DataSci 520

lesson 5

sampling and the central limit theorem

PROFESSIONAL & CONTINUING EDUCATION

UNIVERSITY of WASHINGTON

today's agenda

- sample vs population
- sources of uncertainty
- sampling methods
- the law of large numbers
- the Central Limit Theorem

Sampling

Sampling is the process of selecting a small number of elements from a larger defined target group

(Population)

of elements such that the information gathered from the small group will allow judgments to be made about the larger groups.

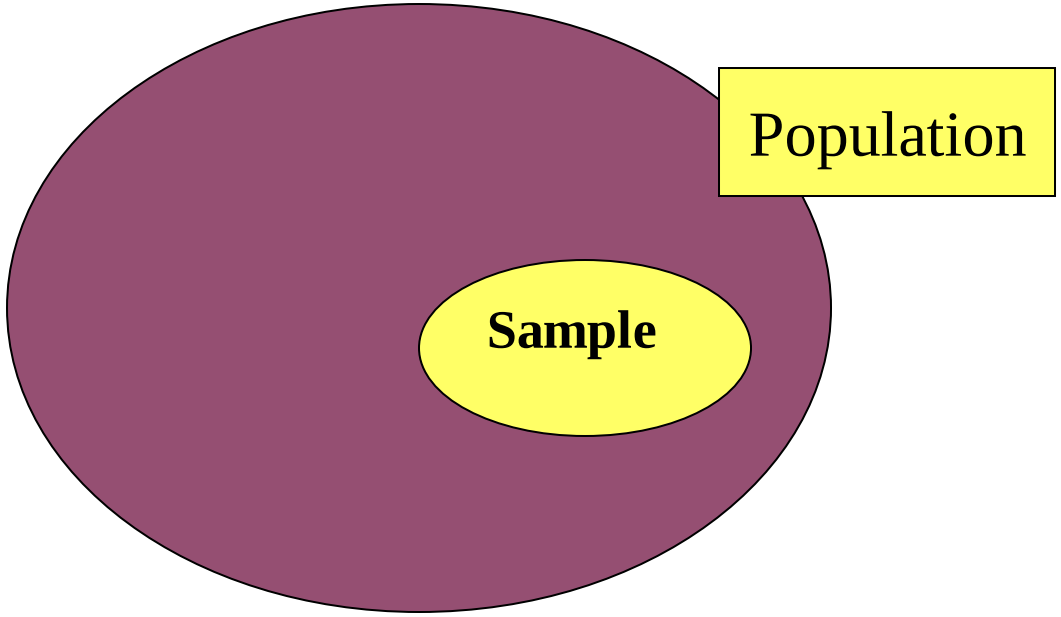
Sampling is the act, process, or technique of selecting a suitable sample, or a representative part of a population for the purpose of determining parameters or characteristics of the whole population.

Why Sampling?

- Empirical testing as the touchstone of scientific method
- To test hypothesis in the ‘real’ world with actual observations
- Observations come from a smaller set of individuals
- Can these observations lead to reliable and valid conclusions?
  - Only if selection process is accurate

Important terminologies

- **Population**: is any well-defined set of units of analysis: people, countries, events, years
- **Sample**: is any subset of units collected in some manner from the population



sample vs population

- a **population** is a group of things / people we are targeting
- there is one population and it is usually very large, even infinite
  - all our current and future customers' age when they joined our app
  - all normally distributed numbers with mean μ and standard deviation σ
- we call μ and σ **population parameters** and they are **fixed** (unless you're a Bayesian, but we return to this later)
- a **sample** is a smaller subset of the population (usually much smaller)
  - the age of the customers I have in the database *right now*
  - n numbers drawn from $N(\mu, \sigma)$ (or any other distribution)
- we can compute **statistics** for our sample, such as the sample mean $\bar{x}$ and the sample standard deviation s, but statistics **vary** because samples vary

Sample Vs. Population

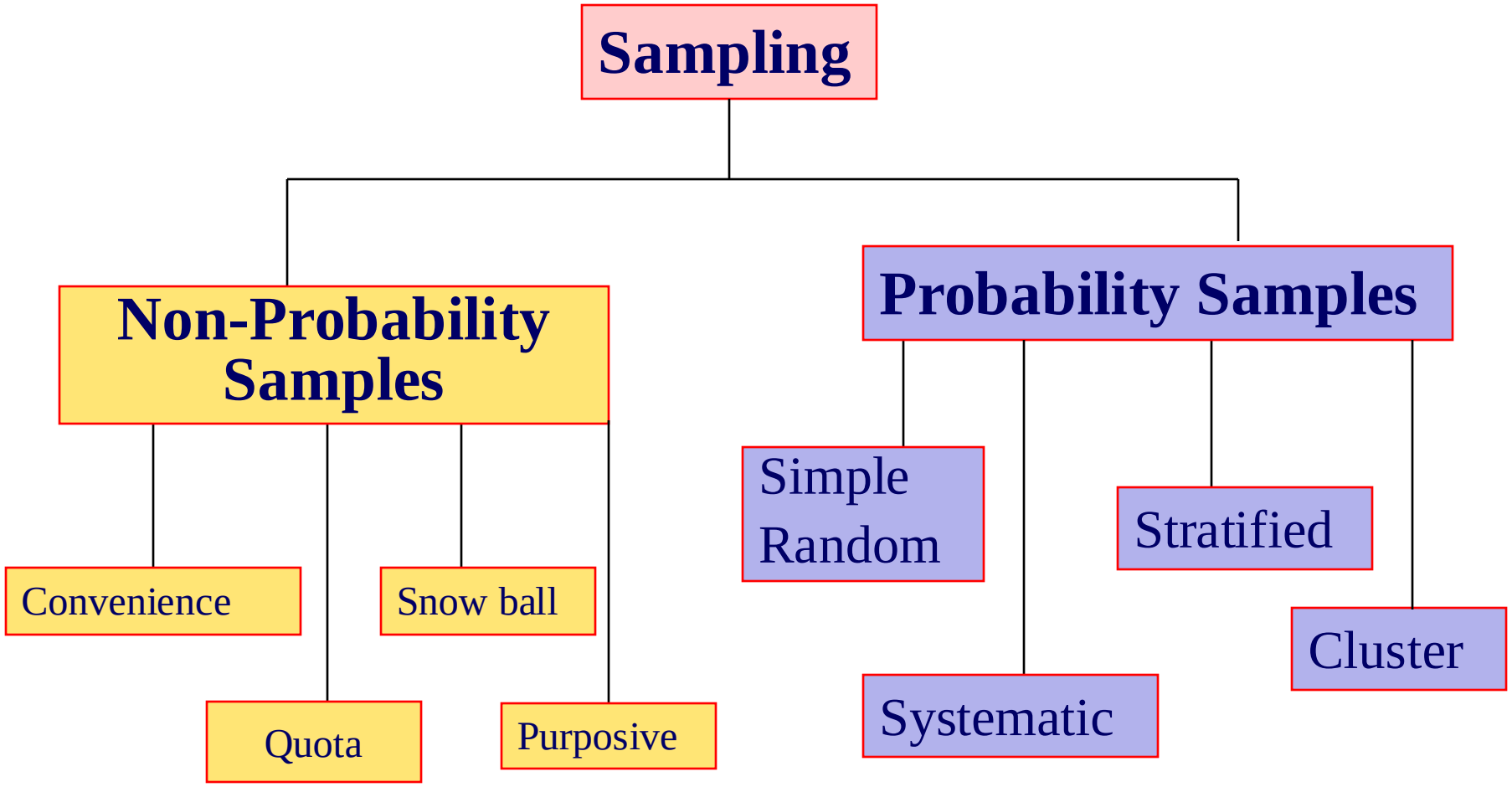
- Sampling is important because we can almost never look at the whole population.
- We use inferences on the sample to say something about the population.
- We always need estimates of variances on the sample calculations to say something about the population.

