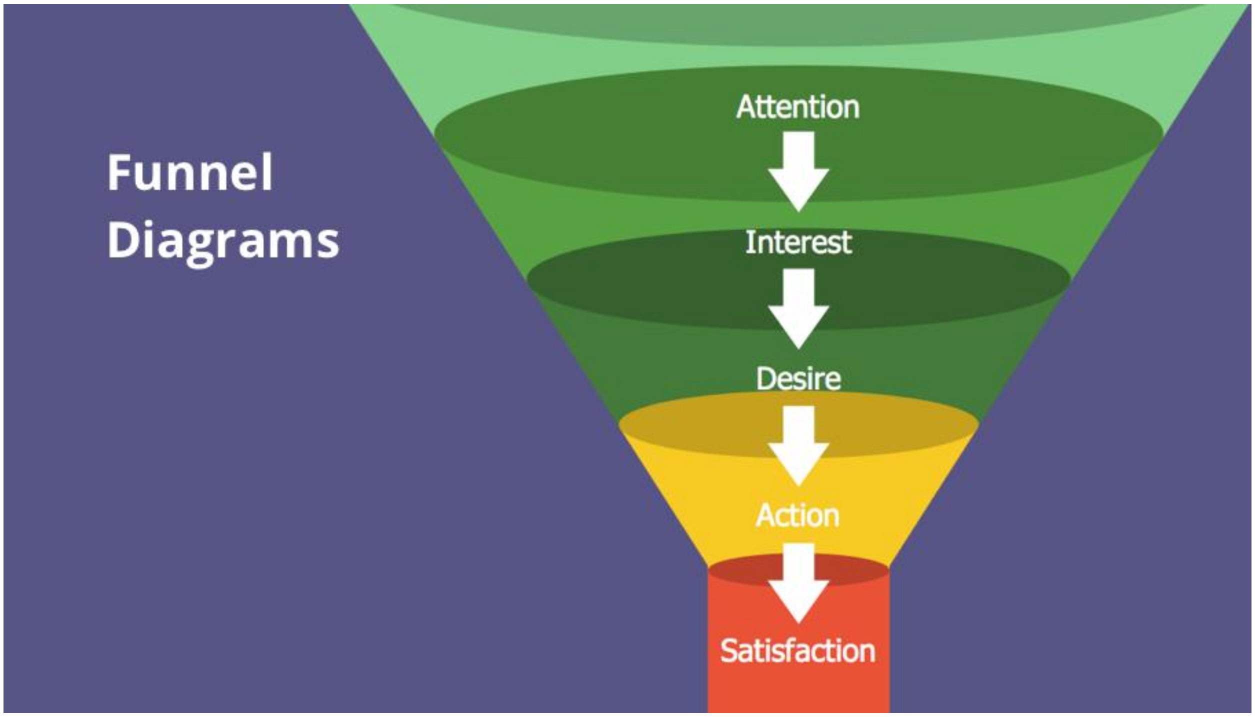
Introduction to DATA ANALYTICS

Point 1.5 – Sampling funnel, central limit theorem, confidence interval and sampling variation



Overview

The **syllabus of Unit 1 (Introduction to Data Analytics)** lists **sampling funnel**, **central limit theorem (CLT)**, **confidence interval** and **sampling variation** as the learning content for point 1.5 of the course's theory learning outcomes. These topics form the bridge between basic descriptive statistics and inferential statistics, enabling analysts to make reliable statements about a population based on a sample.

