Study Design

Representative samples

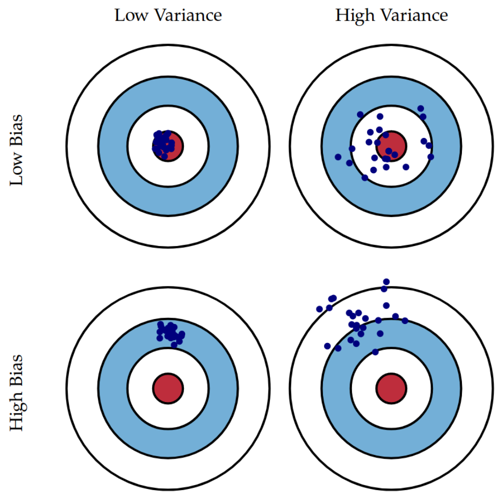
Data is representative if it is:

1. **Proportional Representation** – Subgroups in the population are represented in the dataset in proportions similar to their presence in the population.
2. **Diversity** – The dataset captures the full range of variability(outliers, different subgroups, rare cases) within the population.
3. **Random Sampling** – Data collect through random sampling
4. **Unbiased Selection** – Data is free from selection bias

Bias and Variance

Sources of error

1) **Sampling Bias** – a systemic flaw in how the sample was collected
2) **Sampling Variability** – Differences between samples due to random chance



Low variance – Data are close to each other
High variance – Data are more scattered
Low Bias – Data is close to the true value
High Bias – Data is far away from the true value

Type of Bias:

1) **Selection Bias** – Describes situations in which the method used to collect samples may be associated with the outcome in question.
2) **Non-response Bias** – Describes situation in which willingness to response may be associated with outcome

Sampling Distribution

Standard Deviation(σ)	the average deviation(distance) of individual observations from the mean value.	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ n = number of observation
Standard Error(SE)	Point estimates vary from sample to sample	$SE = \frac{\sigma}{\sqrt{n}}$ n = number of observations in the sample
Margin of Error		$Margin\ of\ Error = c \times \frac{\sigma}{\sqrt{n}}$ $= c \times SE$ c = Calibration value

Central Limit Theorem

The mean, difference in mean, proportion, and difference in proportion of a sample will follow normal distribution if:

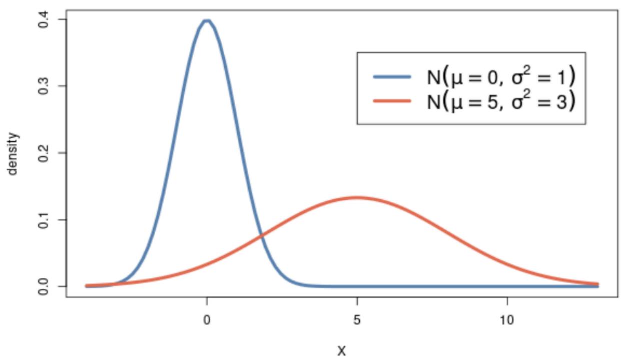
1. Observations in the sample are independent.
   (Samples need to be selected randomly so that their observations do not depend on the values of other observations)
2. The sample is large enough. (Usually when n >= 30)

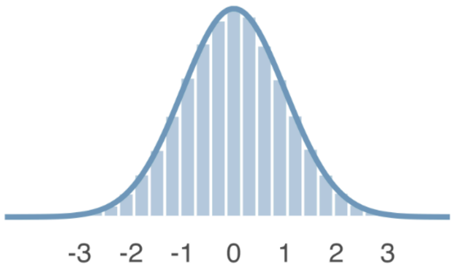
For any variable X with mean μ and population standard deviation σ, the sample mean will have a sampling distribution of: sample mean X ~ N(μ, $\sigma / \sqrt{n}$).

Normal distribution

Normal Distribution Always describes a symmetric, unimodal, bell-shaped curve.	**Standard Normal Distribution (z-distribution)** A type of normal distribution where mean is always 0 and sd is always 1.

X ~ N(μ, σ)
μ = mean
σ = standard deviation
	X ~ N(0, 1)
	
	To standardize a normal distribution, compute the distribution of the z-score of the data.





The peak of the distribution is always at 0

Confidence Intervals

a plausible range of values for the population parameter

Interpretation: something has a 95% confidence interval → the process that constructed this interval has the property that, on average, it contains the true value of the parameter 95 times out of 100.

Confidence Interval are affected by:

- Standard deviation → Greater, width of CI bigger
- Size of the sample → Greater, width of CI smaller
- Point estimate → Move the value of the CI
- Confidence level (99.7%, 95%, 68%)
- Method of computation(bootstrap, point estimate)

Method	**Explanation**	**Pros and Cons**
Point estimate method	% CI = Point estimate ± $c\frac{\sigma}{\sqrt{n}}$ c = critical value $\frac{\sigma}{\sqrt{n}}$ = margin of error	Have to be normal distribution
Bootstrapping method:	80% CI = from the value in 10th percentile to 90th percentile 1. Resampling with replacement from original data n times. 2. For each bootstrap calculate the mean(x*). 3. Sort the bootstrapped mean in ascending order 4. Find the 10th percentile and 90th percentile.	Useful when the distribution is unknown

