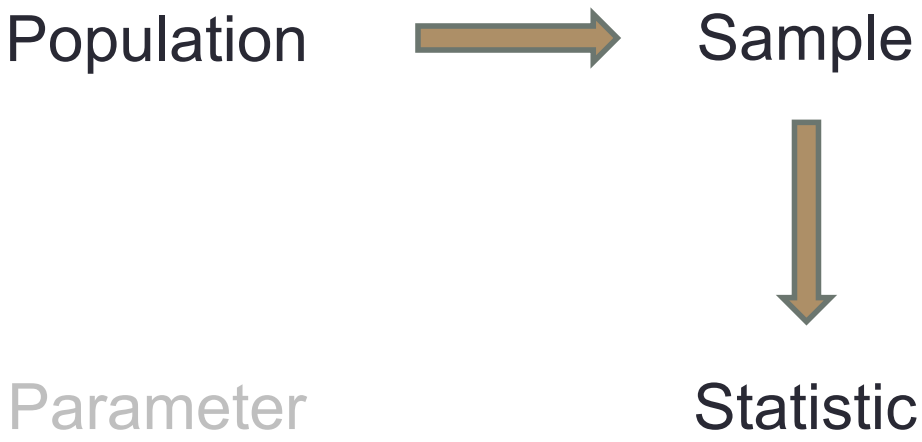
Population	Sample
Parameter	Statistic

Parameters vs. Statistics

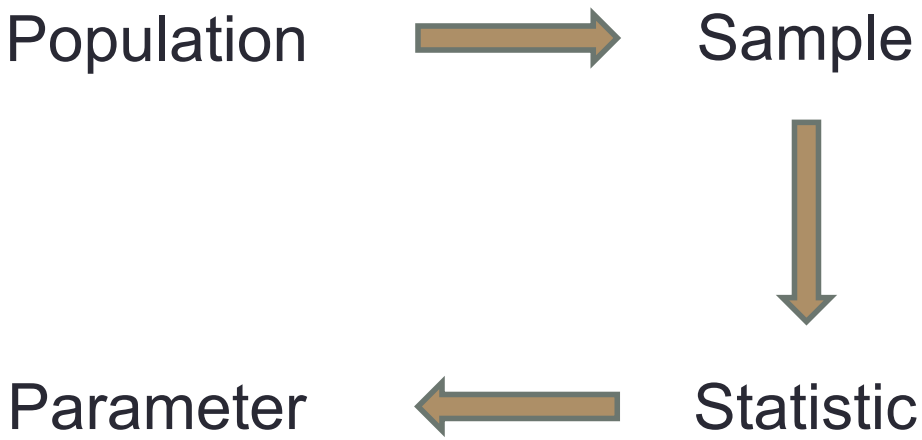
Population  Sample

Parameter Statistic

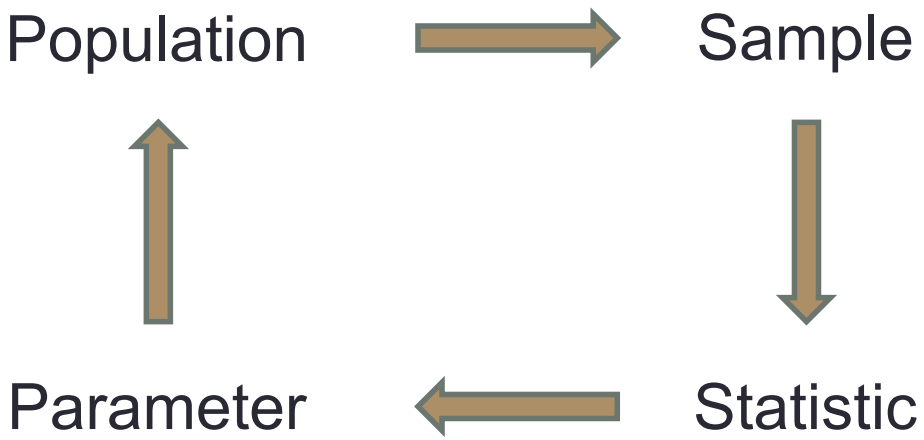
Parameters vs. Statistics



Parameters vs. Statistics



Parameters vs. Statistics



Example

- A retail chain is trying to determine if a new product they introduced is selling well across their stores. The retail chain has 2135 stores nationwide. The analyst in charge of this project is tasked to estimate the average daily sales of this new product across all stores. Older computing technology forces the company to randomly pick 179 stores spread evenly throughout the nation to calculate gather data from. The average daily sales from these 179 stores is $129.19.

- Identify population, sample, parameter, statistic.
- Any sampling frame issues?

Example

- A retail chain is trying to determine if a new product they introduced is selling well across their stores. The retail chain has **2135 stores nationwide**. The analyst in charge of this project is tasked to estimate the average daily sales of this new product across all stores. Older computing technology forces the company to randomly pick 179 stores spread evenly throughout the nation to calculate gather data from. The average daily sales from these 179 stores is $129.19.

- Identify **population**, sample, parameter, statistic.
- Any sampling frame issues?

Example

- A retail chain is trying to determine if a new product they introduced is selling well across their stores. The retail chain has 2135 stores nationwide. The analyst in charge of this project is tasked to estimate the average daily sales of this new product across all stores. Older computing technology forces the company to randomly pick **179 stores spread evenly throughout the nation** to calculate gather data from. The average daily sales from these 179 stores is $129.19.

- Identify population, **sample**, parameter, statistic.
- Any sampling frame issues?

Example

- A retail chain is trying to determine if a new product they introduced is selling well across their stores. The retail chain has 2135 stores nationwide. The analyst in charge of this project is tasked to estimate **the average daily sales of this new product across all stores**. Older computing technology forces the company to randomly pick 179 stores spread evenly throughout the nation to calculate gather data from. The average daily sales from these 179 stores is $129.19.

- Identify population, sample, **parameter**, statistic.
- Any sampling frame issues?

Example

- A retail chain is trying to determine if a new product they introduced is selling well across their stores. The retail chain has 2135 stores nationwide. The analyst in charge of this project is tasked to estimate the average daily sales of this new product across all stores. Older computing technology forces the company to randomly pick 179 stores spread evenly throughout the nation to calculate gather data from. The average daily sales from these 179 stores is **$129.19**.

- Identify population, sample, parameter, **statistic**.
- Any sampling frame issues?

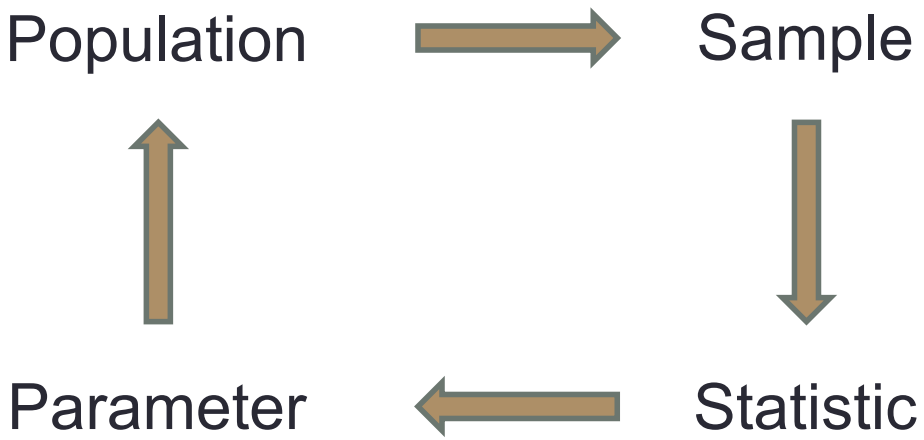
Example

- A retail chain is trying to determine if a new product they introduced is selling well across their stores. The retail chain has 2135 stores nationwide. The analyst in charge of this project is tasked to estimate the average daily sales of this new product across all stores. Older computing technology forces the company to randomly pick 179 stores spread evenly throughout the nation to calculate gather data from. The average daily sales from these 179 stores is $129.19.

