Data Science 1
Statistical Inference for One Population
Sample Statistics

Ann Maharaj

Statistical Inference - Sample Statistics

1. Introduction
2. Sampling Distribution of the Sample Mean
3. Sampling Distribution of the Sample Proportion
4. Sampling Distribution of the Sample Variance
5. Summary of Sampling Distributions

Introduction

- Statistical inference is the process by which conclusions are drawn about a population from analysing a representative sample drawn from this population. This involves estimation and hypothesis testing.

- In this part of the topic, we focus on estimation of parameters for a single population.

- A point estimator is a single value that estimates an unknown population parameter such as the population mean, variance, proportion, and it is the corresponding sample statistic.

	Population Parameter	**Sample Statistic**
Mean	μ	$\overline{X}$
Variance	σ^2	s^2
Standard Deviation	σ	s
Proportion	π	p

Introduction

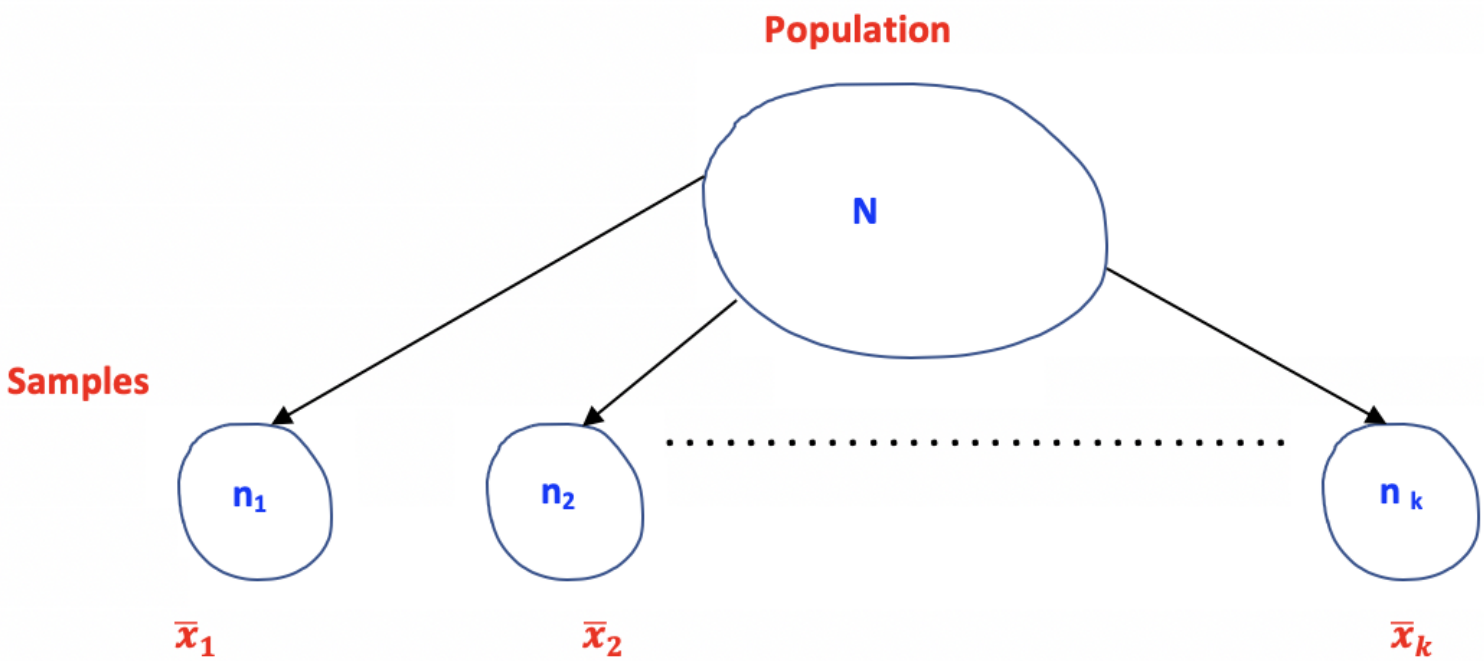
- Proper analysis and interpretation of a sample statistic requires knowledge of its distribution.
- In particular, in order to obtain a confidence interval estimator of a population parameter, we must know the sampling distribution of the corresponding sample statistic.
- The sampling distribution of a statistic is the probability distribution of the statistic over all possible samples.
- We will briefly discuss the sampling distributions of the sample mean, sample proportion and sample variance.

