NRES 776 Lecture 2

Descriptive statistics and sampling methods

Sunny Tseng

Our first (official) lecture!

Our schedule today

- Announcement (5 min)
- Population vs Sample (10 min)
- Descriptive statistics (20 min)
- Sampling methods (10 min)
- Wrap up (5 min)

Announcement

- Considering keep the video on (but you have absolute right for privacy)!
- Will be recording the lecture today
- UNBC Applied Analysis Hub
- Workshop: use R to create study area map (Sep. 19th 15:00 - 16:30pm)
- How was your week? smoke, school life?

Why we need statistics

To make our life harder?

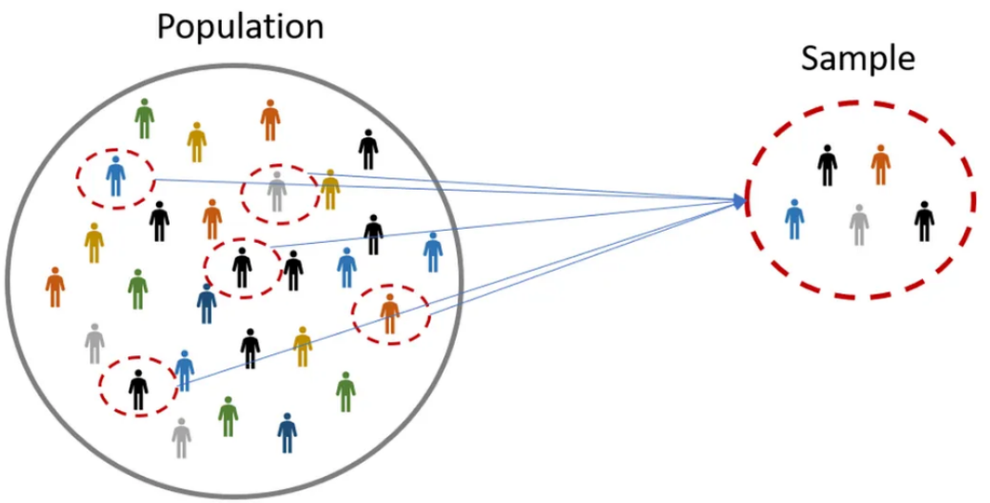
Population is often…

Too big, too many, and unrealistic to measure

- How many moose are there in the research forest?
  - Find all the moose, and count them, in the forest
  - Mark-recapture **estimate**
- How many bird species are in mature forest versus clear cut?
  - Stand there 24 hours, 365 days, and count all the bird species
  - Standardized point counts for **estimation** on both habitats

We need samples to …

Get more understanding about the population



Statistics: a branch of mathematics dealing with collection, analysis, interpretation, presentation of (sampled) data

- We can make sure the estimation is close enough to the true value
- So that we can make actions in real life (conservation, medical purposes)

Descriptive statistics

Summarize data into several numbers to provide you with an idea what the data look like.

Attributes in Population vs Sample

Population parameters

(Greek letters)

- mean (μ)
- standard deviation (σ)
- variance (σ^2)
- correlation coefficient (ρ)
- number of elements (N)
- …

Sample statistics

(Roman letters)

- mean ($\bar{x}$)
- standard deviation (s)
- variance (s^2)
- correlation coefficient (r)
- number of elements (n)
- …

Mean

Or average, refers to the central value of a set of numbers. The most common one is the Arthmetic mean