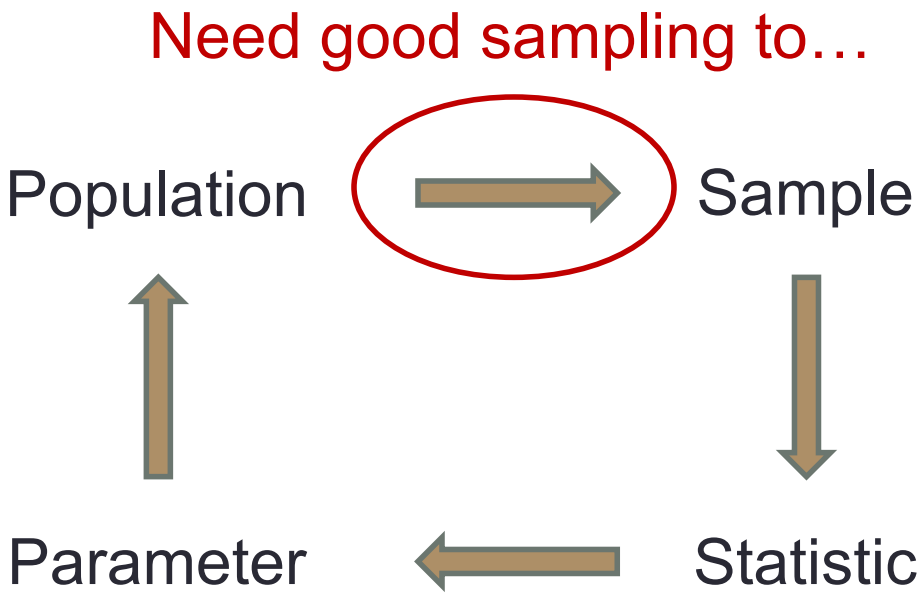
- Identify population, sample, parameter, statistic.
- Any sampling frame issues? **NO**

SAMPLING TECHNIQUES

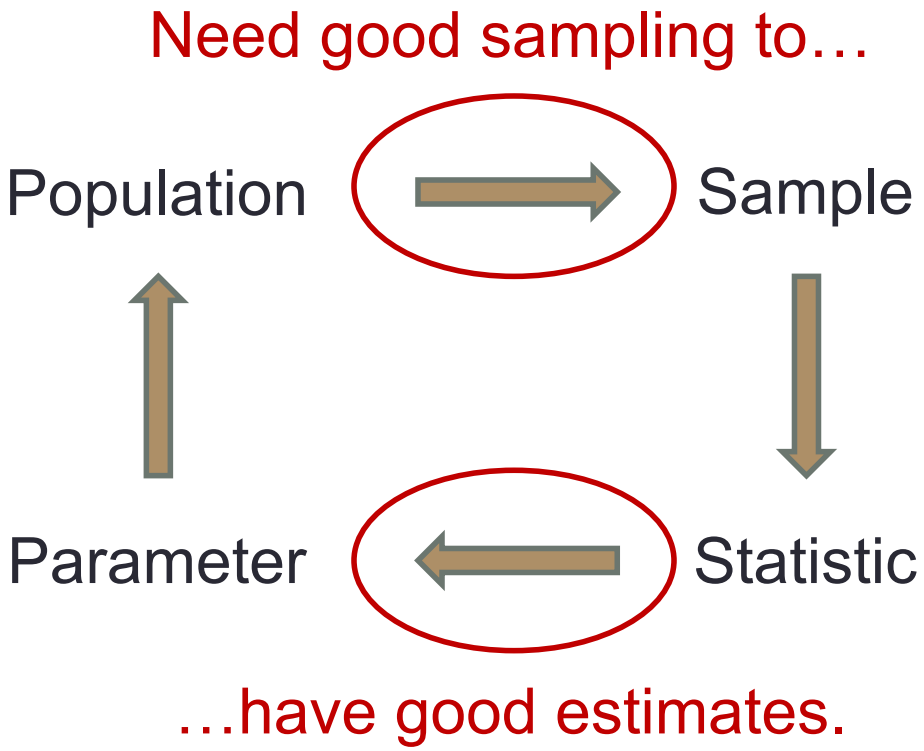
Parameters vs. Statistics



Parameters vs. Statistics



Parameters vs. Statistics



Sampling

- There are many different ways to sample data from population.
- Mistakes in sampling can lead to bias in the sample.

- **Bias** – certain outcomes are favored over other outcomes in samples.

Types of Bias

- **Bias** – certain outcomes are favored over other outcomes in samples.
- 2 Common Types of Bias:
  1. Selection Bias
  2. Sampling Bias

Types of Bias

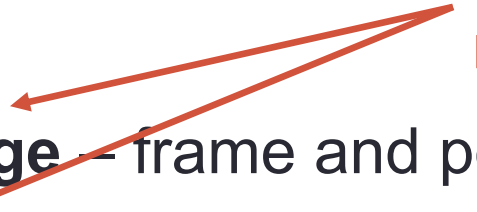
- **Bias** – certain outcomes are favored over other outcomes in samples.
- 2 Common Types of Bias:
  1. Selection Bias
  2. Sampling Bias

Types of Bias

- **Bias** – certain outcomes are favored over other outcomes in samples.
- 2 Common Types of Bias:
  1. Selection Bias
     a) **Undercoverage** – frame and population are not equal (ex. Phone book)
     b) **Nonresponse** – subject in sample cannot / will not respond or be measured (ex. Telemarketer)
  2. Sampling Bias

Types of Bias

- **Bias** – certain outcomes are favored over other outcomes in samples.
- 2 Common Types of Bias:
  1. Selection Bias
     a) **Undercoverage** – frame and population are not equal (ex. Phone book)
     b) **Nonresponse** – subject in sample cannot / will not respond or be measured (ex. Telemarketer)
  2. Sampling Bias



Types of Bias

- **Bias** – certain outcomes are favored over other outcomes in samples.
- 2 Common Types of Bias:
  1. Selection Bias
  2. Sampling Bias

Types of Bias

- **Bias** – certain outcomes are favored over other outcomes in samples.
- 2 Common Types of Bias:
  1. Selection Bias
  2. Sampling Bias
     a) **Convenience sampling** – technique that selects subjects from population based on accessibility and ease.
     b) **Voluntary sampling** – technique where subjects volunteer themselves to sample.

Statistical Techniques

- Statistical sampling techniques use selection methods based on chance selection instead of convenience or judgement.
- 4 Common Techniques:
  1. Simple Random Sampling (SRS)
  2. Stratified Random Sampling
  3. Cluster Sampling
  4. Systematic Sampling

Simple Random Sampling (SRS)

A method of sampling items from a population such that every possible sample of a specified size has an equal chance of being selected.

Simple Random Sampling (SRS)

- Artistic Constructions
- Constructing Wonders
- The Able Contractors
- Gorilla Builders
- Hammer Studios
- Rhino Builders
- Constructionals
- Constructive Partners
- The Remodelers
- Shaking Hand Builders
- Construction Agents
- We Build For U
- Conceptual Home Designs
- Inspired By Nature
- Natural Builders
- We Make Foundation
- Builder Brothers
- Built It
- Pro Builders
- Proof Modelers
- Blue Ladder Builders
- Heavenly Constructions
- Hammering Creativity
- Quality Certified
- The Premium Bricks
- Golden Bricks
- New Foundation
- High Voltage Foundation
- Engineering The World
- Power Home Builders
- Sunrise Builders
- Nailed It Contractors
- Eco-Fri Construction Co.
- The Pyramid Contractors
- We Build Pyramids
- Redesigning Creativity
- Remarkable Builders
- Success Constructions
- Sweet Sweet Home
- Evergreen Engineers
- Five Star Construction
- Well Being Builders
- Visionary Builders
- Builders Choice
- Wonder Makers
- Sparkling Constructions
- Sovereign Steels
- Maestro Builders
- Limited Edition Contractors
- Bossy Builders
- Tribal Contractors
- Jungle Projects
- Evergreen Renovations
- Chief Designs
- New View Constructions
- Builders
- Power Creators
- Rebuild Me
- Building Blocks
- Smart Roof
- Trusted Walls
- Eyeing For Builders
- Star Constructions
- Home Expert Builders
- Block At The Moon
- Building Buddy
- Brick Quick
- Concreting Planet
- The Invisibles
- Game of Builders
- The Throne Makers
- The Steel Foundation
- Building Buddies
- Urbanizing
- The Thor Hammers
- Skyscrapers Constructions
- Zooming Buildings
- Beauty Builders
- Ballistic Contractors
- Booked Builders
- Craning Contractors
- Big Bang Company
- Creative With Clay
- The Crown Contractors
- The Best Choice Builders
- Building The Nation
- Make Construction Great
- Re Structuring
- Tiles & Bricks
- Road Runners
- Diamond Construction
- The Owl Construction
- American Dream Builder
- Square Contractors
- Team of Brilliants
- Adam & Eve Constructions
- All The Way Homes
- The Desert Engineers
- Legions of Creatives

Sampled 20 of 99 construction companies!