Why sampling is needed

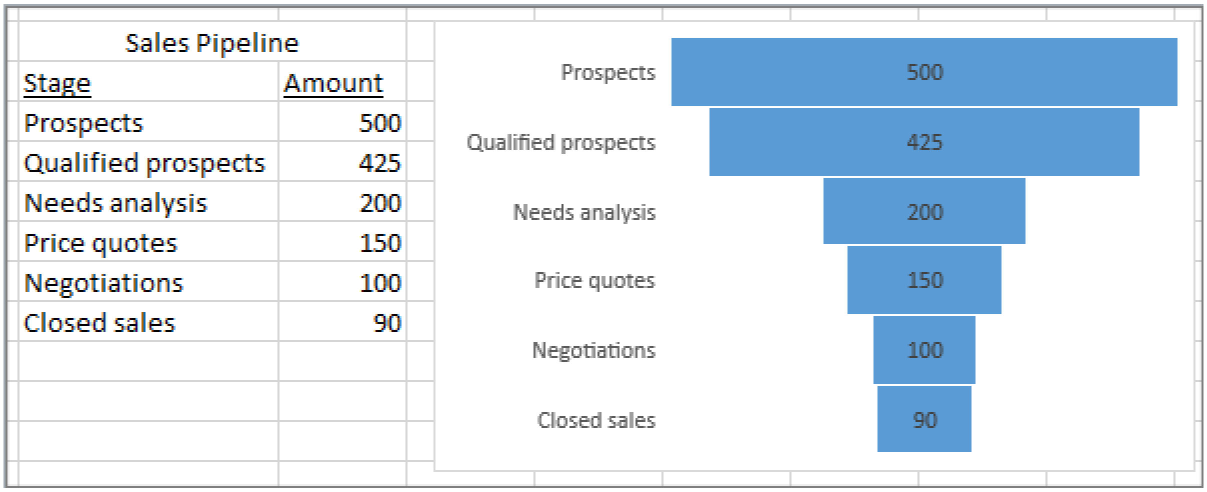
In real-world analytics, it is seldom possible or efficient to measure every member of a population. Instead, analysts draw a **sample**, calculate statistics (such as means and proportions), and generalize these to the population. Sampling introduces **sampling error**—the difference between a sample statistic and the unknown population parameter. Sampling error occurs because the sample does not include all population members[1]. Even well-designed random samples have variability; taking only a few observations produces widely varying results, whereas larger samples yield smaller sampling error[2].

1. Sampling funnel

A **sampling funnel** is a conceptual model that visualizes how analysts narrow down from a broad population to a manageable sample. At each stage of the funnel, decisions are made about **target population**, **sampling frame**, **sampling method** (probability or non-probability sampling) and **sample size**.

1. **Define the population and objectives.** Clearly specify the population (e.g., all customers, all manufactured units) and what parameter you wish to estimate (mean, proportion, distribution).
2. **Identify the sampling frame.** Construct a list or frame that approximates the population. For example, a list of registered users or shipments.
3. **Select a sampling method.** Common probability techniques include simple random sampling, stratified sampling (grouping into strata and sampling within each stratum), systematic sampling and cluster sampling. Non-probability methods (e.g., convenience sampling) may be used when frames are unavailable but can introduce bias.
4. **Apply the sampling funnel.** Starting from the sampling frame, successive filters (stratification, inclusion/exclusion criteria and random selection) reduce the set to the final sample. Each filter needs justification to preserve representativeness.
5. **Check sample size and variation.** Larger samples reduce sampling error and ensure that the sampling distribution of the statistic approximates normality (see CLT), which is necessary for confidence-interval estimation.

Why a funnel? The funnel metaphor emphasises that only a small portion of the population is measured, but the aim is to ensure that the sample still represents the broader population. A poorly designed funnel (e.g., non-random filters) can introduce **sampling bias**, producing inaccurate estimates and misleading decisions.



Sales Pipeline

Stage	Amount
Prospects	500
Qualified prospects	425
Needs analysis	200
Price quotes	150
Negotiations	100
Closed sales	90

2. Central limit theorem

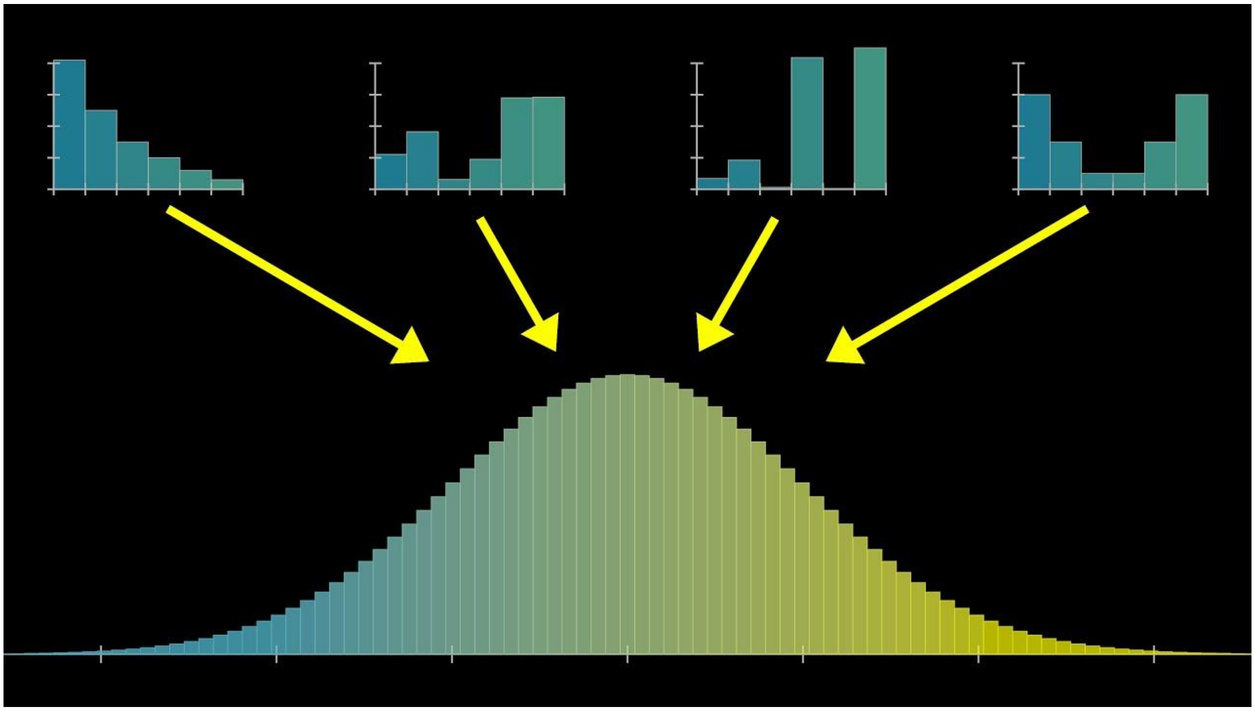
The **central limit theorem (CLT)** provides the theoretical justification for using sample statistics to draw inferences about populations. It states that, under appropriate conditions, the distribution of a normalized sample mean approaches a normal distribution as the sample size increases[3]. In particular, if $X_1, X_2, \dots, X_n$ are independent and identically distributed random variables with finite mean μ and variance σ^2 , then the normalized sample mean