	Sample	Population
AB Testing	The users we show A and B versions of the website.	All users that visit our site: Past, Present, and Future.
World Cup Soccer Game Predictions	Only 32 teams post qualification in each season.	All national teams in the world for four years.
Average Height of a Data Science Student	UW Methods for DS class, ~25 observations.	All DS Students.

Sample Vs. Population

- If we sampled 4 beers and the ABV was [4%, 5%, 5%, 6%], then the sample mean would be 5%.
  - While there is variance in the SAMPLES, there is NO variance in the mean of that sample!!!
- But if we want to say something about a **population** of beers, we provide the mean with a variance statistic.
  - This allows us to say something to the effect of, “There is a 90% chance that the mean of all 4-sample beers lies between 4.5% and 5.5%”.
  - In order to say something about the population, we have to know how the sample was generated.

Types of Sampling Methods



Types of Probability Sampling

- Simple random
- Systematic sampling
- Stratified random
- Cluster sampling
- Multi-stage sampling

Simple Random Samples

- Every individual or item from the frame has an equal chance of being selected
- Selection may be with replacement or, without replacement
- Samples obtained from table of random numbers or computer random number generators
- Random samples are unbiased and, on average, representative of the population

Systematic sample

- This method is referred to as a systematic sample with a random start.
- This is done by picking every 5th or 10th unit at regular intervals.
- For example to carry out a filarial survey in a town, we take 10% sample. If the total population of the town is about 5000. The sample comes to 500.

Stratified Random sample

- This involves dividing the population into distinct subgroups according to some important characteristics, such as age, or socioeconomic status, religion and selecting a random number from each subgroup. (e.g. African voodoo healers)
- Especially important when one group is so small (say, 3% of the population) that a random sample might miss them entirely.
- Population divided into two or more groups according to some common characteristic
- Simple random sample selected from each group
- The two or more samples are combined into one

Cluster sample

- A sampling method in which each unit selected is a group of persons (all persons in a city block, a family, etc.) rather than an individual.
- Used when (a) sampling frame not available or too expensive, and (b) cost of reaching an individual element is too high
  - E.g., there is no list of automobile mechanics in the Myanmar. Even if you could construct it, it would cost too much money to reach randomly selected mechanics across the entire Myanmar : would have to have unbelievable travel budget
- In cluster sampling, first define large clusters of people. Fairly similar to other clusters. For example, cities make good clusters.
- Once you've chosen the cities, might be able to get a reasonably accurate list of all the mechanics in each of those cities. Is also much less expensive to fly to just 10 cities instead of 200 cities.
- Cluster sampling is less expensive than other methods, but less accurate.

Non- Probability Sampling /(Non-Random)

- This is where the probability of inclusion in the sample is unknown.
  - Convenience sampling
  - Purposive sampling
  - Quota sampling
  - Snow ball sampling

Convenience Sample

- Man-in-the-street surveys and a survey of blood pressure among volunteers who drop in at an examination booth in public places are in the category.

- It is improper to generalize from the results of a survey based upon such a sample for there is no known way of knowing what sorts of biases may have been operating.

Purposive/Judgment

- Selecting sample on the basis of knowledge of the research problem to allow selection of appropriate persons for inclusion in the sample
- Expert judgment picks useful cases for study
- Good for exploratory, qualitative work, and for pre-testing a questionnaire.