A method of sampling items from a population such that every possible sample of a specified size has an equal chance of being selected.

Simple Random Sampling (SRS)

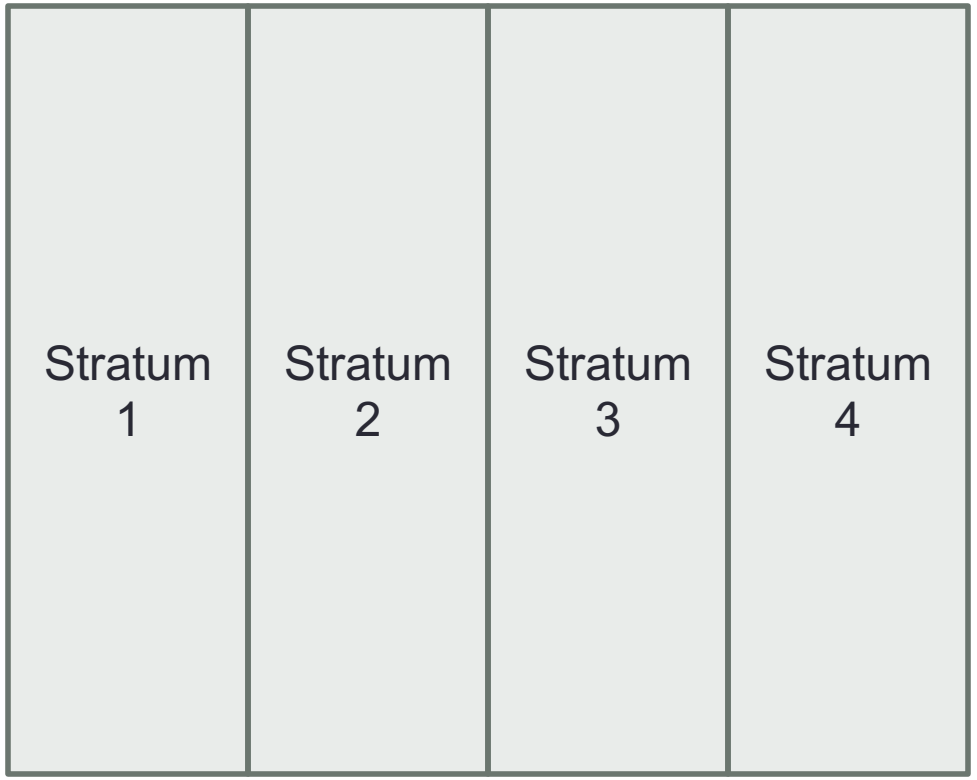
A method of sampling items from a population such that every possible sample of a specified size has an equal chance of being selected.

Advantages: No statistical bias, no previous information about sample needed ahead of time

Disadvantages: Expensive, time consuming, hard to implement, need list of population

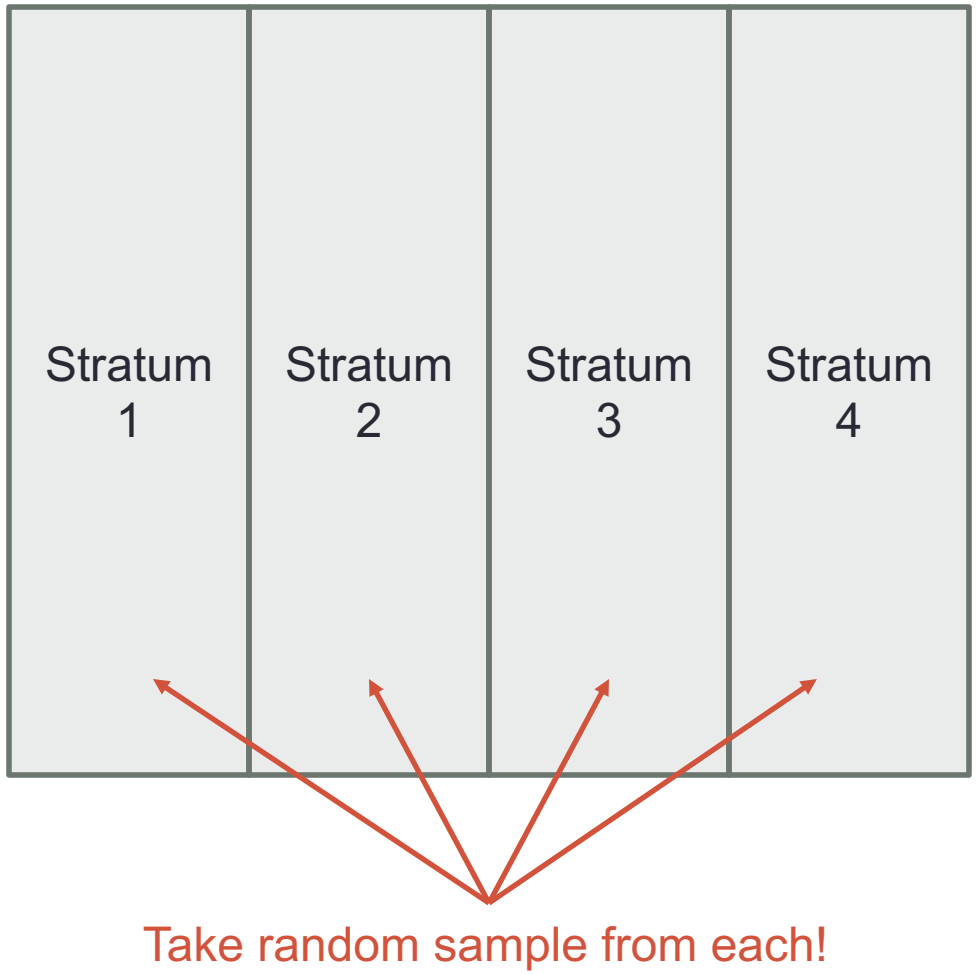
Stratified Random Sampling (STS)

A method of sampling items where the population is divided *a priori* into subgroups, called **strata**, so that each member in the population belongs to only one strata. Sample items from **every** strata (with SRS for example).



Stratified Random Sampling (STS)

A method of sampling items where the population is divided *a priori* into subgroups, called **strata**, so that each member in the population belongs to only one strata. Sample items from **every** strata (with SRS for example).



Stratified Random Sampling (STS)

A method of sampling items where the population is divided *a priori* into subgroups, called **strata**, so that each member in the population belongs to only one strata. Sample items from **every** strata (with SRS for example).

Advantages: Smaller sample sizes can achieve same accuracy as SRS, more information about parts of population

