## Single Sample Test

Null Hypothesis, H0 is a statement about the **parameter** taking a certain value.

* It assumes it to be true, and rejects H0 if proven false (by the hypothesis test)
* Does not assure H0 is true

1. Point Estimate, , for the respective parameters p & μ

|  |  |  |
| --- | --- | --- |
|  | Proportion, p | Mean, μ |
| Assumptions | * Variable is categorical * data obtained using randomisation * sample size, n sufficiently large, sampling distribution of sample proportion , is approximately normal when null is true * np(1-p) ≥ 5 | * variable is quantitative * data obtained using randomisation * population distribution is approximately normal (if n is small)  OR  n ≥ 30, by CLT, is approximately normal |
| Test Statistic | * ) * Z ~ N(0, 1) Null distribution | * T = * T ~ **tn-1** T distribution w/ **n-1 degrees of freedom** |
| R Code |  | t.test(df\_column, mu , alternative = “two.sided”, conf.level = 0.95)  alternative <- c(“two.sided”, “less”, “greater”) |

1. Test Statistic, describes how far from point estimate is from the parameter value given in the null hypothesis

|  |  |
| --- | --- |
| **Presentation format** | **Examples** |
| assumptions are met or not | Assume :   1. data was sampled using randomisation  population distribution is normal 2. distribution of the population is normal  * histogram of values * qqplot (qqnorm(diff), qqline(diff)) * else, if n ≥ 30 or np(1-p) ≥ 5, by CLT, is approximately normal |
| state what is H0, H1 | Null hypothesis H0 : the two samples are from 2 populations with the same variance |
| Report test statistic | F-value = 0.67988  (T-value = -1.0275, with null distribution t1194.8) |
| Report p-value & conclusion | p-value = 1.33e-5. P-val < level of significance. **H0 is rejected**.  Variance of the hours spent on the internet for the 2 groups are unequal. |

**Interpretation of Confidence Interval**  
95% confident that interval (a,b) contains the unknown population parameter, p or μ

**Significance level, α not provided**

Conclude as: Since p-value = 0.0003. Data provides very strong evidence against H0 and supporting H1

**DO NOT REJECT** H0 since significance level not provided

**p-value Interpretation**

Conditional probability test statistic takes a value given that H0 is true (population parameter = a)

**Errors**

|  |  |
| --- | --- |
| **Type 1** | **Type 2** |
| Reject H0 when it is true | Do not reject H0 when it is false |
| α | Β  Power : 1- B, probability of rejecting when it is false |

Cannot reduce both types of errors simultanously

* when α small, H0 is rejected less often
* Higher likelihood that H0 is accepted when it is false

## Two-sample test

|  |  |
| --- | --- |
| Dependant Samples | Independent Samples |
| Observations in one sample give no clue about values in other sample | Two samples comprising same set of subjects  EX: before & after weight loss |

|  |  |  |
| --- | --- | --- |
| **Independent Samples** | | |
| Variance | Equal Variance | Unequal Variance |
| Assumptions | * quantitative response for 2 groups * 2 independent samples, random sampling. **randomised** experiment * population distribution of each group is approximately normal (small n)  1. check distribution of plots – if is skewed 2. if skewed, if n ≥ 30, CLT, is approximately normal, hence violation of normality will not affect test  * equal/different population variance for each group (by var.test) | |
| Check histogram to ensure normality assumption fo the data distribution of the sample is valid | |
| Test for variance equal | Whether the 2 samples are from 2 populations with the same variance.  Test using var.test(x, y), an F test  H0 : 2 samples are from 2 populations with the same variance  H0 : 2 samples are from 2 populations with equal variance  H1 : 2 samples are from 2 populations with unequal variances  p-value smaller than 0.05, test is significant, the 2 variance are unequal     * can always choose the group that has a smaller variance as the first group (numerator group). This way F test is left-tailed, then check if p-value < significance level * var.test(x,y) & var.test(y,x) gives different F-stat * However, final p-value is the same | |
| Test Statistic | Sp estimates σ²   * Under H0, T follows a **t-distribution** with **(n1 + n2) − 2** degrees of freedom * n1,n2 not required to be identical   H0 : µ1 - µ2 = 0  H1 : µ1 - µ2 ≠ 0  t.test(x, y, **alternative = ‘two.sided’,** **var.equal = TRUE**, conf.level = 0.95) | If H0 is true, then T follows a t-distribution with a complicated number of degrees of freedom (might not be an integer). Let’s call it df  H0 : µ1 - µ2 = 0  H1 : µ1 - µ2 ≠ 0  t.test(x, y, alternative = ‘two.sided’, **var.equal = FALSE**, conf.level = 0.99)  **One-sided test**   * Greater than/ Less than * only look at p-value, if p-value lower than alpha, reject |
|  | Default Params: var.equal:FALSE, alternative: “two.sided”, paired = FALSE | |

**Dependant Samples**H0 : µ1 = 0

|  |  |
| --- | --- |
| **Test for Differences in R** | **Test for Matched Samples in R** |
| t.test(diff, mu = 0, alternative = ‘greater’, conf.level = 0.99)  H0 :  H1 : | t.test(Yes, No, alternative = ‘greater’, paired = TRUE, conf.level = 0.99)  H0 :  H1 :  The 2 tests are equal |

# Confidence Intervals

|  |  |  |
| --- | --- | --- |
| **Terms** | **Points** | **Formula** |
| Point estimates, | Vary from sample to sample   * not mean of population * taking random sample each time |  |
| Margin of Error | accuracy of point estimate in estimating a parameter | quantile x Standard Error(test statistic) |
| Standard Error | **estimated** standard deviation of a sampling distribution |  |
| Interval Width | Size of the interval  S can be estimated from similar study/equal probability ()  **D = 2 x margin of error** | 2 x quantile x se  For a (1- α)100% CI with width ≤ D |
| t-Distribution | df ≥ 30, t-distribution approximately equal to normal distribution  can approximate using qnorm() | tn-1,0.975 x |

* increases with n (SE decreases, more certain)
* variance of underlying distribution (cannot control)
* increases with large confidence level (wider interval)

# Estimation

|  |  |  |
| --- | --- | --- |
|  | **Parameter** | **Estimator** |
|  | μ, p |  |
| Test Statistic |  |  |
| SD  (estimator) |  | Standard Error, SE( estimator ) |

**Data Distribution**

Distribution of individual observation, X in sample

**Sampling Distribution**

Distribution of point estimates derived from each sample

**Approximating Sampling Distribution of test statistic to normal distribution**

|  |  |  |
| --- | --- | --- |
|  | mean | Proportion |
|  | n ≥ 30 | np(1-p) ≥ 5 |
|  | Variance of sampling distribution is exact if population is normally distribution | Variance of sampling distribution is estimated using test statistic,, |

Diagram, engineering drawing, schematic

Description automatically generated

* estimated using s, hence use tn-1
* in p case, estimated using , hence not required

**R code:**  
prop.table(table(data))  
linear reg:  
sr = rstandard(model)  
hist(sr)  
qqnorm(sr)  
qqline(sr)

Table

Description automatically generated with low confidence