Snowball

> Recruiting people based on recommendation of people you have just interviewed

> Useful for studying invisible/illegal populations, such as drug addicts

Types of Sampling - Summary

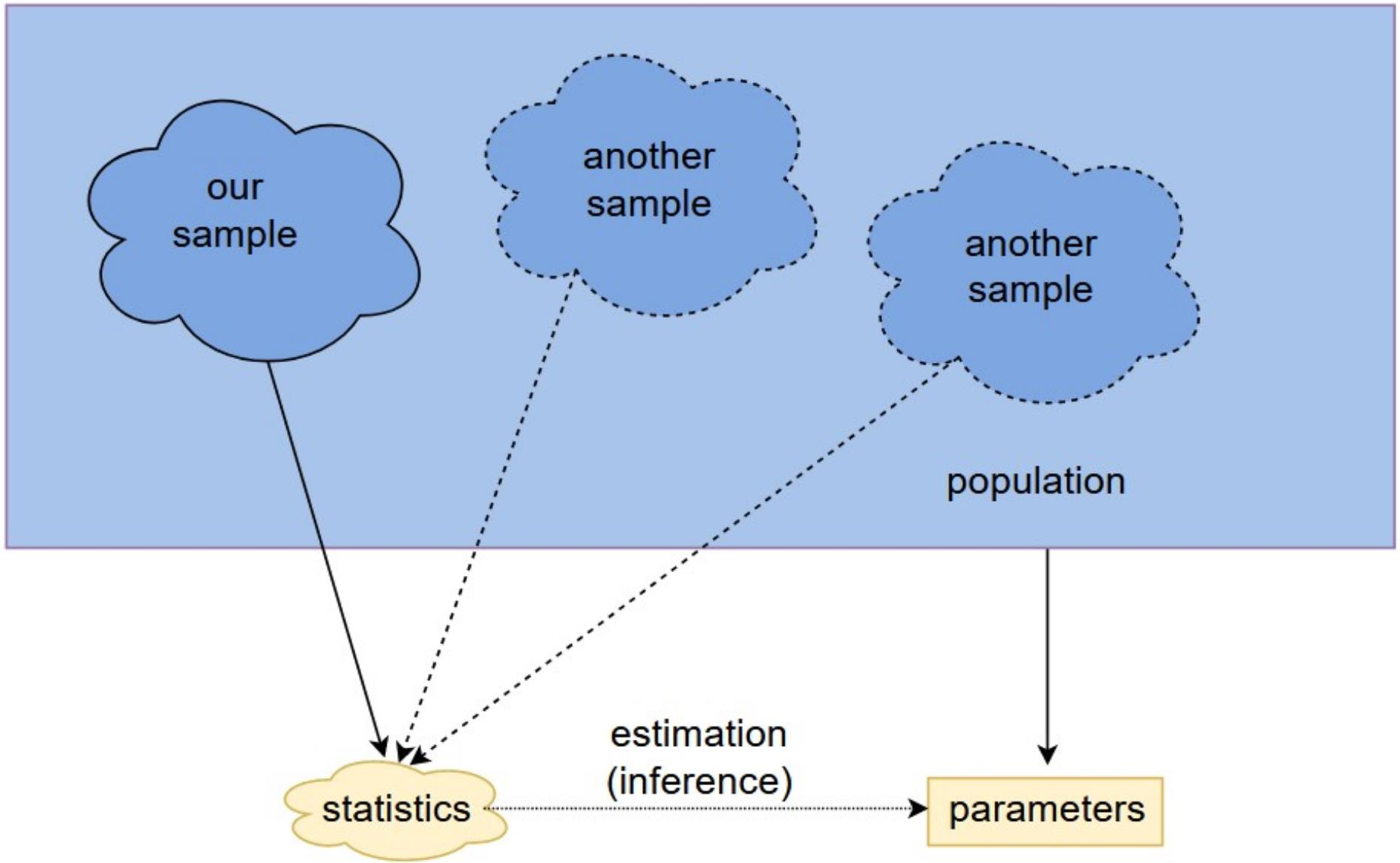
- Convenience or Accidental Sampling (This is bad).
  - Grabbing whatever is easier.
  - Grabbing whatever is available/possible.
- Cluster Sampling
  - Divide the data into clusters and sample select few clusters.
- Simple Random Sample (Most common)
  - Every point is subjected to a probability of being sampled.
- Stratified Sampling
  - Sampling subpopulations in a representative fashion.
  - This is sometimes important for class-imbalance problems.
- Systematic Sampling
  - Sampling every k-th element of a population. (Common in SQL servers, “table sampling”)
  - Bad for time series data.

Sample Design

- A set of rules or procedures that specify how a sample is to be selected
- This can either be probability or non-probability
- Sample size: The number of elements in the obtained sample
- How do we define populations? Answer: We use distributions

sources of uncertainty

- if we had access to the **whole population** we could measure anything we wanted, and our uncertainty about it is called epistemic uncertainty
- but that's usually not possible, or at least very expensive (such as the census), so in practice, we only have a **sample**, and we hope that the sample is **representative** of the larger population (otherwise we have a **biased sample**)
- so what we wanted to measure on the population (the **parameter**), we measure it on the sample (the **statistic**): the statistic is an **estimate** of the parameter
- because samples vary, we now have additional uncertainty, called **statistical uncertainty**, to account for: this is what **statistical inference** is all about


our sample
another sample
another sample
population
estimation (inference)
statistics
parameters

sampling methods

Suppose the data has N rows, and can be grouped into K subsets

- **simple random sample:** give me n rows at random ($n < N$)
- **Bernoulli sample:** for each row of the data, toss a coin with probability p of heads to determine if it should be in the sample or not
- **stratified sample:** for each of the K subsets, give me m rows at random
- **cluster sample:** first choose $k < K$ subsets at random, and choose m samples for each of those subsets
- **systematic sample:** give me every ith row of the data

discussion

You have been charged with finding the share of an American household's budget that goes towards child care

- let's say you conduct a **census**
  - what is your population and your **population parameter**?
  - what are some examples of **epistemic uncertainty**?
- let's say you conduct a **survey** instead
  - what is your statistic?
  - what kind of **sampling** would you perform?
  - what are some **sampling bias** problems you could run into?
  - what are some examples of **statistical uncertainty**?

discussion

Say you have a labeled data set with a binary target that you want to train a supervised learning algorithm on, and the data happens to be un-balanced, meaning the "positive" class for the target is very small compared to the "negative" class

- what sampling strategy should we use? (assuming sampling is needed)

Suppose you intentially corrupt the data by **mislabeling** 5% of the labels

- let's say you train a model and get training accuracy of 99%, what can we say about the accuracy of the model?
- what is the minimum we should expect for **epistemic uncertainty**? and how can we get rid of this uncertainty?

the law of large numbers

- a fundamental property of statistics (not just statistics the science, but more specifically statistics as measurements derived from **samples**, such as mean, median, IQR, variance, min, max, and so on)
- recall that **statistics** are *estimates* for **parameters**:
  - the **sample mean** is an estimate for the true **population mean**
  - the **sample variance** is an estimate for the true **population variance**
  - and so on, and so forth
- the **law of large numbers** states that the *larger* the **sample size**, the *better* the **estimate**: as we increase the sample size the sample becomes more representative of the population (and eventually the sample *is* the population)