Sampling distribution of the sample mean

- The population mean μ is a parameter and it is a fixed value.
- The sample mean $\overline{X}$ is the point estimator of the population mean and it a sample statistic.
  - It varies according to which sample is actually taken.
  - For a particular sample size, the sampling distribution of the sample mean is the probability distribution of the sample mean over all possible samples.

Sampling distribution of the sample mean

- Take all possible simple random samples of size n from a population of size N.



- The frequency distribution of all the sample means $\overline{x}_1, \overline{x}_2, ...\overline{x}_k$ leads to the sampling distribution of the sample mean $\overline{X}$.

Sampling distribution of the sample mean

Given a random variable of interest X with mean μ and standard deviation σ, the sample mean $\overline{X}$ has

mean

$$E(\overline{X}) = E\left(\frac{\Sigma_{i=1}^{n} X_i}{n}\right) = \frac{1}{n} E(\Sigma_{i=1}^{n} X_i) = \frac{1}{n} n\mu = \mu$$

variance

$$Var(\overline{X}) = Var\left(\frac{\Sigma_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2} Var(\Sigma_{i=1}^{n} X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

and hence **standard deviation** also referred to as the **standard error of the mean**

$$SE(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

$E(\overline{X}) = \mu \Rightarrow \overline{X}$ is an unbiased estimator of μ

Sampling distribution of the sample mean

- If X is normally distributed, i.e., $X \sim N(\mu, \sigma^2)$, then the sample mean $\overline{X}$ is normally distributed.

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

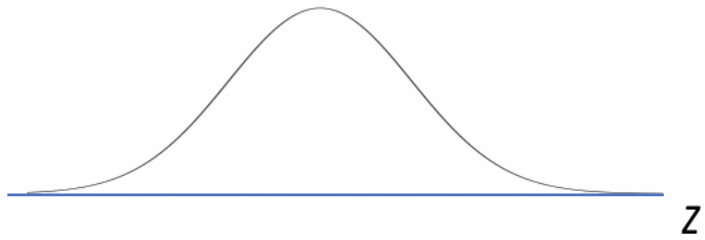
- If X is not normally distributed or if its distribution is unknown, then provided that n is large enough, $\overline{X}$ is approximately **normally distributed**. This is as a consequence of the **Central Limit Theorem**.

Sampling distribution of the sample mean

- Knowing the sampling distribution of the mean, allows us to make probability statements about the sample mean $\overline{X}$.
- Just as we standardise a random variable X that is normally distributed, we can standardise the sample mean $\overline{X}$.

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

- Z is normally distributed with mean 0 and variance 1, that is, $Z \sim N(0, 1)$.



Sampling from a finite population

- $\overline{X}$ is still approximately normally distributed with mean μ but with standard error

$$SE(\overline{X}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

where N is the population size.

$$\sqrt{\frac{N-n}{N-1}}$$

is called the finite population correction factor (fpc).

- If n is small relative to N, the fpc factor is close to 1 and SE($\overline{X}$) is approximately equal to $\frac{\sigma}{\sqrt{n}}$.
- As a rule of thumb, if the population size N is known and if $\frac{n}{N} > 0.05$, then the fpc should be included.