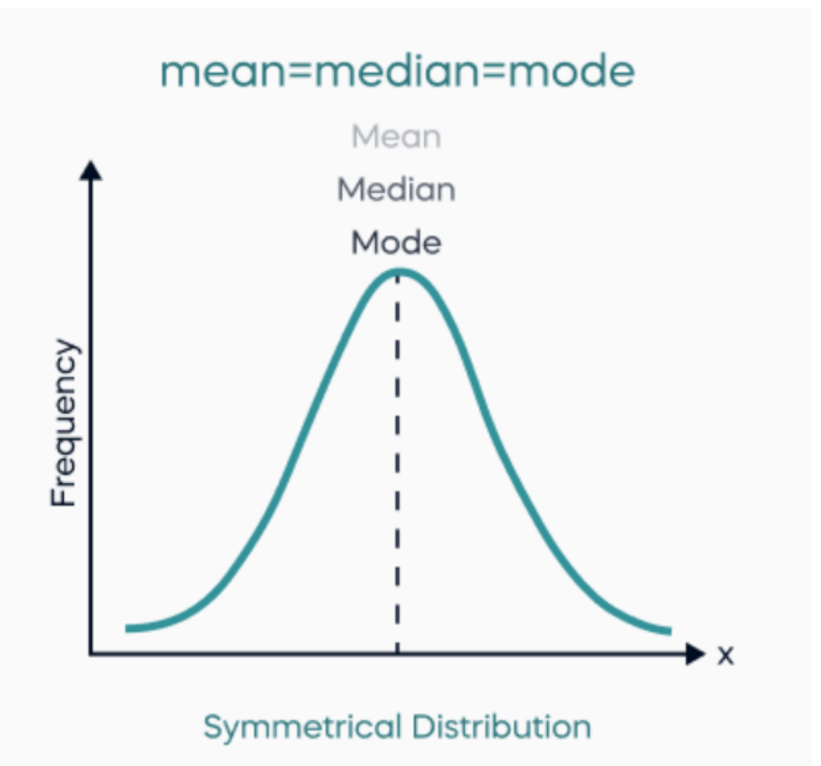
$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- unit: the same as x
- population parameter: μ

Median

The number separating the higher half of a set of data from the lower half (50% of data above the median; 50% lower)