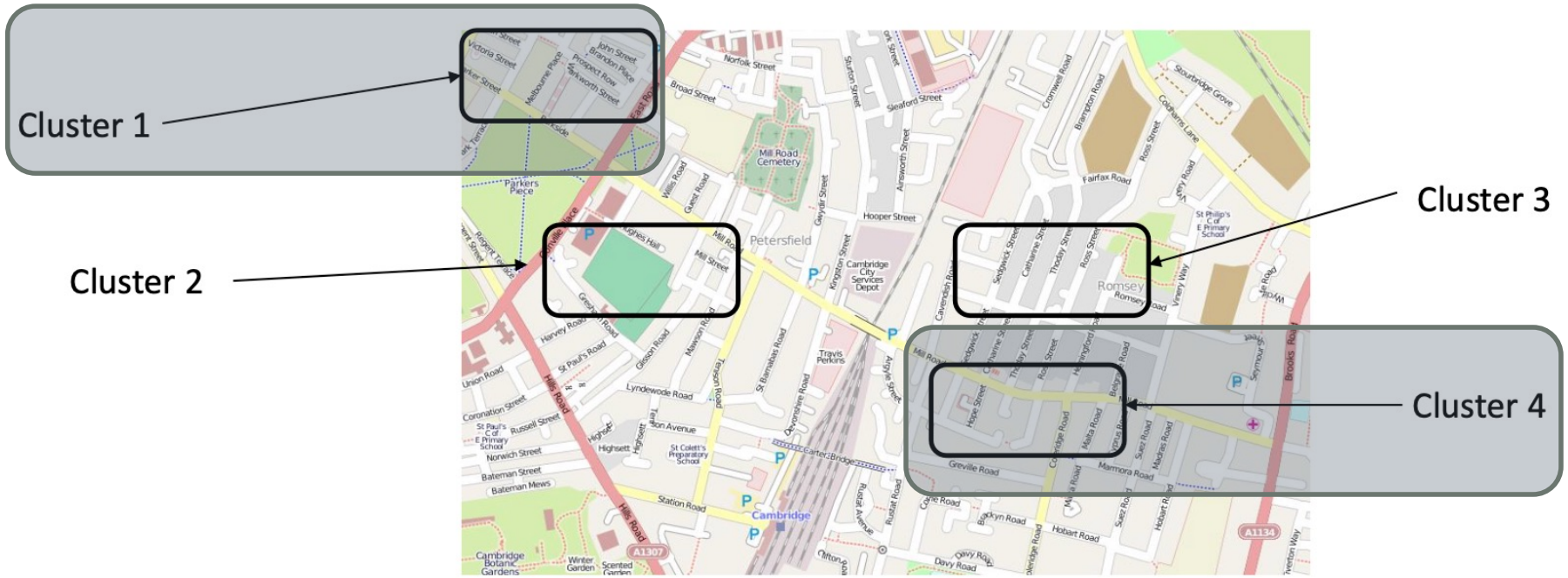
Disadvantages: Need information about population ahead of time to split on!

Cluster Sampling

A method of sampling items where the population is divided *a priori* into subgroups, called **clusters**, so that each member in the population belongs to only one cluster. Sample items from **a sample** of *m* clusters selected randomly.

Cluster Sampling

A method of sampling items where the population is divided *a priori* into subgroups, called **clusters**, so that each member in the population belongs to only one cluster. Sample items from **a sample** of *m* clusters selected randomly.



Cluster Sampling

A method of sampling items where the population is divided *a priori* into subgroups, called **clusters**, so that each member in the population belongs to only one cluster. Sample items from **a sample** of *m* clusters selected randomly.

Advantages: Overcome issues with travel, time, and expense; Easier to implement than SRS or STS

Disadvantages: Need information about population ahead of time to split on – but not total list!; May have slight bias if random clusters aren't representative

Systematic Sampling

- Artistic Constructions
- Constructing Wonders
- The Able Contractors
- Gorilla Builders
- Hammer Studios
- Rhino Builders
- Constructionals
- Constructive Partners
- The Remodelers
- Shaking Hand Builders
- Construction Agents
- We Build For U
- Conceptual Home Designs
- Inspired By Nature
- Natural Builders
- We Make Foundation
- Builder Brothers
- Built It
- Pro Builders
- Proof Modelers
- Blue Ladder Builders
- Heavenly Constructions
- Hammering Creativity
- Quality Certified
- The Premium Bricks
- Golden Bricks
- New Foundation
- High Voltage Foundation
- Engineering The World
- Power Home Builders
- Sunrise Builders
- Nailed It Contractors
- Eco-Fri Construction Co.
- The Pyramid Contractors
- We Build Pyramids
- Redesigning Creativity
- Remarkable Builders
- Success Constructions
- Sweet Sweet Home
- Evergreen Engineers
- Five Star Construction
- Well Being Builders
- Visionary Builders
- Builders Choice
- Wonder Makers
- Sparkling Constructions
- Sovereign Steels
- Maestro Builders
- Limited Edition Contractors
- Bossy Builders
- Tribal Contractors
- Jungle Projects
- Evergreen Renovations
- Chief Designs
- New View Constructions
- Builders
- Power Creators
- Rebuild Me
- Building Blocks
- Smart Roof
- Trusted Walls
- Eyeing For Builders
- Star Constructions
- Home Expert Builders
- Block At The Moon
- Building Buddy
- Brick Quick
- Concreting Planet
- The Invisibles
- Game of Builders
- The Throne Makers
- The Steel Foundation
- Building Buddies
- Urbanizing
- The Thor Hammers
- Skyscrapers Constructions
- Zooming Buildings
- Beauty Builders
- Ballistic Contractors
- Booked Builders
- Craning Contractors
- Big Bang Company
- Creative With Clay
- The Crown Contractors
- The Best Choice Builders
- Building The Nation
- Make Construction Great
- Re Structuring
- Tiles & Bricks
- Road Runners
- Diamond Construction
- The Owl Construction
- American Dream Builders
- Square Contractors
- Team of Brilliants
- Adam & Eve Constructions
- All The Way Homes
- The Desert Engineers
- Legions of Creatives

A method of sampling items that involves selecting every k^{th} item in the population after randomly selecting a starting point between 1 and k. The value k is determined as the ratio of the population size over the desired sample size.

Systematic Sampling

A method of sampling items that involves selecting every k^{th} item in the population after randomly selecting a starting point between 1 and k. The value k is determined as the ratio of the population size over the desired sample size.

Systematic Sampling

- Artistic Constructions
- Constructing Wonders
- The Able Contractors
- Gorilla Builders
- Hammer Studios
- Rhino Builders
- Constructionals
- Constructive Partners
- The Remodelers
- Shaking Hand Builders
- Construction Agents
- We Build For U
- Conceptual Home Designs
- Inspired By Nature
- Natural Builders
- We Make Foundation
- Builder Brothers
- Built It
- Pro Builders
- Proof Modelers
- Blue Ladder Builders
- Heavenly Constructions
- Hammering Creativity
- Quality Certified
- The Premium Bricks
- Golden Bricks
- New Foundation
- High Voltage Foundation
- Engineering The World
- Power Home Builders
- Sunrise Builders
- Nailed It Contractors
- Eco-Fri Construction Co.
- The Pyramid Contractors
- We Build Pyramids
- Redesigning Creativity
- Remarkable Builders
- Success Constructions
- Sweet Sweet Home
- Evergreen Engineers
- Five Star Construction
- Well Being Builders
- Visionary Builders
- Builders Choice
- Wonder Makers
- Sparkling Constructions
- Sovereign Steels
- Maestro Builders
- Limited Edition Contractors
- Bossy Builders
- Tribal Contractors
- Jungle Projects
- Evergreen Renovations
- Chief Designs
- New View Constructions
- Builders
- Power Creators
- Rebuild Me
- Building Blocks
- Smart Roof
- Trusted Walls
- Eyeing For Builders
- Star Constructions
- Home Expert Builders
- Block At The Moon
- Building Buddy
- Brick Quick
- Concreting Planet
- The Invisibles
- Game of Builders
- The Throne Makers
- The Steel Foundation
- Building Buddies
- Urbanizing
- The Thor Hammers
- Skyscrapers Constructions
- Zooming Buildings
- Beauty Builders
- Ballistic Contractors
- Booked Builders
- Craning Contractors
- Big Bang Company
- Creative With Clay
- The Crown Contractors
- The Best Choice Builders
- Building The Nation
- Make Construction Great
- Re Structuring
- Tiles & Bricks
- Road Runners
- Diamond Construction
- The Owl Construction
- American Dream Builders
- Square Contractors
- Team of Brilliants
- Adam & Eve Constructions
- All The Way Homes
- The Desert Engineers
- Legions of Creatives

A method of sampling items that involves selecting every k^{th} item in the population after randomly selecting a starting point between 1 and k. The value k is determined as the ratio of the population size over the desired sample size.

Systematic Sampling

A method of sampling items that involves selecting every k^{th} item in the population after randomly selecting a starting point between 1 and k. The value k is determined as the ratio of the population size over the desired sample size.

Advantages: Very easy to get sample

Disadvantages: May be biased, especially if order of list of population matters

Example

- A large worldwide financial company wants to develop a new retirement plan for the company. They want to survey different managers of branches around the world to find out the most important strategies the new retirement plan should contain. They have 5000 branches worldwide and want to personally interview these branch managers. They have information about the branch size (small, medium, large), and the state/province location of the branch. They want to talk to 50 branch managers.

- Develop four separate strategies to sample these branch managers based on the four different statistical sampling techniques discussed previously.