$$Z_n = \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma}$$

converges in distribution to a standard normal $\mathcal{N}(0,1)$ as $n \to \infty$ [4]. This means that, regardless of the original distribution, the distribution of $\bar{X}_n$ becomes approximately normal with mean μ and variance σ^2/n for sufficiently large n [4].

Implications for data analytics:

- **Approximate normality**: even when the underlying data are skewed or non-normal, the sampling distribution of the sample mean becomes nearly normal for moderate sample sizes. A rule of thumb is that $n \geq 30$ yields a reasonable approximation.
- **Enables inference:** the normality of $\bar{X}_n$ allows analysts to compute probabilities, quantiles and confidence intervals using the normal (or t-) distribution.
- **Supports hypothesis testing:** many statistical tests (t-tests, ANOVA, regression) rely on CLT assumptions for large samples.



3. Confidence interval

A **confidence interval (CI)** is a range of values used to estimate an unknown population parameter[5]. Instead of reporting a single point estimate, analysts provide an interval, together with a confidence level (e.g., 95 %), indicating the long-run proportion of repeated samples that would produce intervals containing the true parameter[5].

For a population mean, when the sample size is large (so that the CLT holds) or the population is normally distributed, a two-sided $100(1-\alpha)$ % confidence interval for μ is

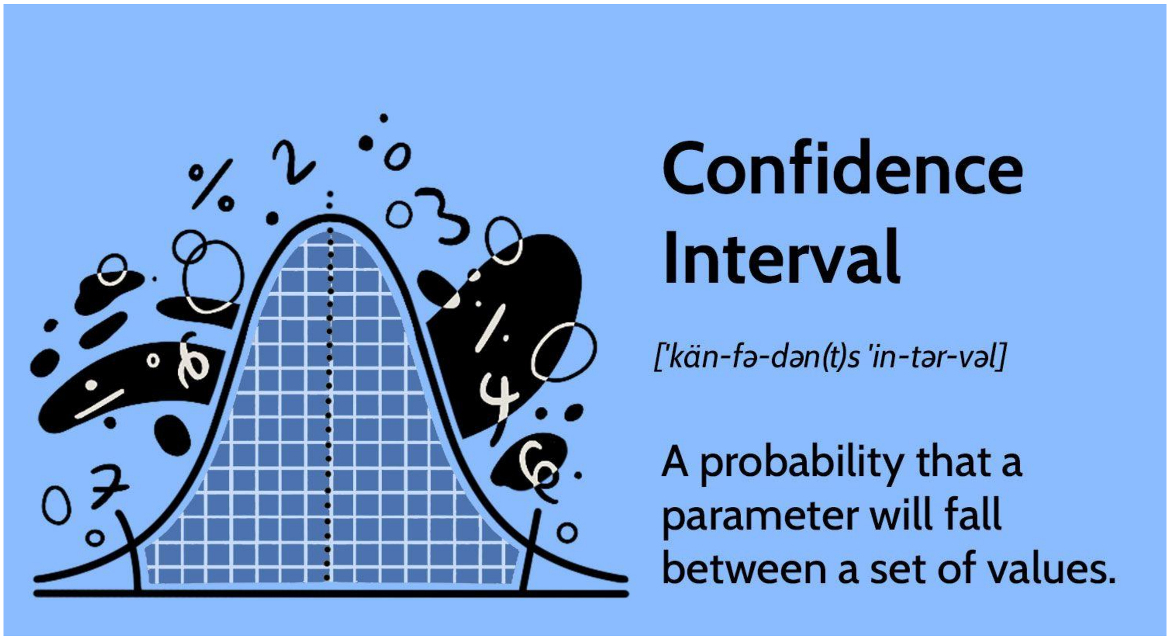
$$\bar{X} \pm z_{\alpha/2}\,\frac{S}{\sqrt{n}},$$