In the set of data `1, 2, 3, 5, 7, 8, 9`, the median is `5`

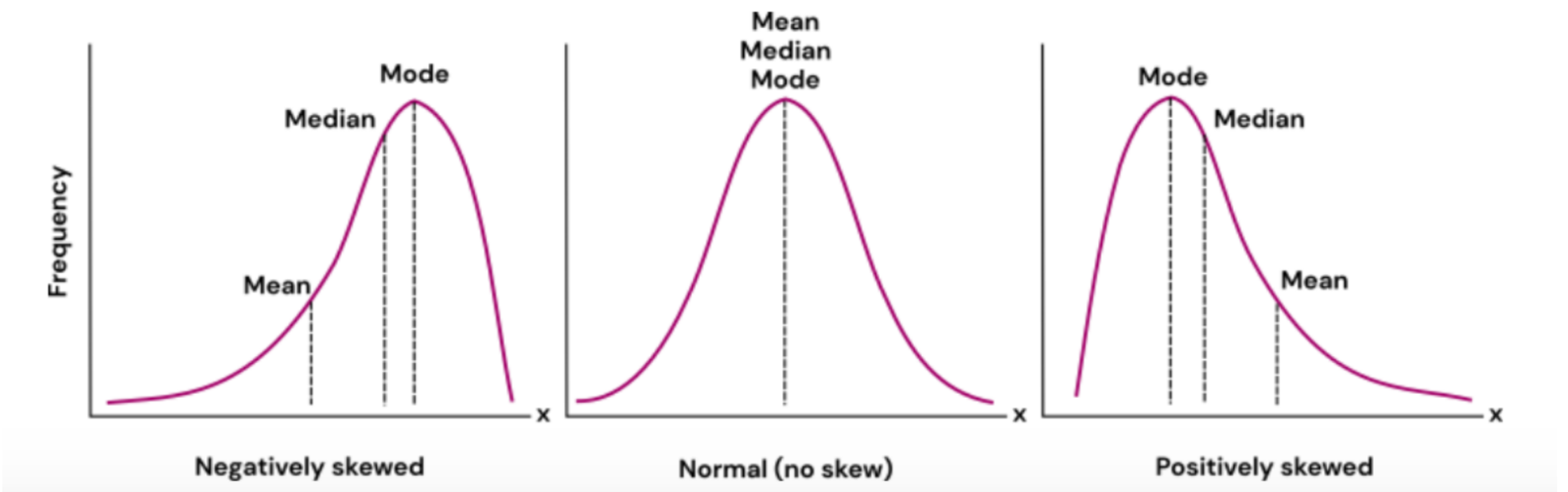


Mode

The value that appears with highest frequency

In the set of data `3, 4, 4, 5, 4, 6, 7, 8`, the mode value is `4`

- unit: the same as x

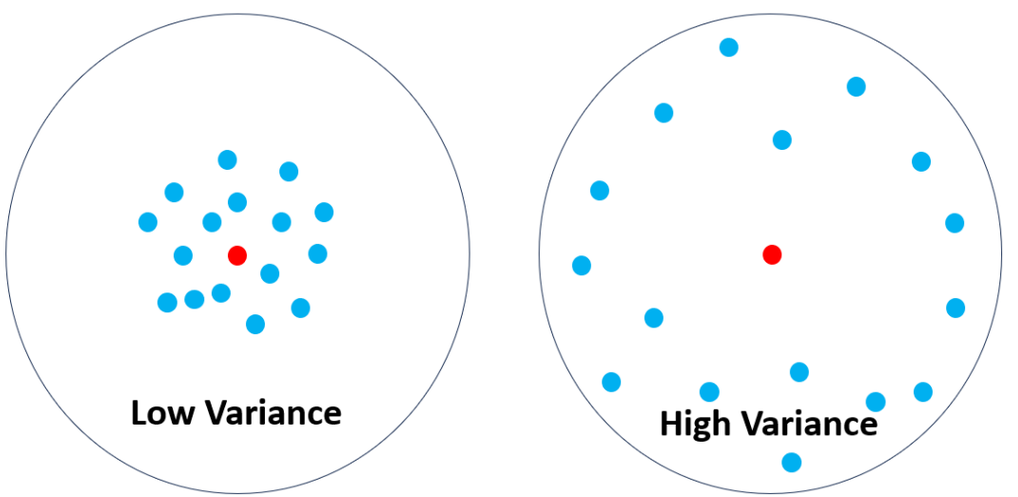


Variance

Measures how far a set of numbers is spread out

$$var(x) = S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

1. $(x_i - \bar{x})$: distance between each value and mean
2. square: make everything positive
3. sum: total distance square
4. $n - 1$: mean distance square



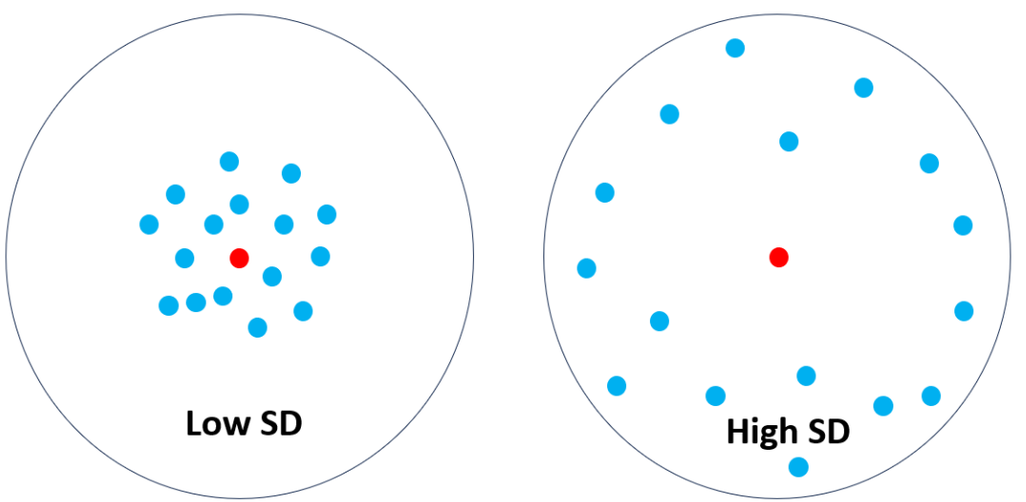
- unit: squared x
- population parameter: σ^2