Example

- Develop four separate strategies to sample these branch managers based on the four different statistical sampling techniques discussed previously.
  1. SRS – Randomly sample 50 branches to interview their managers
  2. STS – Stratify by size and select SRS from each
  3. Cluster – Randomly select sample of states/provinces, then select branches at random from those states/provinces
  4. Systematic – Select every 100^{th} branch in list of branches

TYPES OF DATA

4 Types of Data

- There are four main types of data people typically deal with in data analysis.
- These four types are split into two groups
  1. Qualitative vs. Quantitative
  2. Time Series vs. Cross-sectional

Quantitative vs. Qualitative

- **Quantitative**:
  - Data that are numeric that define value or quantity.
  - Easy check → Must be able to do basic arithmetic and have it make sense.

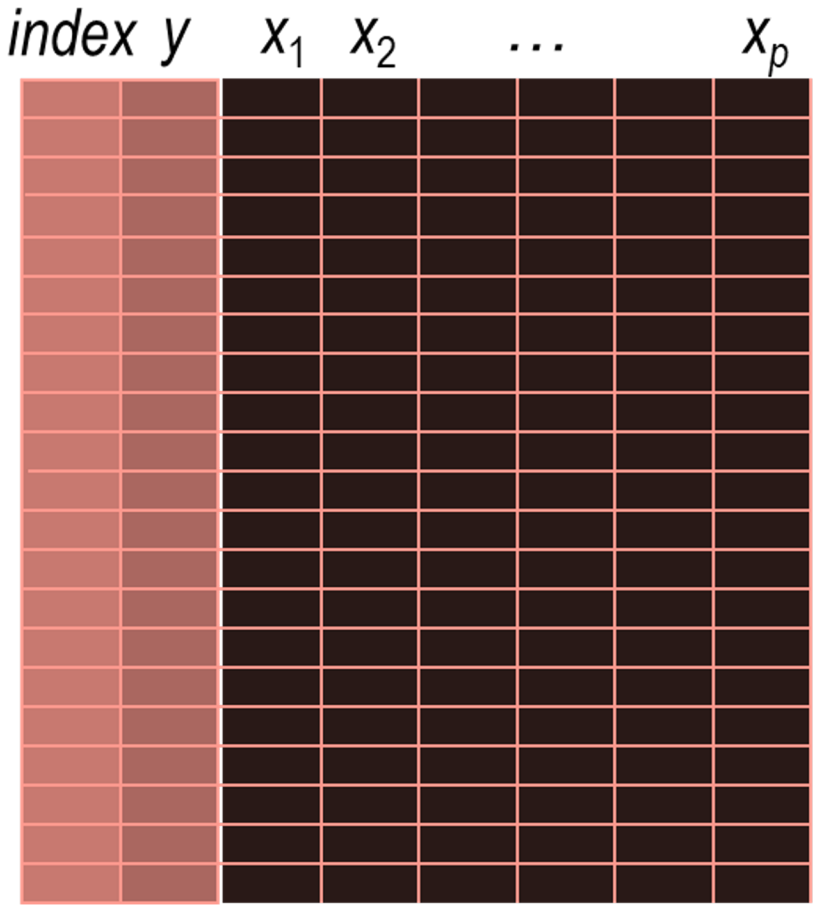
- **Qualitative**:
  - Data whose measurement scale is inherently categorical.
  - **Nominal** – categories with no logical ordering
  - **Ordinal** – categories with a logical order / only two ways to order the categories (binary IS ordinal)

Time Series vs. Cross-sectional

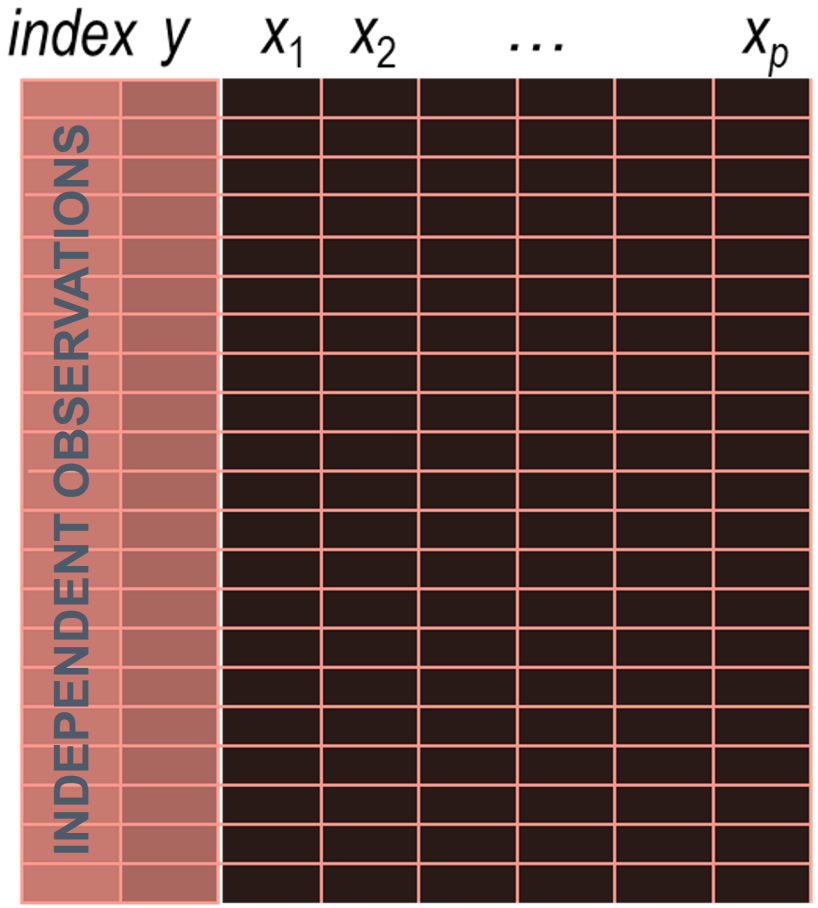
- **Time Series**:
  - Set of ordered data values observed at successive points in time.

- **Cross-sectional**:
  - Set of data values observed at a fixed point in time, or where time is of no significance.