where $\bar{X}$ is the sample mean, S is the sample standard deviation, n is the sample size and $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal distribution[6]. In practice, when n is small, the Student's t distribution is used in place of z .

Interpreting a 95 % CI: if we repeatedly draw samples and compute 95 % CIs, about 95 % of those intervals would contain the true mean[5]. The confidence level is about the reliability of the procedure, not the probability that a specific interval contains the parameter.

Factors affecting interval width

- **Sample size (** n **)**: larger samples reduce the standard error $S/\sqrt{n}$, producing narrower intervals.
- **Population variability (** σ **)**: higher variability leads to wider intervals.
- **Confidence level**: higher confidence (e.g., 99 % vs. 95 %) requires a larger $z_{\alpha/2}$, hence a wider interval.



Confidence Interval

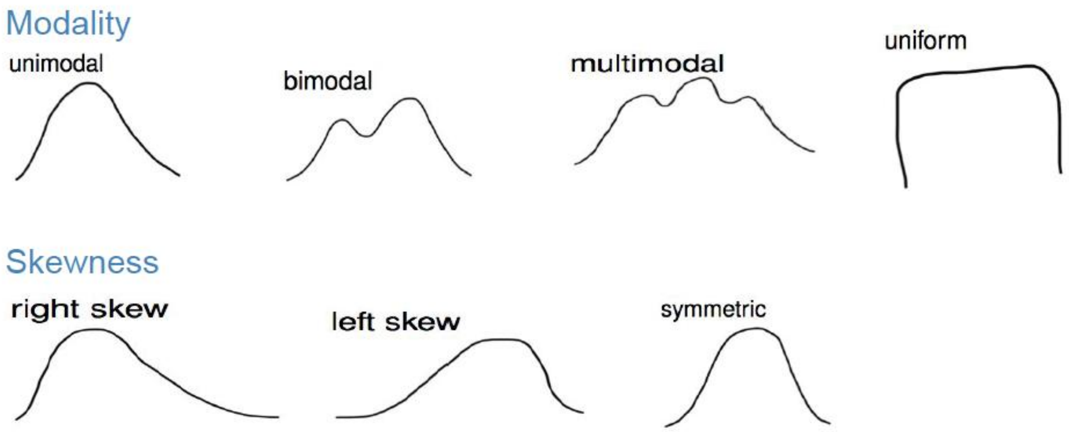
['kän-fə-dən(t)s 'in-tər-vəl]

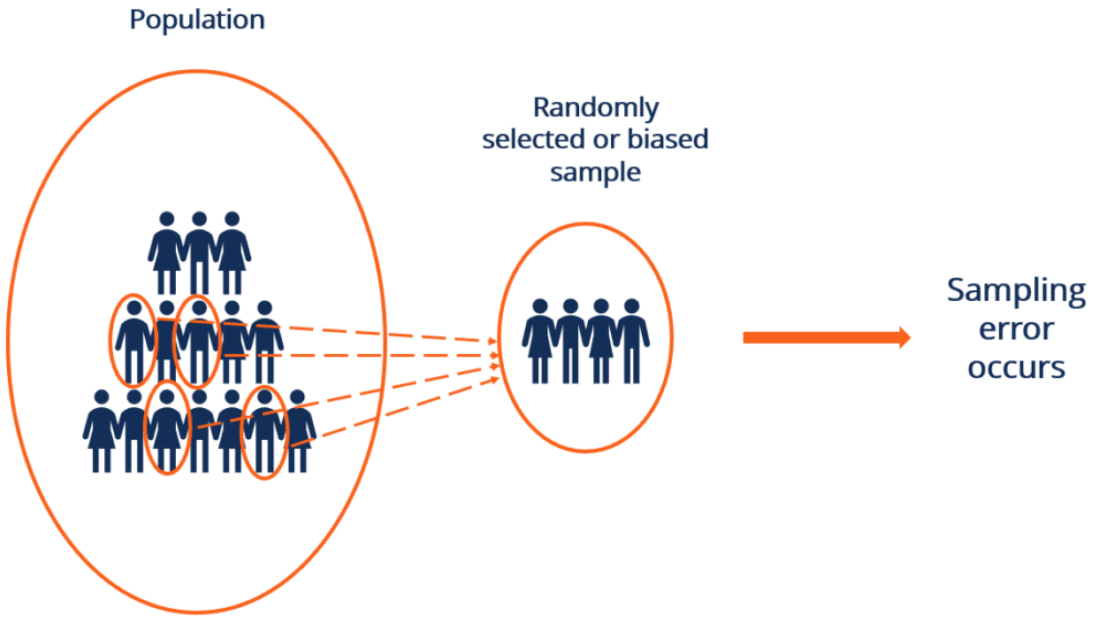
A probability that a parameter will fall between a set of values.