Population standard deviation is unknown

- If the population standard deviation σ is unknown, it is estimated from the sample, i.e., the sample standard deviation, s. Then the estimated standard deviation of the sample mean $\overline{X}$ which is also referred to as standard error of the sample mean is

$$SE(\overline{X}) = \frac{s}{\sqrt{n}}.$$

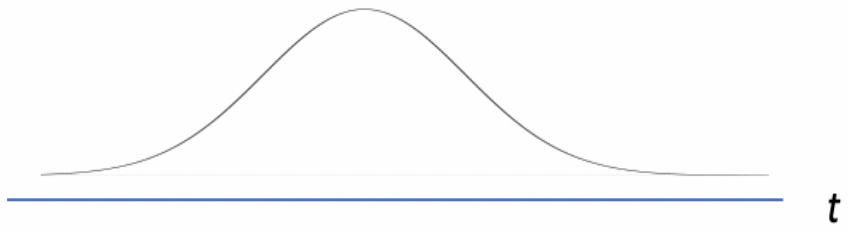
- If the population size N is known and, we include the *fpc*

$$SE(\overline{X}) = \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

Population standard deviation is unknown

If $X \sim N(\mu, \sigma^2)$, it can be shown that the standardised sample mean follows a ***t*-distribution** with n-1 degrees of freedom.

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$



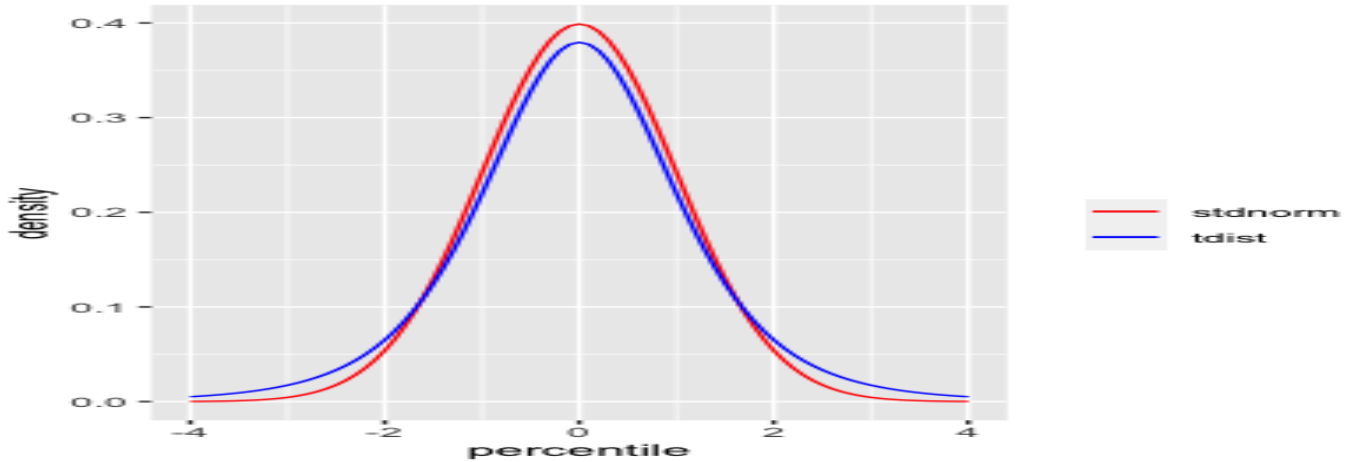
t- distribution

- The term degrees of freedom df refers to the number of values of a variable that are free to vary, given some restriction on the data.
  - One degree of freedom is lost for every prior estimate that is included in the new computation.
  - For example in calculating the sample variance, one degree of freedom is lost since the sample mean is a prior estimate that is included in the computation.
  - In general the degrees of freedom
    df = number of sample observations - number of estimated parameters.
- The standardised sample mean t is a numerical value of the t-distribution that defines the precise shape of the distribution.

t- distribution

The t-distribution looks very much like the standard normal distribution.

- It is symmetrical, bell-shaped and centred at 0.
- It is flatter and slightly more spread out than the normal distribution and the spread is greater for small degrees of freedom.
- As n increases, so do the degrees of freedom, and the t-distribution gets closer to the normal distribution, if n is *large enough*, the t distribution and the standard normal distribution are practically indistinguishable.
- For larger n, the requirement that X is normally distributed becomes less important.