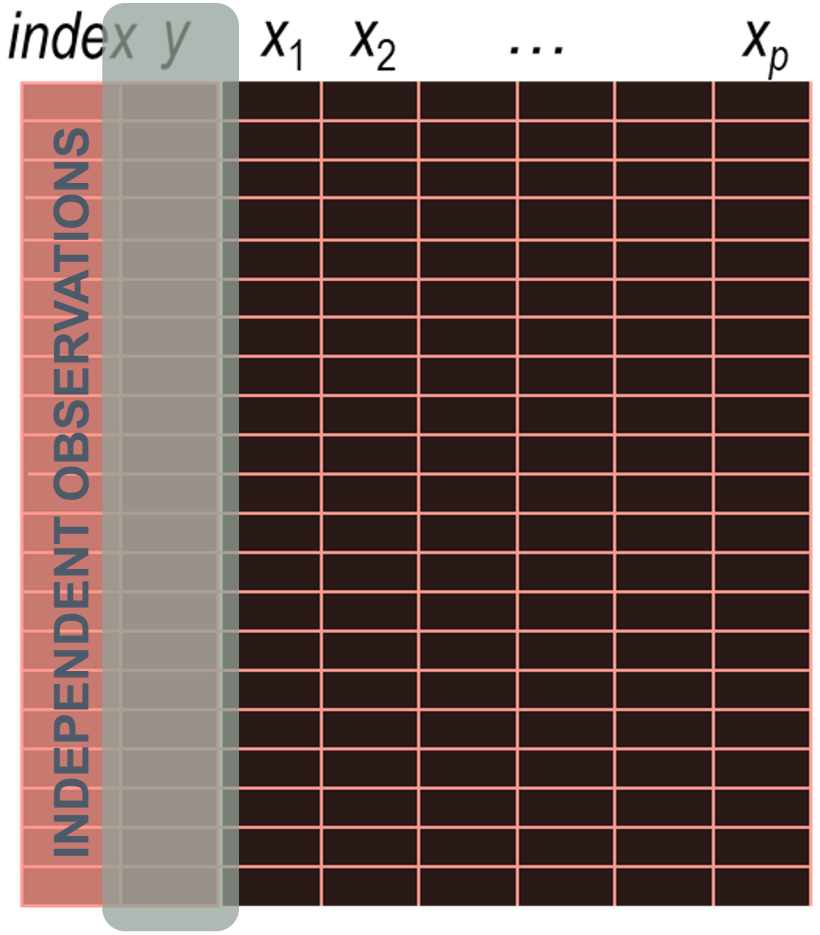
Cross-sectional

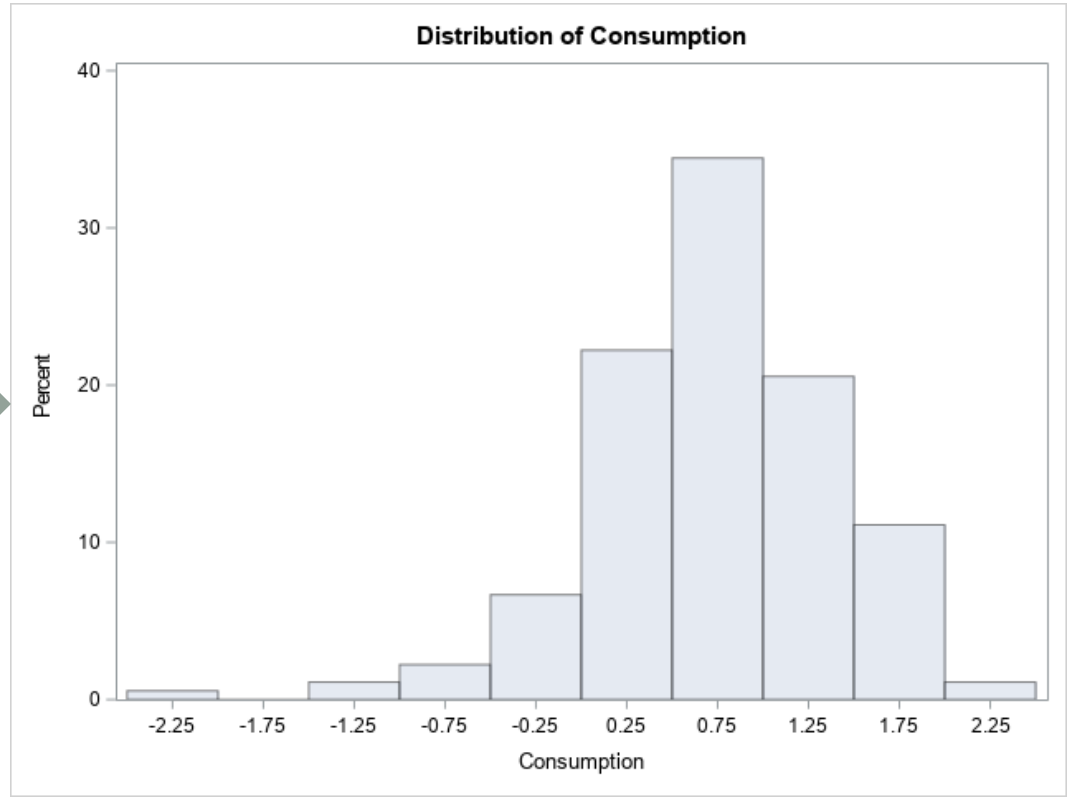


Cross-sectional

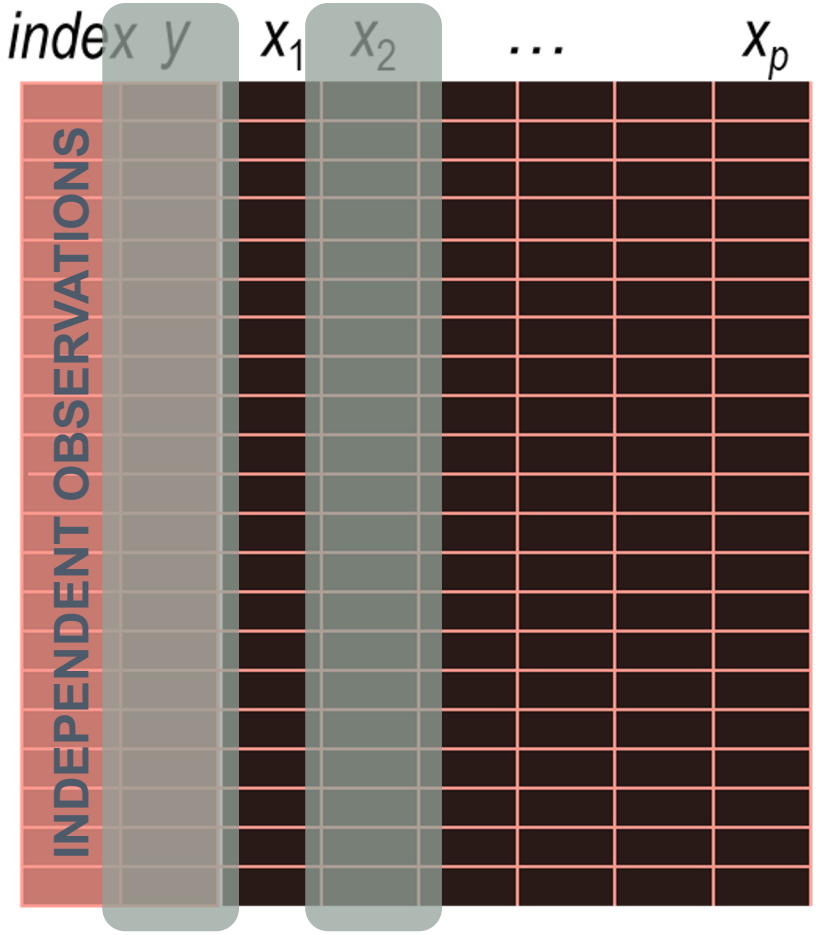


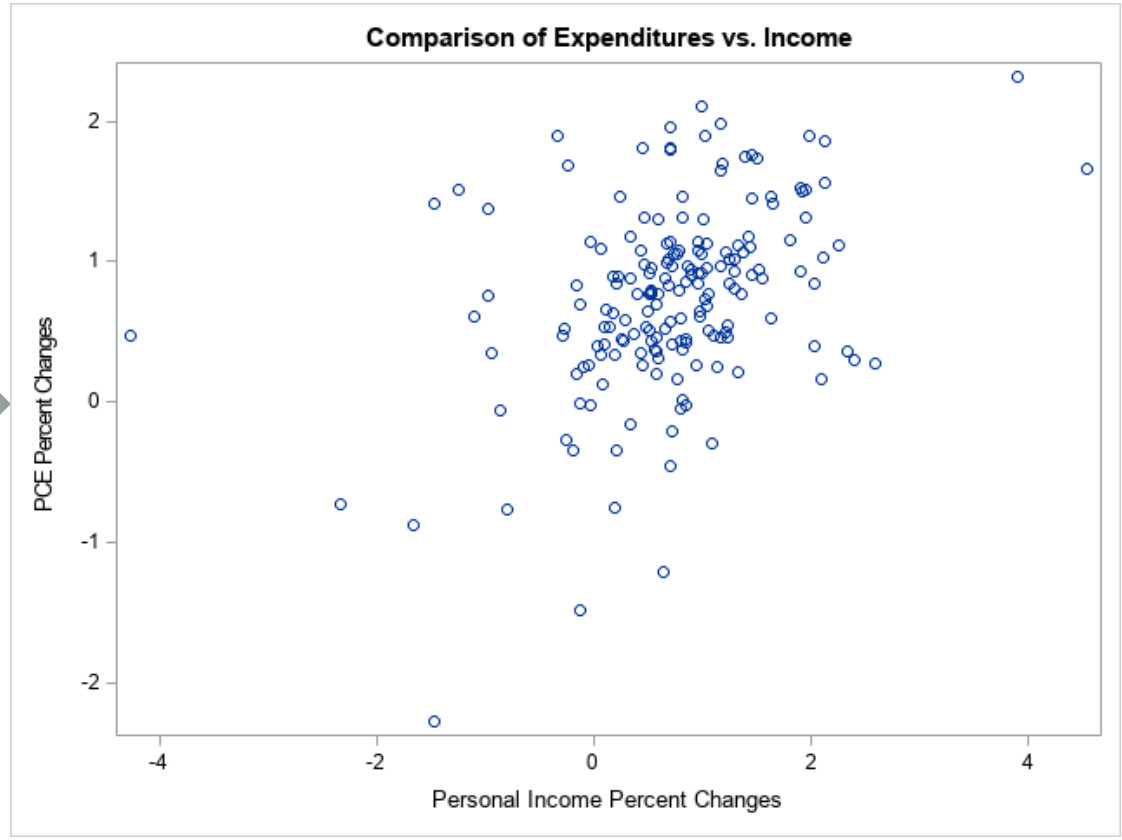
Cross-sectional



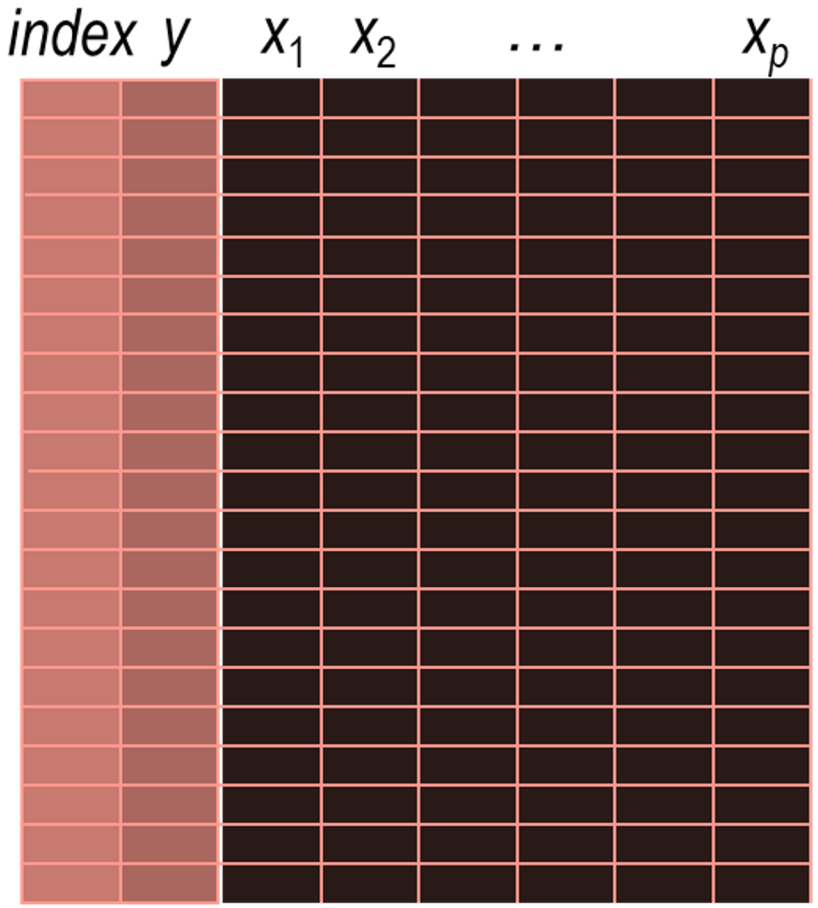


Cross-sectional