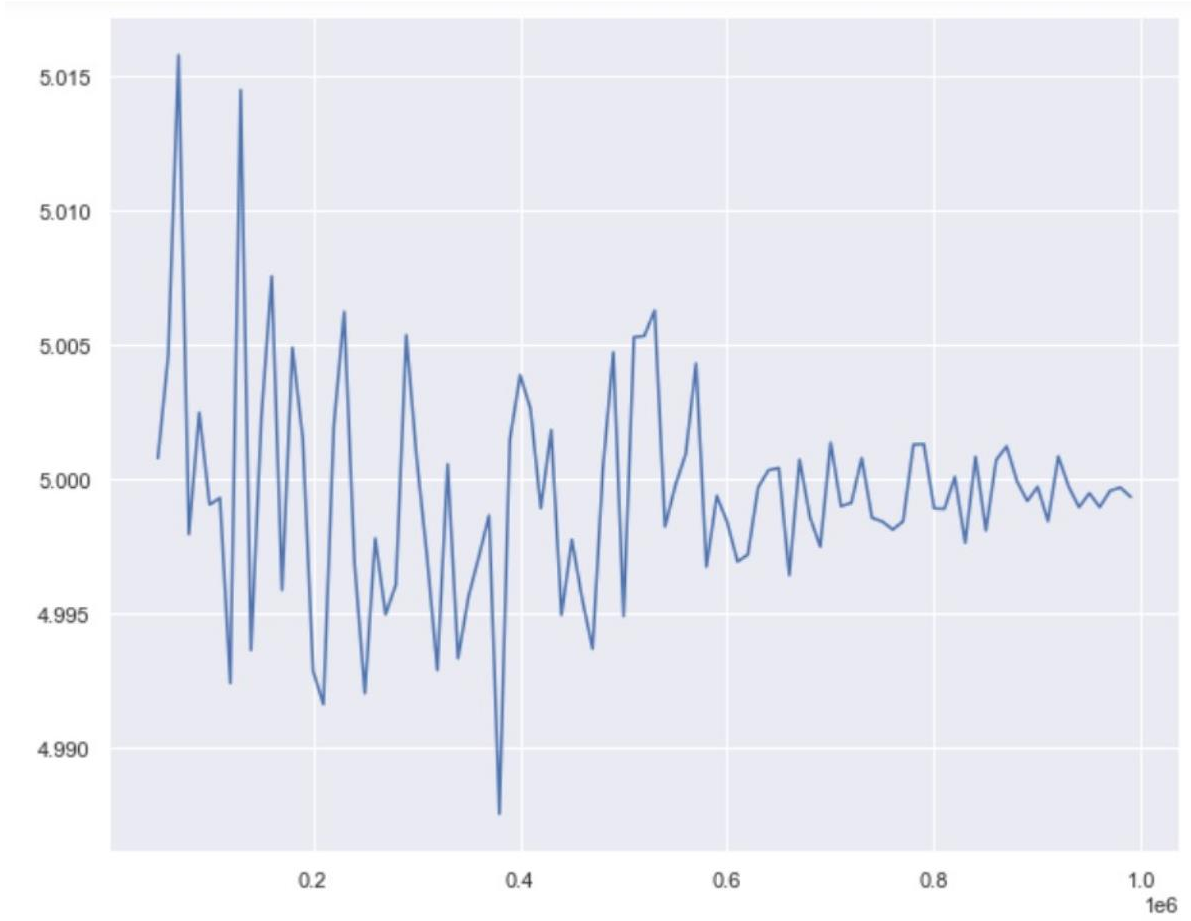
discussion

here is an example of the law of large numbers visualized

```
n, p, size = 100, 0.5, 10**6
pop = pd.DataFrame({'x': nr.binomial(n, p, size)})
n_range = np.arange(1, 10**6, 10000)
out = [pop.sample(n = x)['x'].std(axis = 0) for x in n_range]

sns.lineplot(x = n_range[5:], y = out[5:])
```

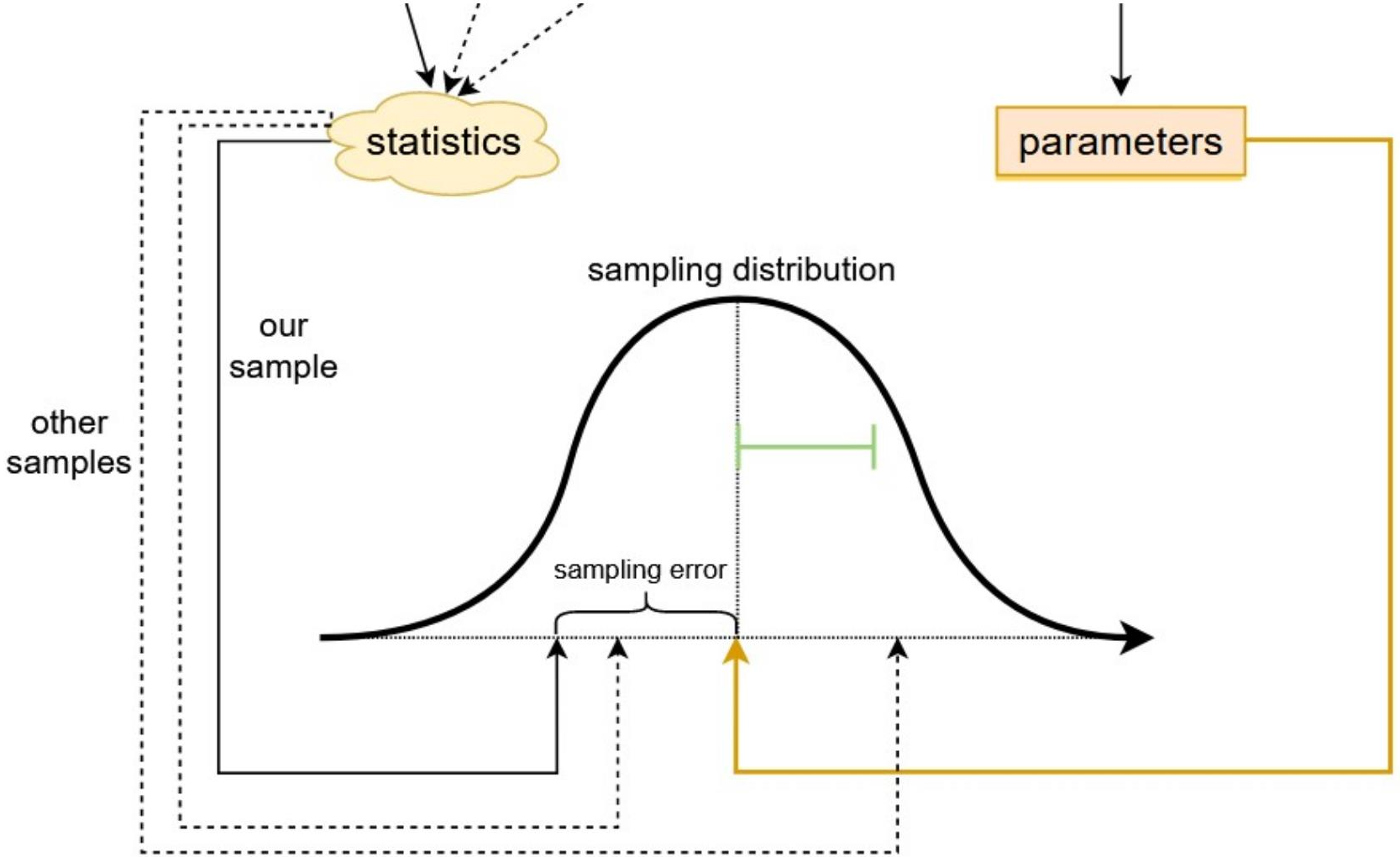
- what is the population?
- what is the population parameter we want to estimate?
- what do the axes on the plot represent?



the Central Limit Theorem (CLT)