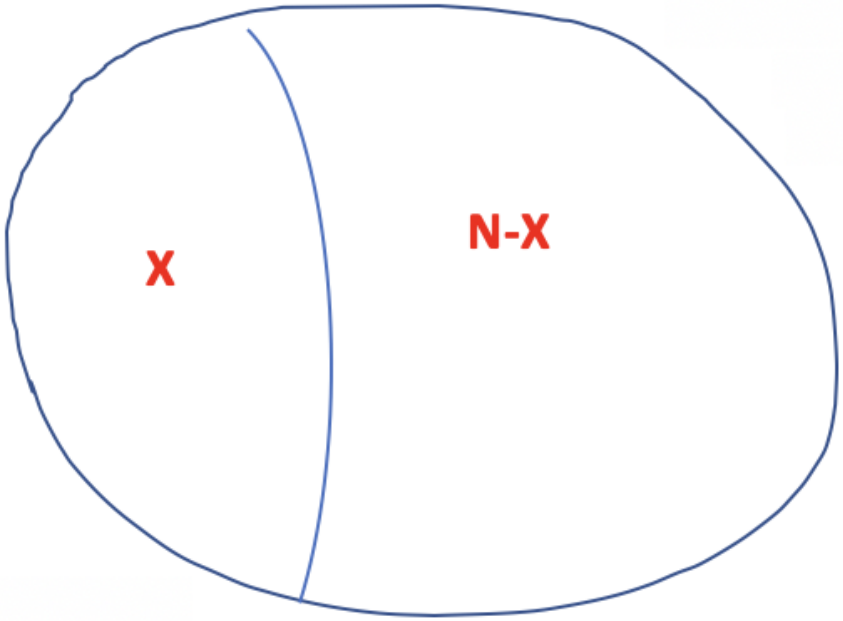


Sampling distribution of the sample proportion

- For categorical variables having only two possible outcomes, such as good or bad, male or female, and so on, we are usually interested in the proportion of observations in a sample that have a certain characteristic.
- The point estimator of a population proportion π is the statistic $p = \frac{X}{n}$, the sample proportion.
- X is the number of observations in the sample having the desired characteristic and n is the sample size.