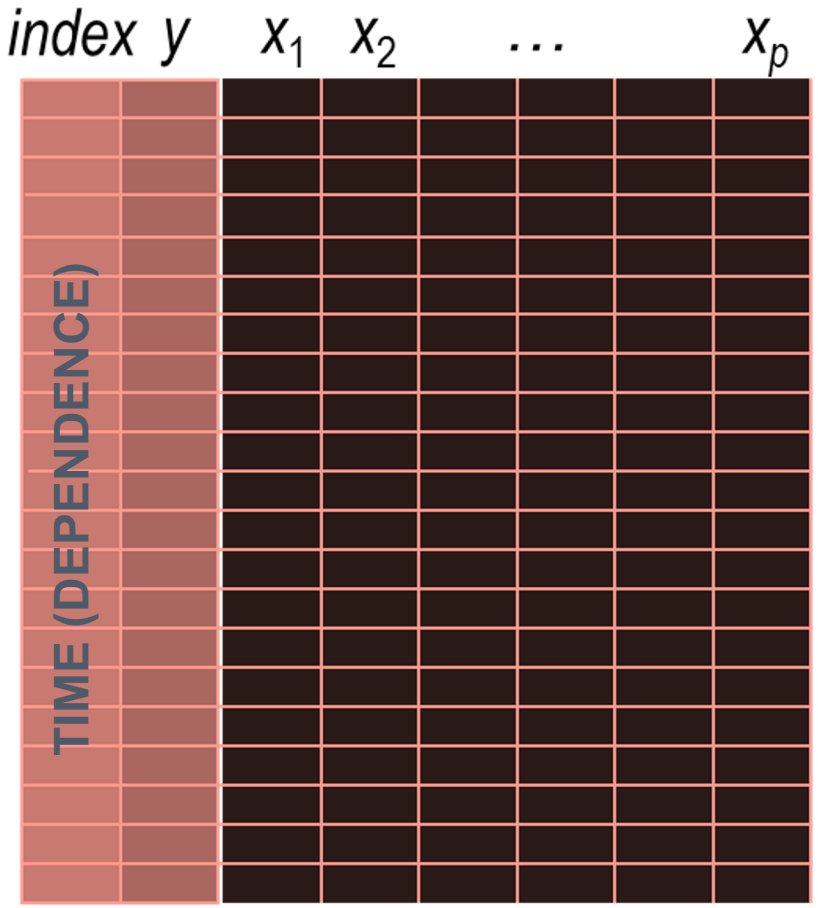




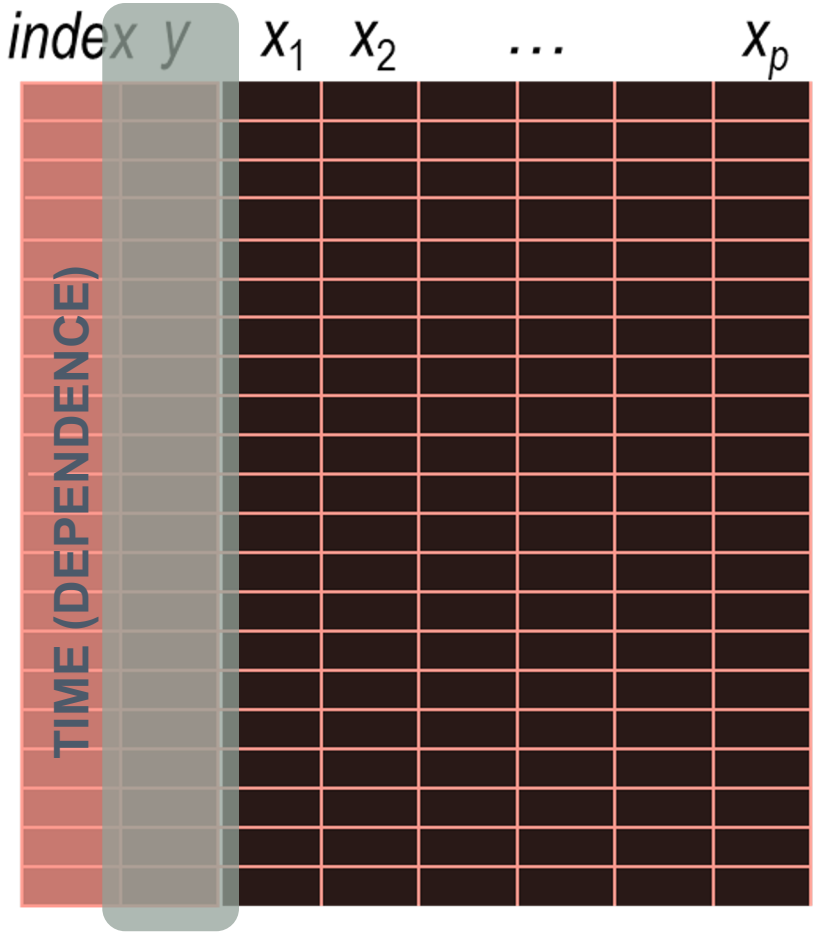
Time Series

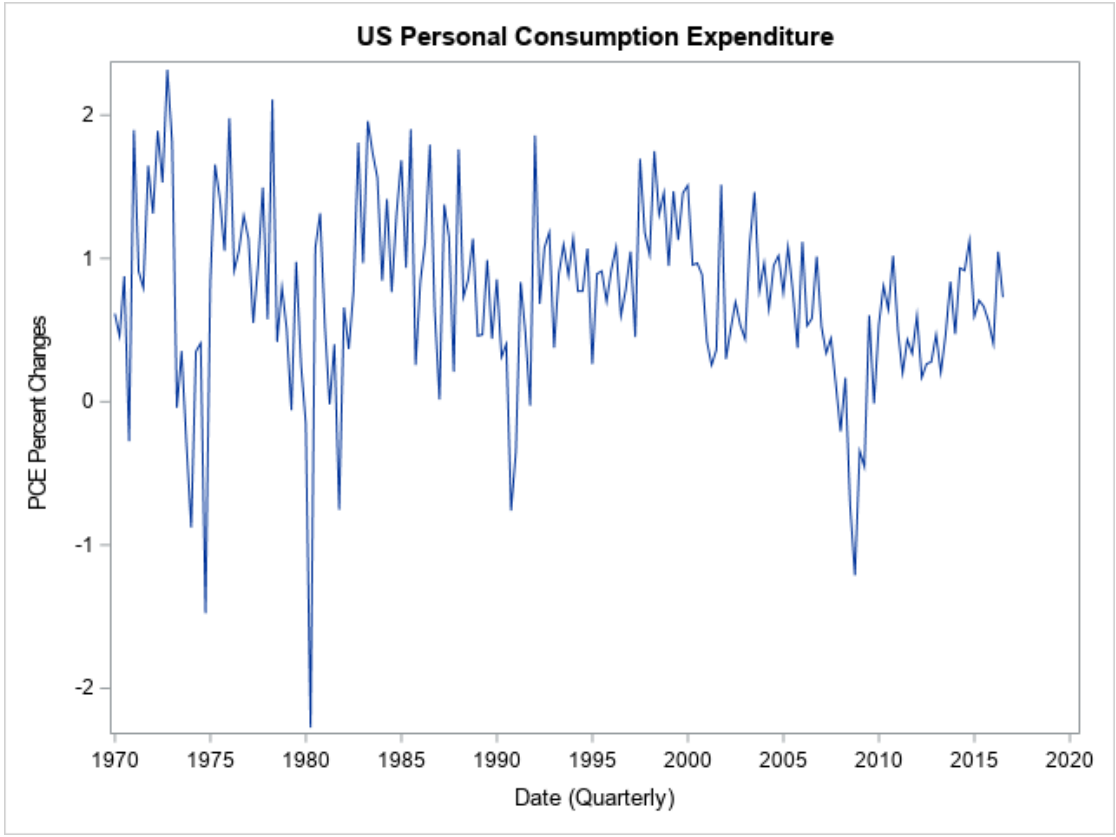


Time Series

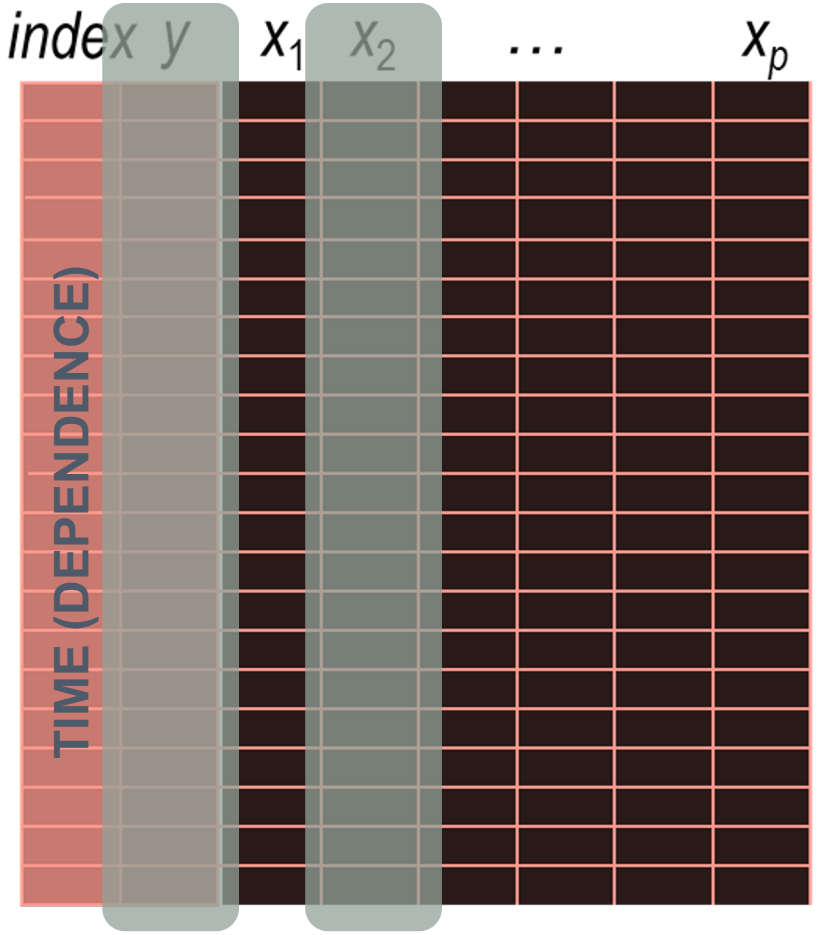


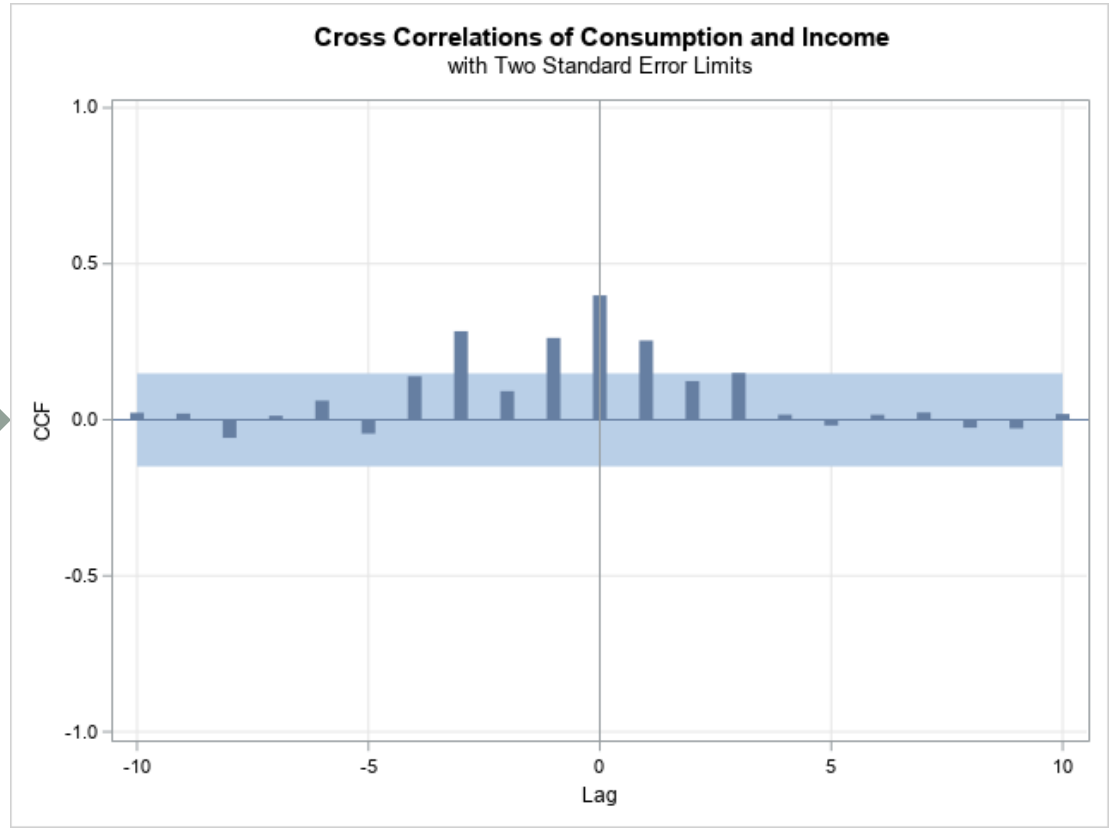
Time Series


index y
x_1
x_2
...
x_p
TIME (DEPENDENCE)


US Personal Consumption Expenditure
PCE Percent Changes
2
1
0
-1
-2
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
Date (Quarterly)

Time Series


index
y
x_1
x_2
...
x_p
TIME (DEPENDENCE)


Cross Correlations of Consumption and Income
with Two Standard Error Limits
CCF
1.0
0.5
0.0
-0.5
-1.0
-10
-5
0
5
10
Lag