Standard deviation

Square root of variance (i.e., mean distance between each value and mean)

$$S_x = \sqrt{\frac{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- unit: the same as x
- population parameter: σ

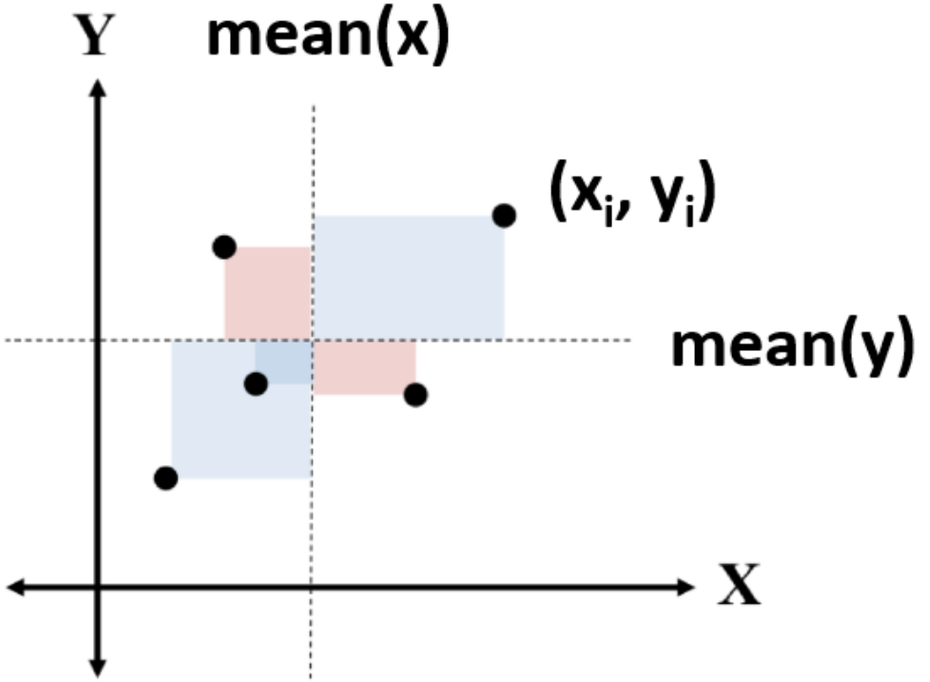


Covariance

A measure of how much two sets of data associated

$$cov(x, y) = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

- similar to the equation of the variance!
- sum of the area, positive and negative
- very positive: positively related
- very negative: negatively related
- 0: no linear relationship
- hard to tell the magnitude of relationship because of the unit difference

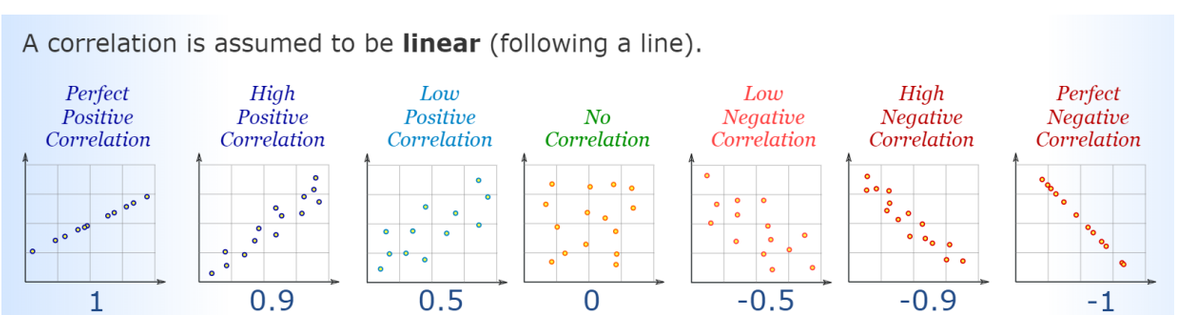


Correlation coefficient

Standardized covariance

$$r(x,y) = \frac{cov(x,y)}{S_x S_y} = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}}$$