Sampling distribution of the sample proportion

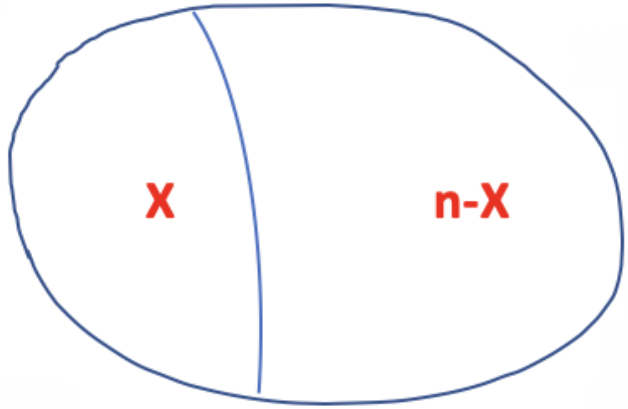
Population



Population Proportion

$\pi = X/N$

Sample

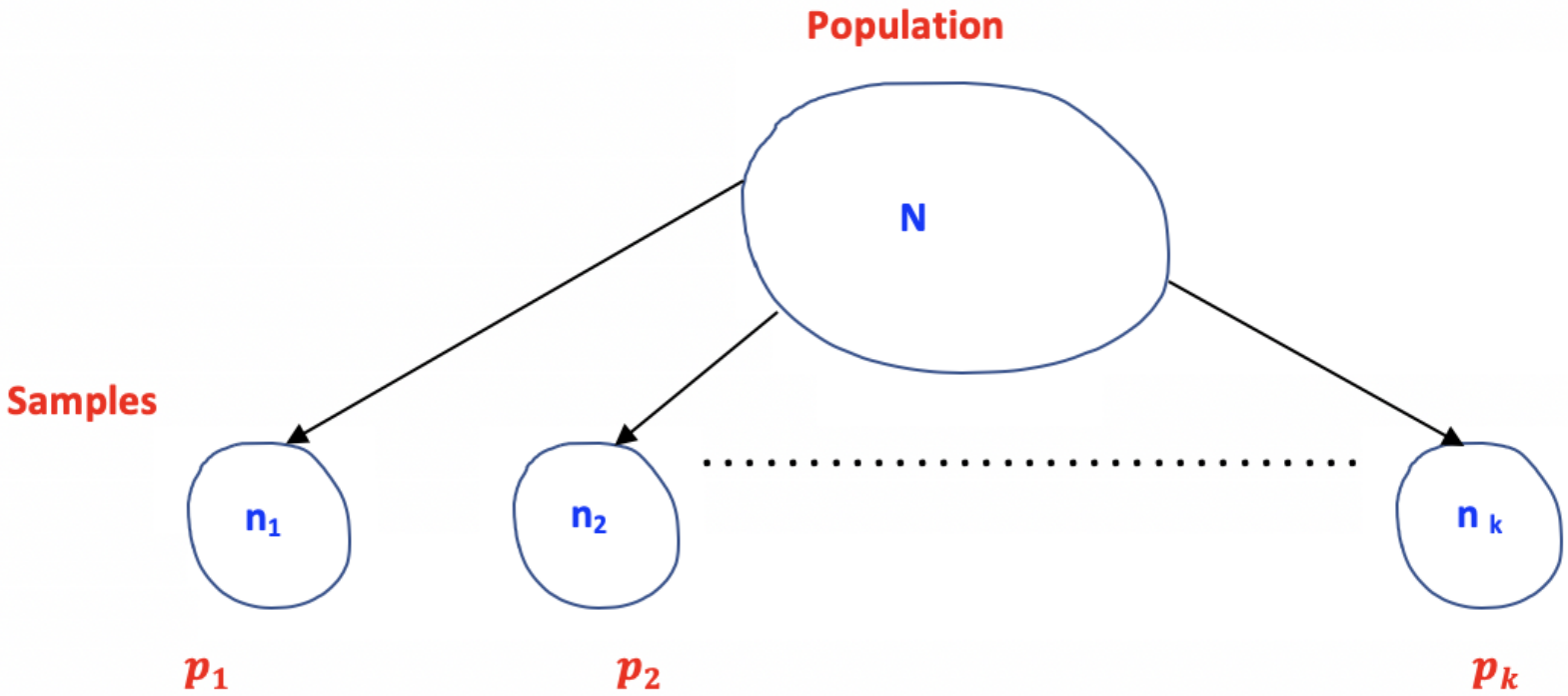


Sample Proportion

$p = X/n$

Sampling distribution of the sample proportion

- Take all possible simple random samples of size n from a population of size N.



- The frequency distribution of all the sample proportions $p_1, p_2, \ldots p_k$ leads to the sampling distribution of the sample proportion p.

Sampling distribution of the sample proportion

- The sampling distribution of the sample proportion p is the probability distribution of all possible values of p.
- If we are sampling with replacement from a finite population, and are interested in number of items X associated with a particular characteristic of the population, the sampling distribution of X follows the binomial distribution with mean $E(X) = n\pi$ and variance $Var(X) = n\pi(1-\pi)$.
- It follows that

$$E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{n\pi}{n} = \pi$$

and

$$Var(p) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}Var(X) = \frac{1}{n^2}n\pi(1-\pi) = \frac{\pi(1-\pi)}{n}.$$

Sampling distribution of the sample proportion

- If follows that the sampling distribution of the sample proportion $p = \frac{X}{n}$ has mean and variance

$$E(p) = \pi$$

$$Var(p) = \frac{\pi(1-\pi)}{n}.$$

- Hence the standard deviation (standard error) is

$$SE(p) = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

- $E(p) = \pi \Rightarrow p$ is an unbiased estimator of π

Sampling distribution of the sample proportion

- If n is sufficiently large, the sampling distribution of p approaches a **normal distribution** as a consequence of the **Central Limit Theorem.** Hence