4. Sampling variation and sampling error

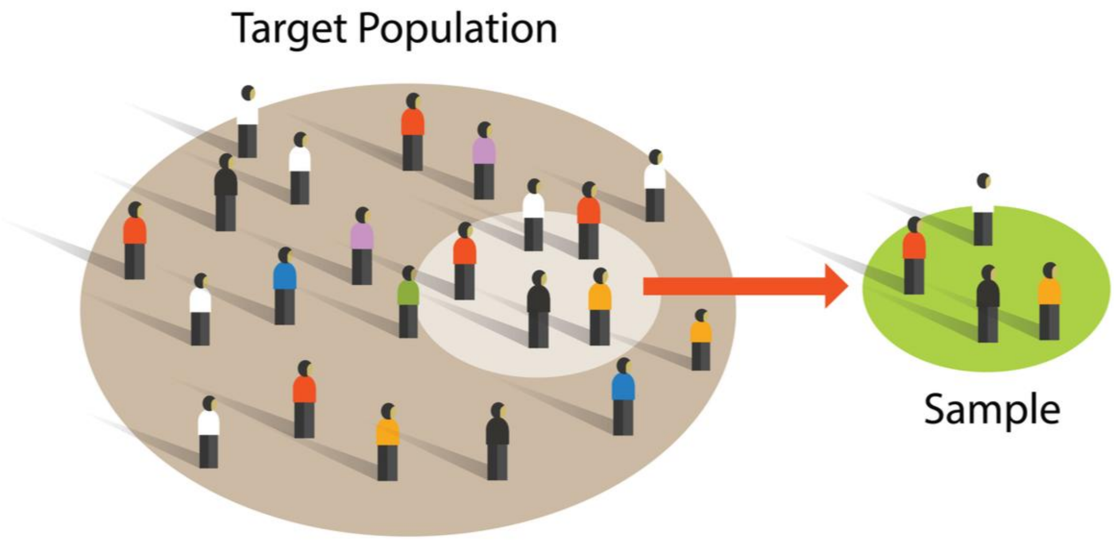
Because samples are only subsets of the population, repeated sampling will produce different estimates. The **sampling error** is the difference between a sample statistic (e.g., sample mean) and the true population parameter[1]. Even in a perfect random sample, variation remains due to chance[2]. Key points:

- **Random variation**: two random samples from the same population rarely yield identical statistics. The spread of the sampling distribution (standard error) quantifies this variation.
- **Reducing sampling error**: increasing the sample size decreases the standard error and sampling variation, improving precision. Stratified or cluster sampling can also reduce variability by controlling heterogeneity.
- **Relation to CLT and CI**: the CLT shows that sampling variation is predictable for large samples; confidence intervals incorporate sampling variation to provide ranges likely to contain the true parameter.





Summary



Point 1.5 of Unit 1 introduces the core **inferential statistics concepts** that underpin data analytics. The **sampling funnel** illustrates how analysts move from the full population to a representative sample. The **central limit theorem** ensures that sample means follow a normal distribution for large samples, enabling the construction of **confidence intervals**, which quantify uncertainty around estimates[3]. Understanding **sampling variation** and sampling error is crucial for designing studies, choosing appropriate sample sizes and interpreting results[1]. Together, these concepts allow data analysts to make reliable inferences from limited data and to communicate the uncertainty inherent in their estimates.

---

[1] [2] Sampling error - Wikipedia

https://en.wikipedia.org/wiki/Sampling_variability

[3] [4] Central limit theorem - Wikipedia

https://en.wikipedia.org/wiki/Central_limit_theorem

[5] [6] Confidence interval - Wikipedia

https://en.wikipedia.org/wiki/Confidence_interval