- unit: no unit
- population parameter: ρ
- **value always between 1 to -1**

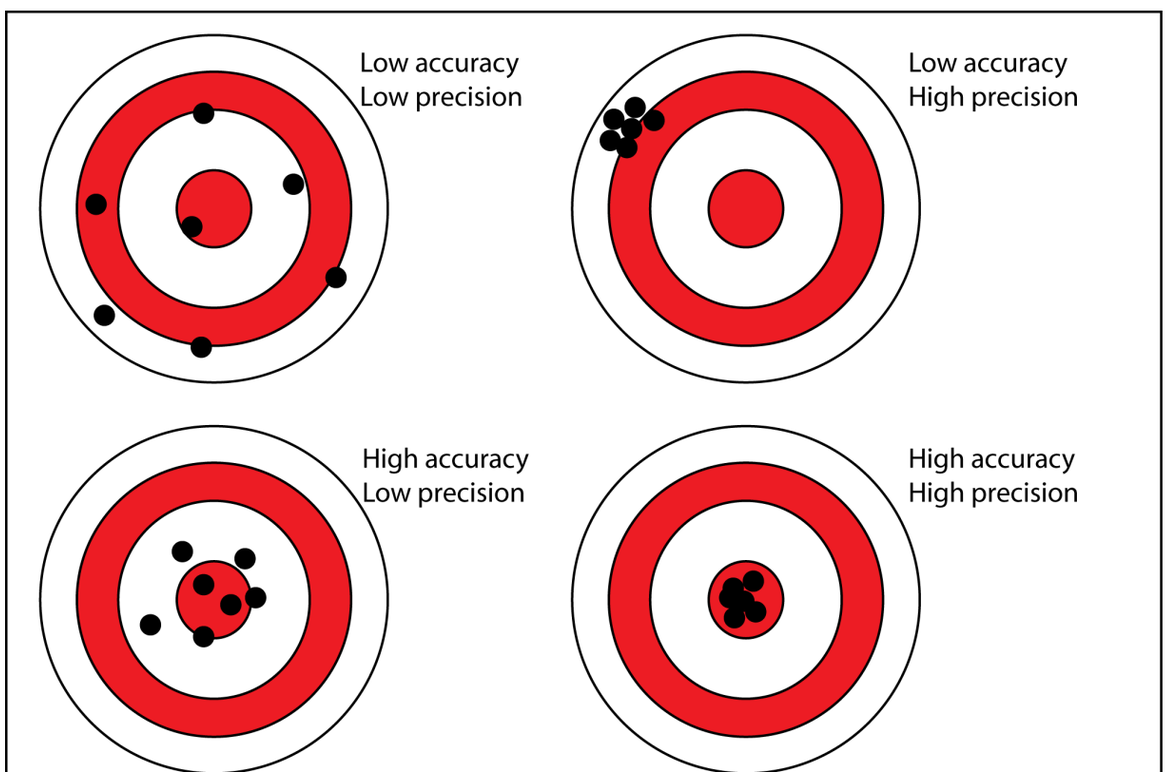


Magic of sampling

How to select samples from a population to make the sampling statistics close to the population parameters?