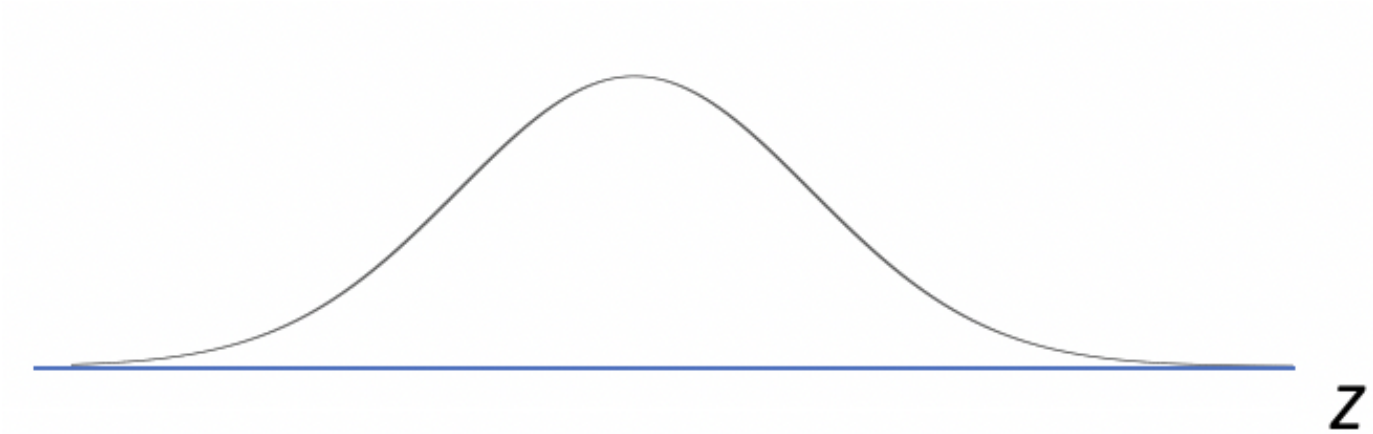
$$p \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right).$$

- As a rule of thumb, we use this normal approximation to the binomial distribution when $n\pi$ and $n(1-\pi)$ are each at least 5.

Sampling distribution of the sample proportion

Just as we standardise a random variable that is normally distributed, we can standardise the sample proportion.

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0, 1).$$



Sampling distribution of the sample variance

- The population variance is σ^2.
- The sample variance is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}.$$

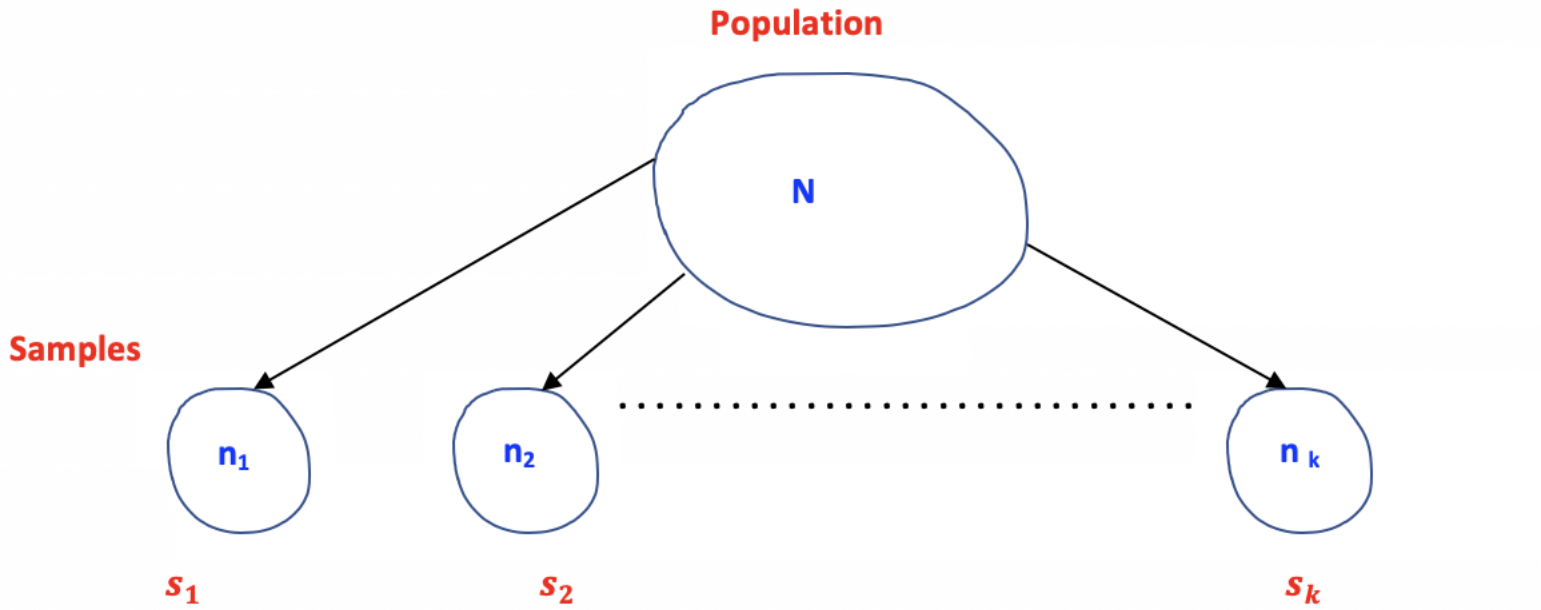
- It can be shown that $E(s^2) = \sigma^2 \Rightarrow s^2$ is an unbiased estimator of σ^2.
- If the random variable of interest X, is normally distributed, i.e., the sample is drawn for a normal population, it can be shown that

