# Comparison Between two populations

Very important in analytics. This type of analysis removes much of the extraneous variations in data collection. We are comparing 1:1 values.

Chubby chunky example

**Matched Samples:** Same person we got two different readings

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Men | | Women | | Differences Men | Differences Women |
| Sl No. | Before | After | Before | After |
| 1 | 72 | 68 | 98 | 92 | 4 | 6 |
| 2 | 84 | 78 | 89 | 84 | 6 | 5 |
| 3 | 83 | 75 | 76 | 65 | 8 | 11 |
| 4 | 96 | 92 | 84 | 77 | 4 | 7 |
| 5 | 78 | 86 | 87 | 95 | -8 | -8 |
| 6 | 83 | 79 | 93 | 88 | 4 | 5 |
| 7 | 102 | 95 | 90 | 84 | 7 | 6 |
| 8 | 74 | 72 | 79 | 76 | 2 | 3 |
| 9 | 82 | 81 | 81 | 77 | 1 | 4 |
| 10 | 80 | 87 | 68 | 77 | -7 | -9 |
| 11 | 88 | 83 | 74 | 70 | 5 | 4 |
| 12 | 76 | 74 | 70 | 66 | 2 | 4 |
| 13 | 79 | 72 | 103 | 94 | 7 | 9 |
| 14 | 70 | 75 | 72 | 80 | -5 | -8 |
| 15 | 91 | 85 | 79 | 73 | 6 | 6 |

## Quadrants of operation

|  |  |  |
| --- | --- | --- |
|  | H0 | H1 |
|  | µD <= 0: i.e. people will not lose weight | µD > 0: i.e. people will lose weight |
| Don’t reject H0 | Good decision | Implication: Don’t join program don’t lose money |
| Reject H0 and Conclude H1: | Implication: Join the program but lose money and gain weight | Good decision. |

## Rejection criteria

I will reject H0 if I find a sample whose differences in weight is higher than critical value as determined by the confidence interval.

## Calculations

|  |  |  |
| --- | --- | --- |
|  | **Men** | **Women** |
| **Average of differences** | 2.4 | 3 |
| **Number of observations (n)** | 15 | 15 |
| **Std. dev. Of differences** | 5.124172407 | 6.21059 |
| **Confidence interval** | 0.95 | 0.95 |
| **Error** | 0.05 | 0.05 |
| **t\_critical for two tailed** | 1.761310136 | 1.76131 |
| **Rated weight difference** | 0 | 0 |
| **Upper limit of single right tailed** | 2.330311285 | 2.82438 |
| **T value for given sample mean** | 1.813982687 | 1.870829 |
| **Probability of commiting Type 1 error (P value)** | 0.045582715 | 0.041209 |

## Conclusion

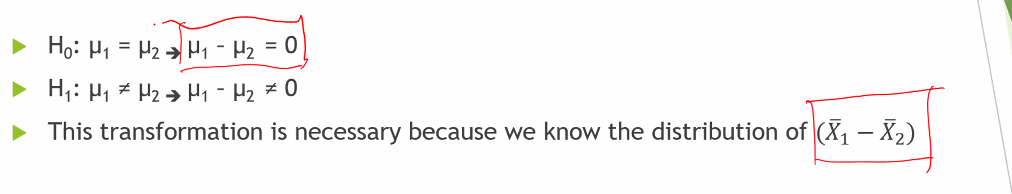
As per the given sample, for both men and women we reject the null hypothesis that the program is worthless as P value < significance level 0.05 in both cases. However the probability of committing type 1 error is lesser in case of women since Pmen > Pwomen based on this sample. Hence Chubby told his wife to go to the program.

# Implication of P value

P value is the error we are likely to be committing if we reject null hypothesis based on the given sample.

# Unpaired Samples

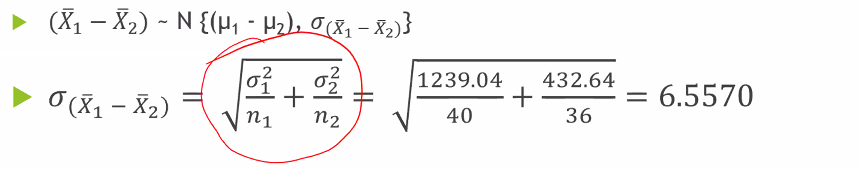
There are often cases when paired samples are hard to get. Suppose we want to measure the average time taken by a machine to produce some parts. To get a paired sample we have to create the same type of parts from two different machines which is impractical.

We know the distribution of X1\_bar and X2\_bar. These are unmatched samples and independent. Then we go for estimate difference.X

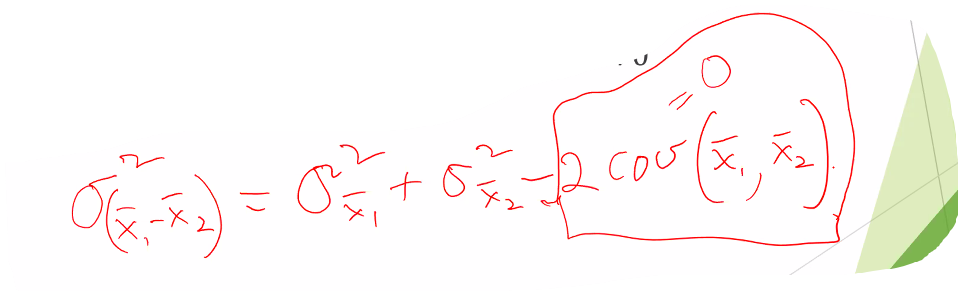
For linear combination of X1\_bar and X2\_bar each being normally distributed is also normally distributed.

Thus if we have matched samples we have to take difference in observations.

If we have unmatched samples we take difference in estimates.



Need to prove the σ(X1\_bar and X2\_bar).

….. self proof required.

Here assumed COV(X1\_bar and X2\_bar) ~ 0.

## Proof Var(A-B) = Var(A) + Var(B) - 2\*Covariance(A,B)

|  |  |  |
| --- | --- | --- |
| A | B | A-B |
| a1 | b1 | a1 - b1 |
| a2 | b2 | a2 - b2 |
| … |  |  |
| an | bn | an - bn |

Now we know that Var(X) = where

Thus V(A - B) =

Also:

So V(A - B) =

=

=

Thus V(A - B) = Var(A) + Var(B) - 2\*Cov(A, B).

We make a simplistic assumption that Cov(A, B) ~ 0.

Thus Var(A - B) = Var(A) + Var(B)

Thus σ2(A - B) = σ2(A) + σ2(B)

Thus σ(A - B) =

Practice problem.