Accuracy and precision

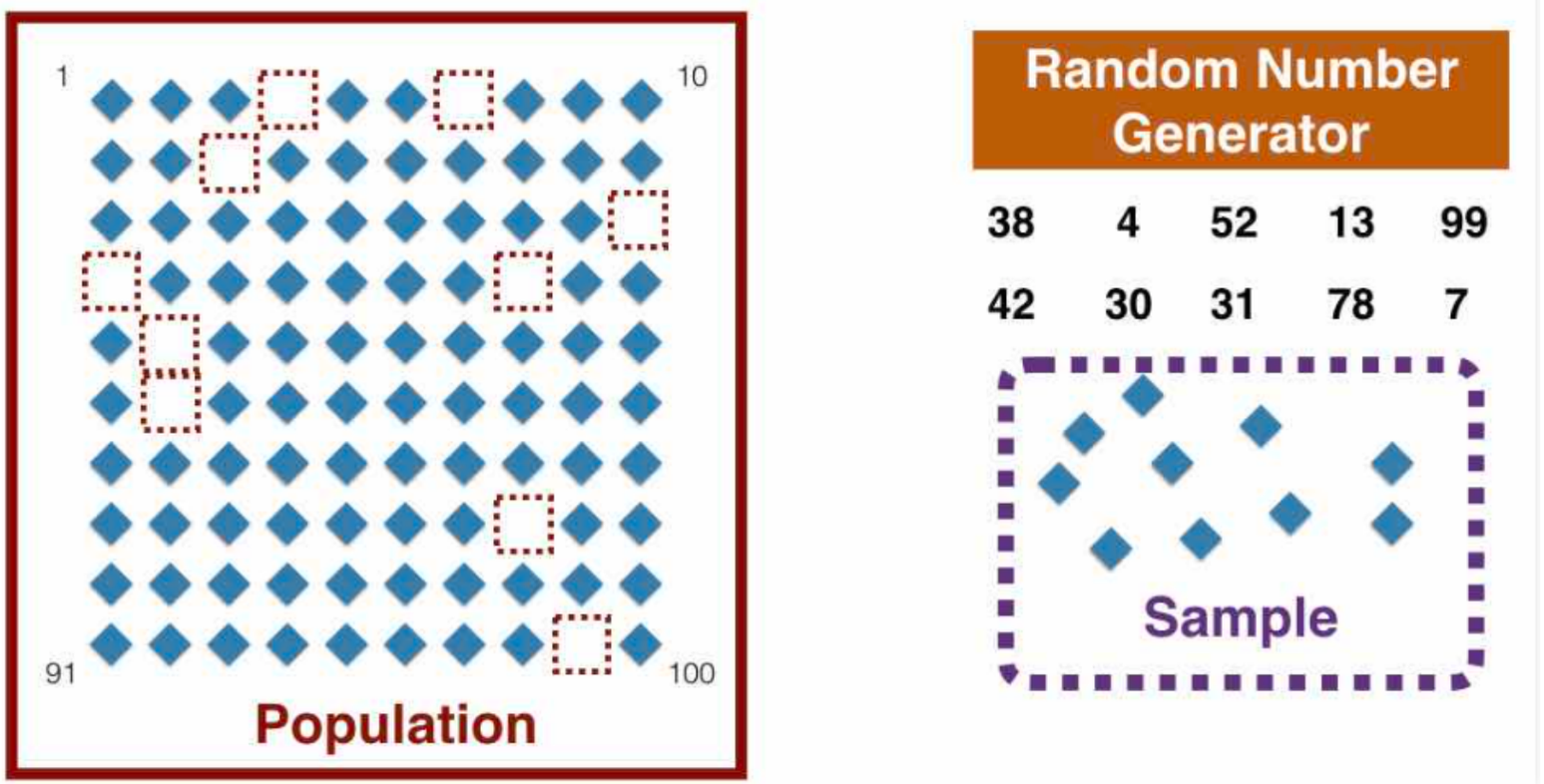
Accuracy refers to how close a measurement is to the true value. Precision refers to how close measurements of the same item are to each other.



We need a sampling method that is both accurate and precise.