- the law of large number is almost too obvious, but the CLT is more tricky!
- law of large number applied to any statistic, but the CLT applies only to the **mean**
- let X be a random variable representing the population
- X is allowed to have **any distribution** (not limited to normal), and let μ be your **true population mean** and σ the **true population standard deviation**
- let $\bar{X}$ be the **sample mean** for a **hypothetical** sample of size n from the population, recall that different samples give different $\bar{X}$ values, so $\bar{X}$ is also a random variable
- the distribution of $\bar{X}$ is called the **sampling distribution**, and according to CLT

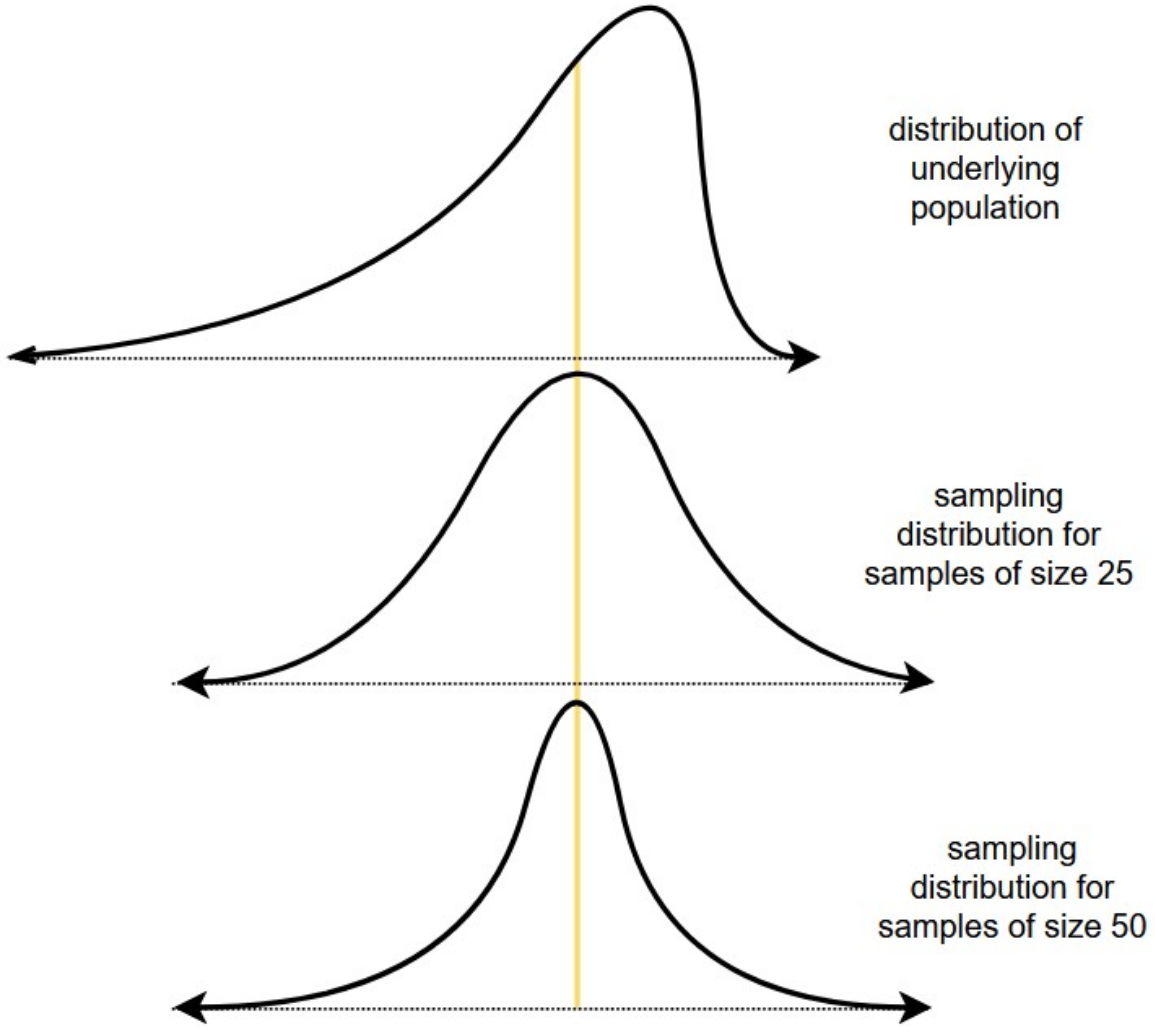
$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$



normal distribution
centered around true parameter
with variance inversely proportional to sample size

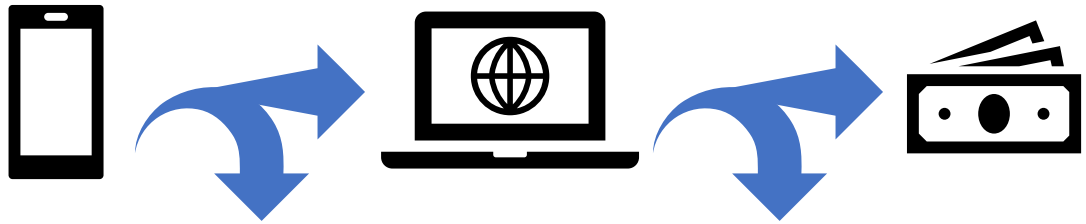
importance of CLT

- CLT is a sort of guarantee
- the distribution of the sample mean doesn't depend on the population the sample was drawn from
- it only depends on the population's mean and variance, and on the sample size
- the CLT is the basis for hypothesis testing (next lesson)

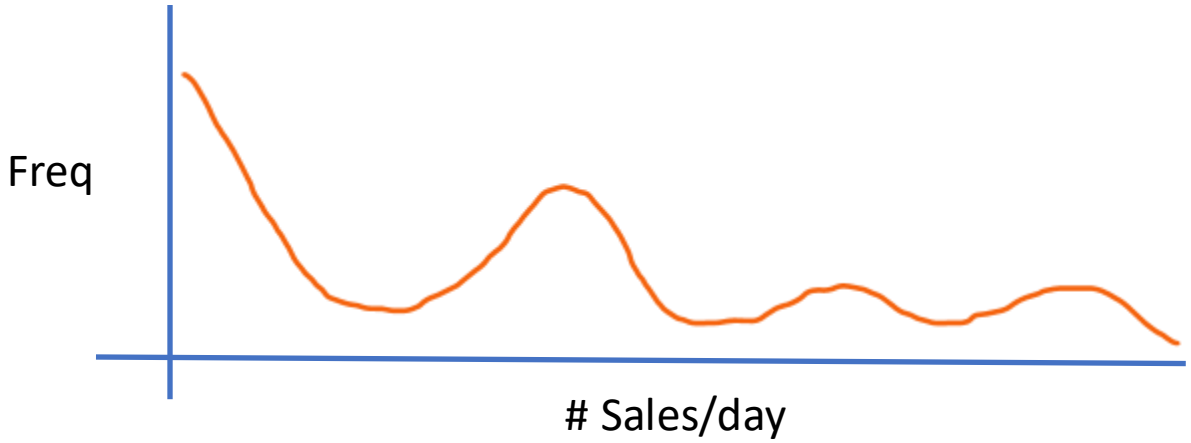


Real World Example: Website Page Views in Checkout

> Let's say there is a website that sells items. In order to get users, they place ads on mobile games.