Bootstrap Procedure

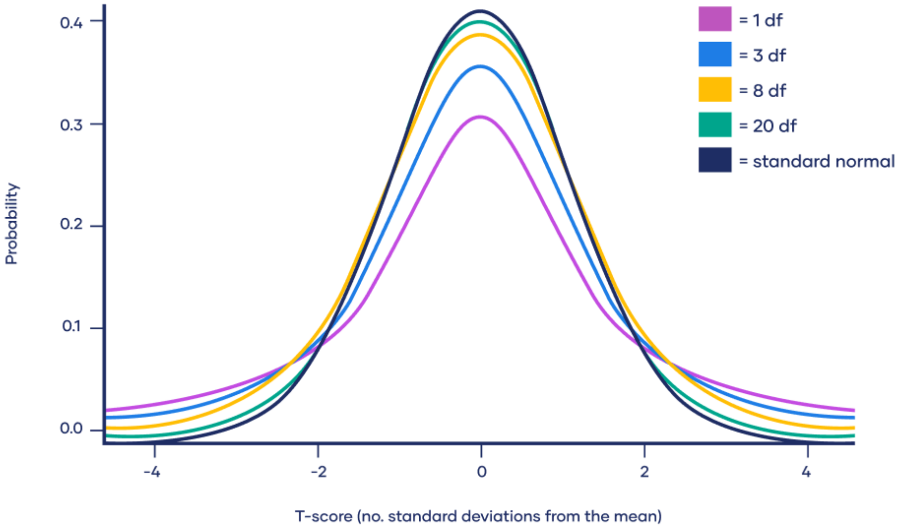
Bootstrap sample (sampling with replacement) – a sample of the original sample.
Ex. Sample with 62 patients. Randomly collect one sample from the 62 patients and find their mean, then randomly sample again from the 62 patients and find their mean until enough samples are taken.

The differences between bootstrap samples are entirely due to natural variability in the sampling procedure.

t-Distribution

a probability distribution used primarily in statistical inference when dealing with small sample sizes or when the population standard deviation is unknown.

As the degree of freedom increases the t-distribution approaches the standard normal distribution.



Formula	$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$	$\bar{x}$ = sample mean
		μ = population mean
		s = sample standard deviation
		n = sample size

Hypothesis Testing

t-test

One sample t-test:	Compare the mean of a single sample to a known value Null hypothesis: $\mu = \mu_0$ (hypothesized mean is the same as true mean) df = n - 1	$t = \frac{\overline{X}-\mu_0}{\frac{s}{\sqrt{n}}}$ $\overline{X}$ = sample mean μ_0 = Hypothesized population mean s = sample standard deviation n = sample size

Steps to Perform a T-Test

1. State Null hypothesis and Alternative hypothesis.
2. Calculate the Test Statistic
3. Determine Degrees of Freedom
4. Find the Critical Value:
   - Use a t-table for the given df and significance level(α).
5. Access the significance of hypothesis test:
   - Compare t-Value to Critical Value – If t-value exceeds the critical value, reject H0.
   - Compare p-value to significance level – If p-value smaller or equal to α, reject H0.

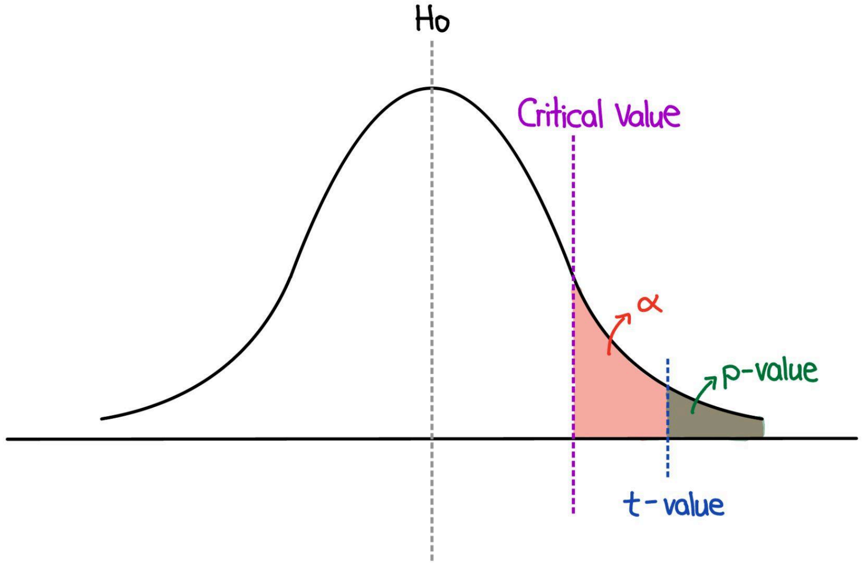
Alternative Hypothesis **H1 or Ha**	The research hypothesis Want to prove it true. Disprove null hypothesis → we have proved the alternative hypothesis
Null Hypothesis **H0**	Claim due to natural variability (no relationship between the variables) Is true if and only if the alternative is false.

Statistic of interest (test statistic) – the summary value of an experiment
Ex. There is a 30% difference of promotion rate between the control and treatment group. The difference in promotion rate is the statistic of interest.

p-value – is the probability of observing data at least as favorable to the alternative hypothesis as our current dataset, given that the null hypothesis were true.
Ex. There is a 30% difference between the control and treatment group. And the probability of having a 30% difference is only 2%. The 2% is the p-value.

Significance level (α) – how rare an event needs to be in order for the null hypothesis to be rejected.

Statistically significant – When the p-value is small (ex. less than a previously set threshold)
If p-value < significance level (ex. <5%) → reject H0
If p-value >= significance level→ fail to reject H0


H0
Critical Value
α
p-value
t-value