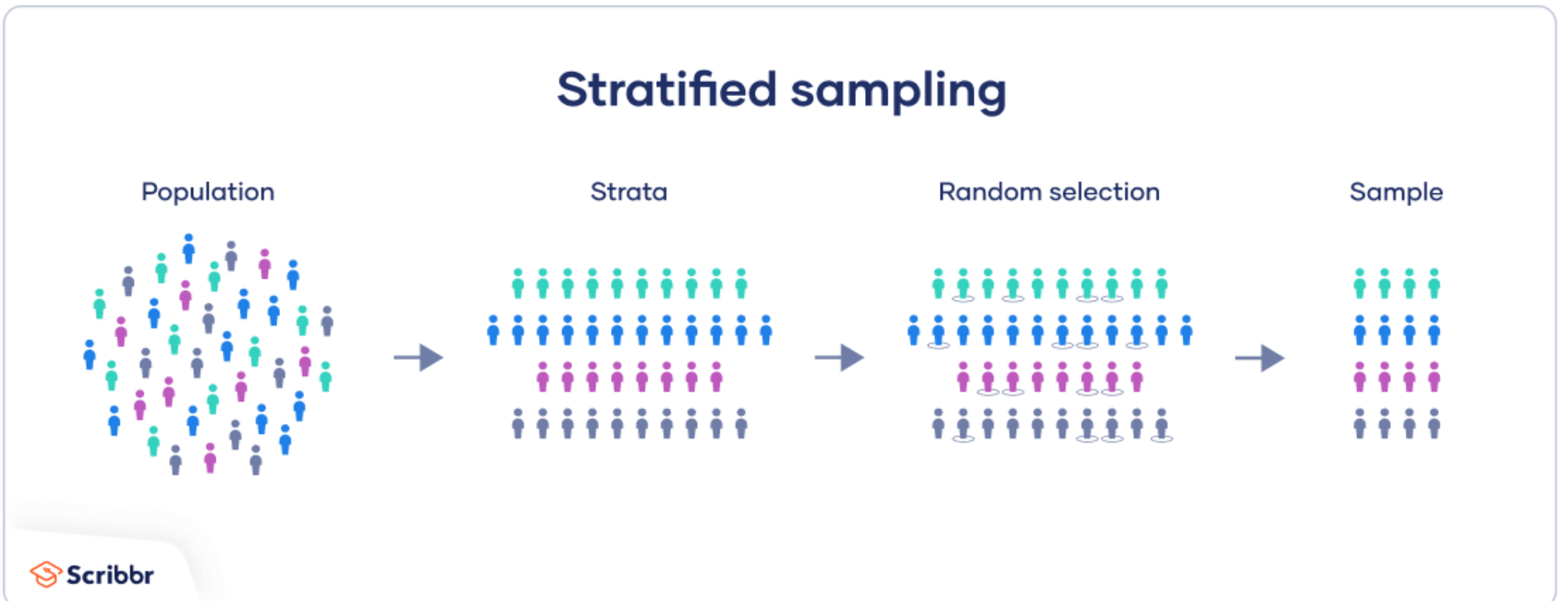
Simple random sampling

A subset of a statistical population where each unit has an equal probability of being chosen.
Each unit has independent identical distribution (i.i.d)



Stratified sampling

Stratification is the process of dividing members of the population into homogeneous subgroups before sampling.



When to use what?

Simple random sampling

When the population is relatively homogeneous or uniform

- Select 20 sites from a forest, regardless of their forest type
- Select 20 recordings from a day, regardless of the time of recording
- Select 40 trees from a plot, regardless of their species

Stratified sampling

When a population’s characteristics are diverse and you want to ensure that every characteristic is properly represented in the sample.

- Select 20 sites out of 4 forest types
- Select 20 recordings from morning and night
- Select 40 trees with equal number of trees of each species

What we learned

- Population (parameters) vs Sample (statistics)
- Descriptive statistics
- Different types of sampling

Wrap up

Before we meet again

- Install R (Window, Mac) and RStudio
- Play with RStudio to make sure it opens properly
- Sign up for discussion paper presentation
- Or I will do it for you 😀

Next time

- Thur. 8am lab, virtual on zoom
- With your morning coffee/tea and relaxed mood