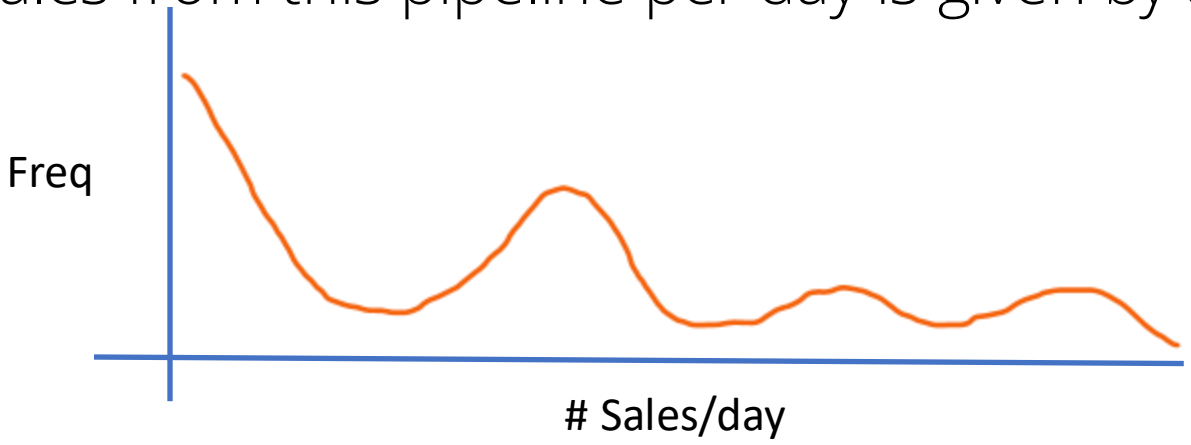


> The number of sales from this pipeline per day is given by a distribution that looks like:

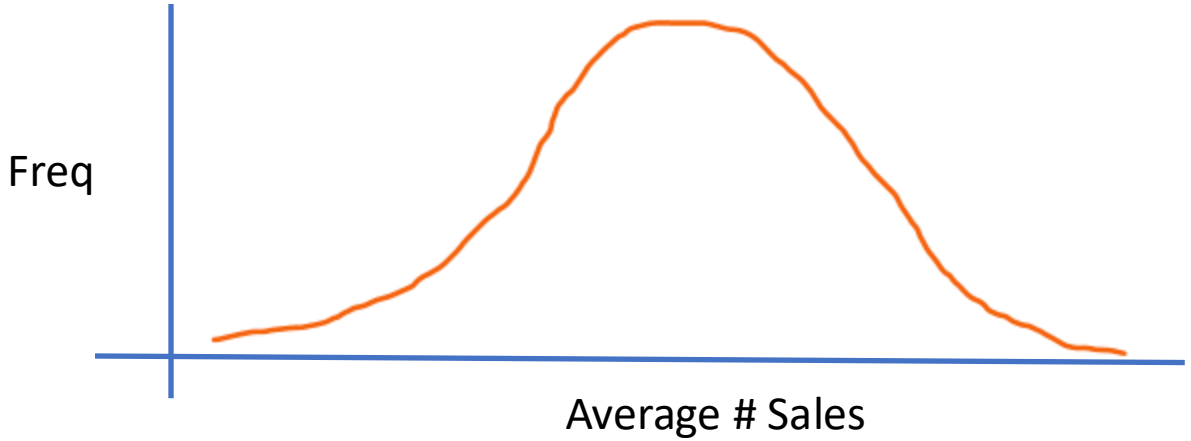


Real World Example: Website Page Views in Checkout

- The number of sales from this pipeline per day is given by a distribution that looks like:

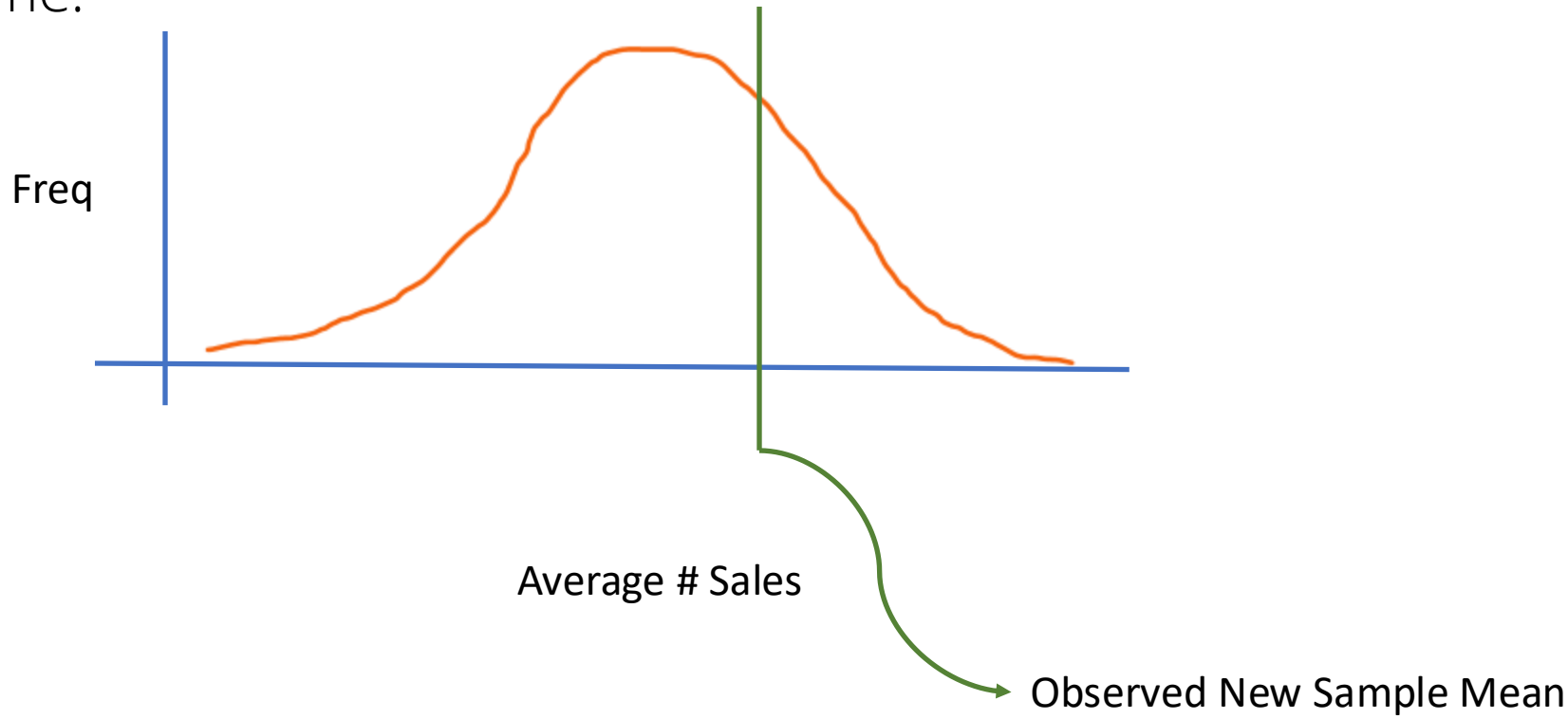


- Look at the MEAN sales per time frame.
  - Yay, this is Normally distributed!! We know this from the Central Limit Theorem!



Real World Example: Website Page Views in Checkout

> NOW, we are going forward with a trail on a NEW advertising platform for the same time frame.



> Discussion: Should we switch all ads or not?

Confidence Intervals

- Confidence intervals are a way to express uncertainty in population parameters, as estimated by the sample.
  - E.g. If we create a 95% confidence interval for the population mean, say

$$\bar{\mu} = \bar{X} = 10 \pm 5$$

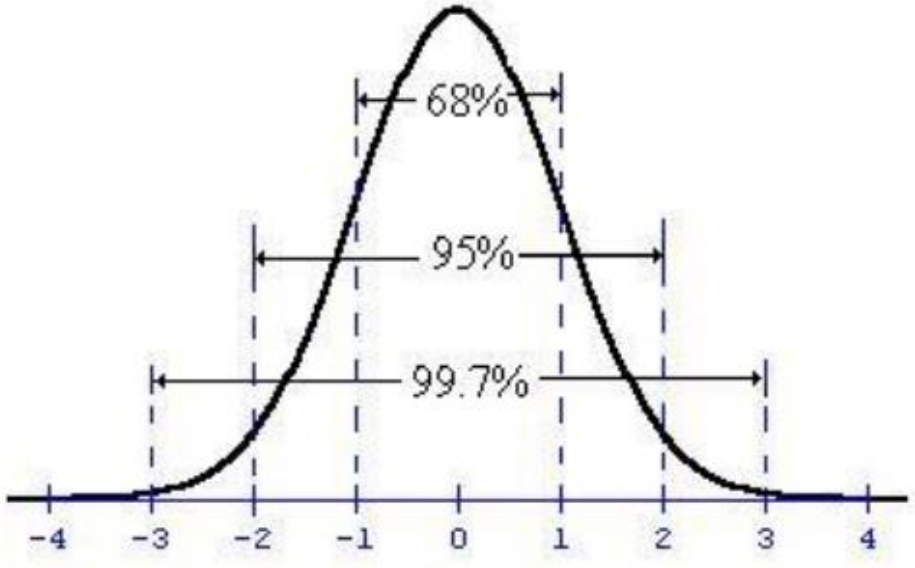
- Then we say that the true population mean, μ , has a 95% chance of being between 5 and 15.

It is not correct to say:

- "~~95% of the sample values are in this range.~~"
- "~~There is a 95% chance that the mean of another sample will be in this range~~."

Confidence Intervals

> To create confidence intervals for population means, we use the central limit theorem and create confidence intervals based on the normal distribution.
  - Repeatedly sample from the population.
  - Calculate the mean for each sample.
  - Use the average of the sample means as the population estimate and create a C.I. based on the s.d. of the sample means.



Dealing with Smaller Datasets

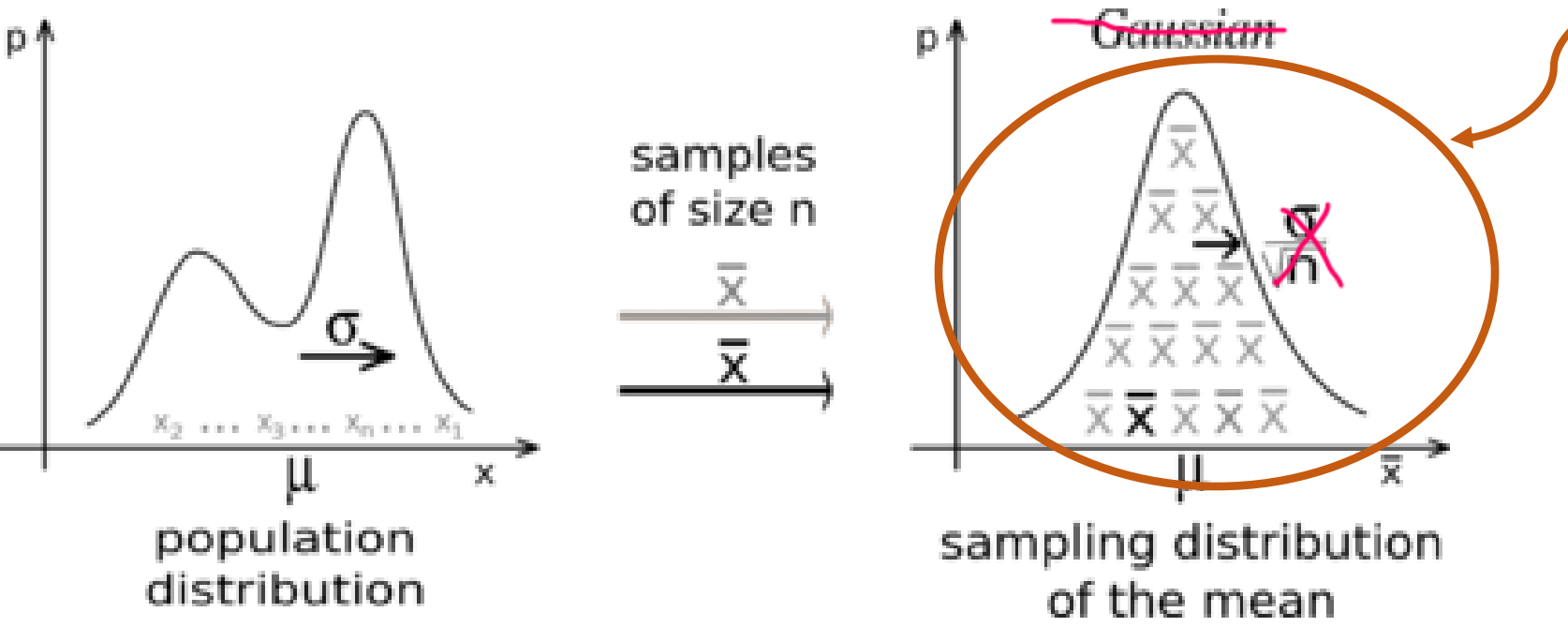
- When dealing with small datasets, it's hard to get a grasp of the true variability in sample statistics.
- Or when we have statistics we are calculating that are hard to know the true distribution (most everything that is not a mean).
- How can we use our small(er) sample to find this variability?
- The idea is to "create data and samples from nothing".
- Key assumption:
  - Our sample, however small, was created by randomly sampling.
  - This means that our sample is 'representative' of the population.
  - We create more data by sampling our sample WITH replacement

Bootstrapping!

