Terms

Population The entire group being researched (e.g. sample, study, target)

Sample A subset of the population, ideally random and large enough to be representative

Sample size The total number of subjects or observations in the sample, represented by n.

Reliability The consistency of the observed measurements from a sample. Data from a sample is considered reliable estimate of the sample statistic when there is very little bias or measurement error.

Validity The degree to which the sample statistic approximates the population parameter. A sample statistic is considered a valid estimate of the population parameter when the sample is large and/or representative of the study population.

Median The middle value in the data separating the top 50% from the bottom 50%. Found by arranging all values from lowest to highest and taking the middle value (or mean of the 2 middle values)

Quartile Each of the 4 equal groups into a which a population can be divided. The divisions between the quartiles are Q1 = 0.25 (25th percentile), Q2 = 0.50 (50th percentile, median), and Q3 = 0.75 (75th percentile).

Interquartile Range (IQR) The difference between the 3rd quartile (Q3 = 0.75) and the 1st quartile (Q1 = 0.25). The middle 50% of the data.

Mean Also called the average. The sum of all values in the sample divided by number of values in the sample. μ (mu) represents the mean of a population, and $\bar{x}$ represents the mean of a sample.

Variance Dispersion (spread) around the mean, determined by averaging the squared differences of all values from the mean. σ^2 (sigma squared) represents the variance of a population, and s^2 represents the variance of a sample.

Standard Deviation The square root of the variance. Also measures dispersion (spread) around the mean, but in the same units as the variable. σ (sigma) represents the standard deviation of a population, and s represents the standard deviation of a sample.

Central Limit Theorem The distribution of a sample statistic approximates the normal distribution $N\left(\text{population parameter}, \text{standard error}\right)$ as $n \to \infty$.

Sampling Distribution the distribution of theoretically possible sample statistics from all samples of size n that can be taken from a population

Standard Error The standard deviation of a sampling distribution. Reflects how variable a sample statistic is expected to be from sample to sample.

Confidence Interval A range of values within which you expect the "true" population parameter to fall if you repeated the study an infinite number of times. The confidence level is the percentage of samples whose confidence interval would capture the "true" population parameter. A confidence intervals upper and lower bounds are found by calculating point estimate $\pm$ critical value $\times$ standard error.

Critical Value The number which defines the upper and lower bounds of a confidence interval from a given distribution. Its value corresponds to the probabilities $\alpha/2$ and $1 - \alpha/2$.

Null Hypothesis There is no meaningful relationship in the data. Represented as H_0, gives the null distribution under which the hypothesis is tested.

Alternate Hypothesis There is something meaningful in the data. Represented as H_A, indicates whether the hypothesis test is one-sided (left- or right-tailed) or two-sided (both tails).

Type I Error The probability of rejecting the null hypothesis H_0 when H_0 is actually "true." Represented by α.

Type II Error The probability of failing to reject the null hypothesis H_0 when H_0 is not actually "true."

Formulas

Proportion

$$\hat{p} = \frac{\text{count (something)}}{\text{count (everything)}} = \frac{\text{count (something)}}{n}$$

Mean

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

Variance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} \qquad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Standard Deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Inference

Means

Table 1: Sample Statistics for Inference of Population Means

Measure	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Paired Difference in Means	$\bar{x}_{\text{difference}}$	$\mu_{\text{difference}}$
Difference in Means	$\bar{x}_1 - \bar{x}_2$	$\mu_1 - \mu_2$
Standard Deviation	s	σ

Assumptions

- ***Independence***: sample observations are independent (i.e. random sample).
- ***Sample size***: the sample size should be greater than 30 ($n \geq 30$) with no extreme outliers.
- ***Normality***: when the sample size n is small, observations come from a normally distribution population. This condition relaxes as $n \to \infty$.
- ***Validity***: sample statistics approximate the population parameters ($\bar{x} \approx \mu$, $s \approx \sigma$)

Single Mean ($\bar{x}$) - One-Sample t-test

- A ***1-sample t-test*** tests if the mean (μ) of a population is different from a null value (μ_0).
- Sample statistics $\bar{x}$ (mean) and s (standard deviation) are used to infer the sampling distribution of the mean $\bar{x} \sim \text{N}(\mu, SE_{\bar{x}})$.
- To account for using $s \approx \sigma$ in the standard error, confidence intervals and hypothesis tests are based on the T distribution (Student's t) with the parameter degrees of freedom (df) $= n - 1$.

Confidence Interval

The confidence interval for the mean $\bar{x}$ estimating μ is...

$$\bar{x} \pm T^*_{\text{df}} \times SE_{\bar{x}}$$

The standard error of $\bar{x}$ is...

$$\text{SE}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

The critical value from the t distribution with degrees of freedom $\text{df} = n - 1$ is...

$$T^*_{\text{df}} = T_{\text{df},\alpha/2} = T_{\text{df},1-\alpha/2}$$

A critical value is calculated from the t distribution in R using the function `qt()`. This function takes a probability `p` ($\alpha/2$ or $1 - \alpha/2$) and degrees of freedom `df` ($n - 1$).

```
qt(alpha/2, df = n-1)

qt(1-alpha/2, df = n-1)
```

Hypothesis Test

The null hypothesis of a 1-sample t-test states that the population mean μ is equal to some null value μ_0.

$$H_0 : \mu = \mu_0$$

The alternate hypotheses of a 1-sample t-test state that the population mean μ is greater than, less than, or not equal to some null value μ_0.

- $H_A : \mu < \mu_0$, left-tailed test (one-sided)
- $H_A : \mu > \mu_0$, right-tailed test (one-sided)
- $H_A : \mu \neq \mu_0$, two-tailed test (two-sided)

The null distribution

Paired Means

Difference in Means

Proportions

Table 2: Sample Statistics for Inference of Population Proportions

Measure	Sample Statistic	Population Parameter
Proportion	$\hat{p}$	p
Difference in Proportions	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$

Single Proportion