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

follows a **chi-square distribution** with $n-1$ degrees of freedom, denoted by χ^2_{n-1}.

Sampling distribution of sample variance

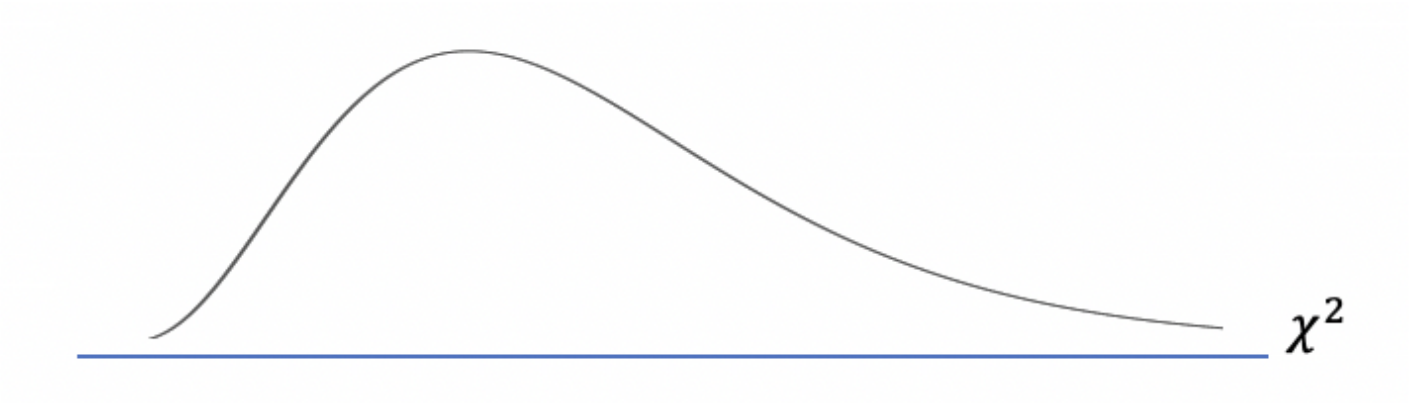
- Take all possible simple random samples of size n from a population of size N.



- Associated with each sample variance s_i^2 is the expression χ_i^2. The frequency distribution of $\chi_1^2, \chi_2^2, ..., \chi_k^2$ leads to the sampling distribution of χ^2.

Sampling distribution of the sample variance

- The chi-square distribution is skewed to the right.



- As the sample size increases, the degrees of freedom increase and the chi-square distribution becomes more symmetrical.

- The values of a chi-square distribution are always positive.

Summary of sampling distributions

- Sample Mean

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

- Sample Proportion

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)$$

- Sample Variance

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$