- Bootstrapping is called as such because of a passage is Ulysses:
  - “There were others who had forced their way to the top from the lowest rung only by the aid of their bootstraps…”
  - We treat these resamples of the data as representatives of the whole population. In fact, under bootstrapping, we think of the population as the set of infinite resamples of the smaller sample.
- E.g. we poll the students in our class for estimates of the instructors age.
- The sample is small, so an error bound on the mean is large. We can reduce this by **bootstrapping.**

Bootstrapping

- The CLT allows us to estimate the distribution of simple statistics, like the mean.
- What if we want to estimate the distribution of other statistics:
  - Median
  - Mode
  - Variance
- Cannot rely on the CLT to know that the sample statistic will be known.

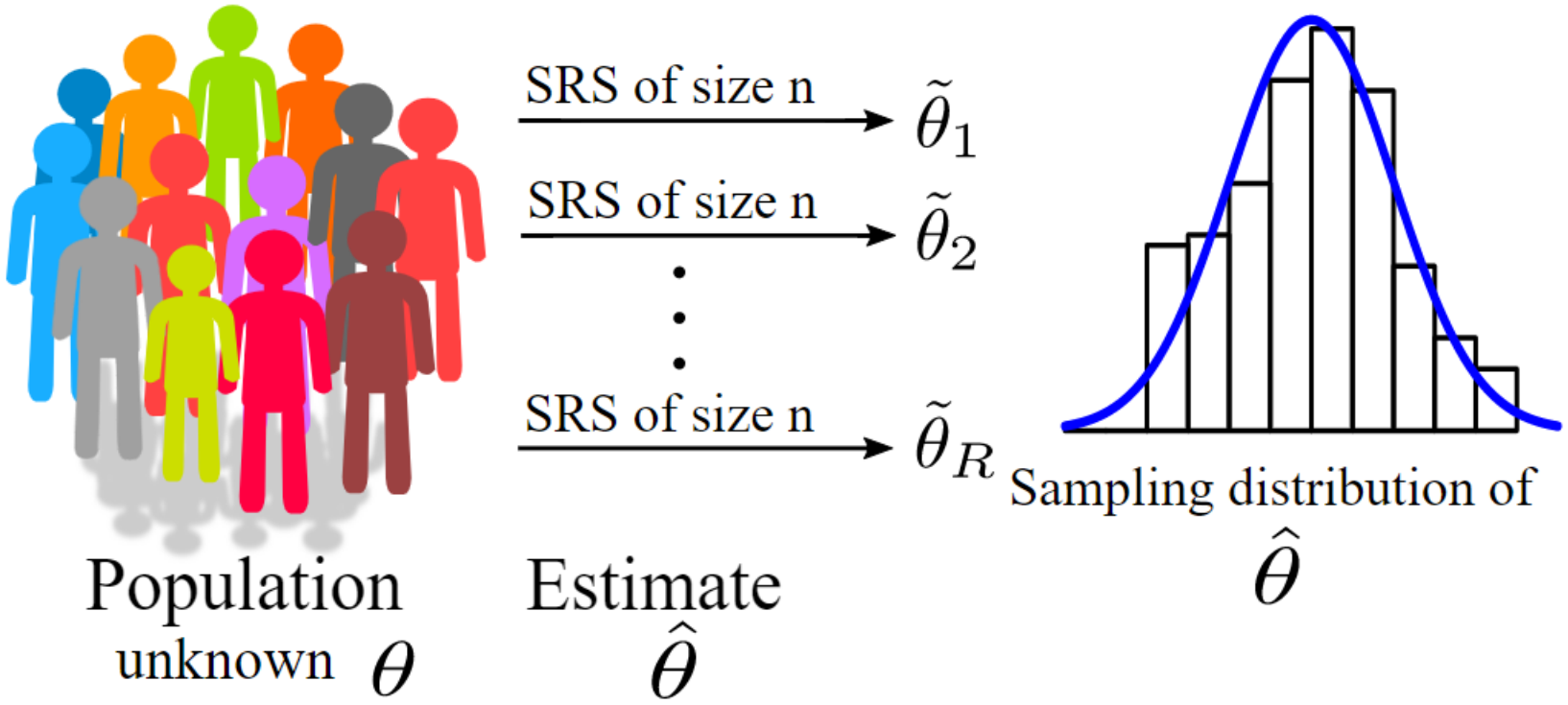


If only we could somehow just take repeated samples from the population and see how these sample statistics behave!

Well, why not…

Bootstrapping

- Bootstrapping is a statistical procedure for estimating the sampling distribution of an estimator by resampling with replacement from the original data.
- We ***assume*** that our sample is representative of the actual population…
  - … then we subsample (with replacement) from that.



Bootstrapping

- Bootstrapping can be used for both parameter estimation and hypothesis testing.
- One of the key advantages of bootstrapping is that it does not require an assumption of normality for the underlying population distribution.
- The basic steps of bootstrapping include:
  - Select a sample with replacement from the original data.
  - Calculating the test statistic of interest for each sample.
  - Repeat the process a large number of times to create a distribution of the test statistic.
- Bootstrapping can be used in combination with other statistical methods, such as regression analysis and hypothesis testing, to provide more robust results.

Estimating Probabilities With Simulation

- If we can generate or simulate large enough data sets, we should be able to use that data, as is, to estimate probabilities.
- We can make basic assumptions and write programs to simulate outcomes. From this set of outcomes we can estimate probabilities.
- Examples:
  - To estimate a new API usage, we can run a simulation with knowledge of prior/existing API activity to sample from.
  - To estimate probabilities of a product selling out, we can simulate customer behaviour from other product releases.
- Cautions:
  - You have to make assumptions to create simulations. These assumptions can easily be broken in new situations.

notebook time

we return to the lecture later

the end