Hypothesis testing

It is the process to assess the plausibility of a hypothesis by using sample data. We make hypothesis about the population parameter and test that with sample data. If obtained sample statistics falls into a region far way from the hypothesized parameter (this region is called region of rejection), then hypothesis can be rejected.

Procedure of Hypothesis testing

STEP1: State the hypothesis about a parameter to be tested. (Specify Null and Alternate hypothesis)

STEP2: Based on the hypothesis, obtain the sampling distribution of a statistics.

STEP3: Decide the region of rejection (Set the significance level α)

STEP4 : Calculate test statistics and p value (If the calculated value of the statistics falls into the region of rejection, the null hypothesis is said to be rejected)

Z-test versus T-test

Both tests are used to compare means of two groups or to compare sample mean to population mean. Point to be noted is both t test and z test will work with mean only if distribution is normal.

Z test should be used when population variance is given, and T test should be used when population variance is unknown. For smaller sample set (<30), mostly we will use t test.

If the sample size is large enough, then the Z test and T test will conclude with the same results. For a large sample size, sample variance will be a better estimate of population variance so even if population variance is unknown, we can use the Z test using sample variance.
Similarly, for a large sample, we have a high degree of freedom. And since t-distribution approaches the normal distribution, the difference between the Z

score and T score is negligible. Therefore, if sample size is large(>30-40), z test and t test both will give almost similar output.

Z test:

$$Z = \frac{\overline{x}-\mu}{\sigma_{\overline{x}}} \quad , \; {\sigma^2}_{\overline{x}} = \frac{\sigma^2}{n}$$

where, σ is population variance
$\overline{x}$ is sample mean
µ is population mean
n is number of sample size

T test:

$$t = \frac{\overline{x}-\mu}{s_{\overline{x}}} \quad , \; {s^2}_{\overline{x}} = \frac{s^2}{n}$$

where s is sample variance and df = n-1
$\overline{x}$ is sample mean
µ is population mean
n is number of sample size

T-Test for single group:

T test for single group helps to compare sample mean to population mean.

For example, if average score in an examination is 230 and we randomly select 15 students with the given scores. We want to know whether sample mean differs from the population mean.

Scores = [78,120,89,92,135,86,99,108,111,97,140,89,86,102,111]

Step1: Find mean and standard deviation of sample data.
mean = 102.86, std = 18.133
Step2: Specify Null and Alternate hypothesis.
Null hypothesis: There is no difference in sample and population mean.
Alternate hypothesis: There is significant difference between sample and population mean.

Step3: Analysis using t test(t value and p value)

t value = $\frac{102.86-230}{18.133/\sqrt{15}}$ = -27.1376

Step4: Calculate 95% confidence interval of population mean.
As sample size=15, degree of freedom = n-1 = 14
With df=14, $t_{0.975}$ = 2.145
Confidence Interval,

$\mu_{upper} = \overline{X} + s_{\overline{x}}\, t_{0.975}$ = 102.86 +2.145(4.6855) = 112.91

$\mu_{lower} = \overline{X} - s_{\overline{x}}\, t_{0.975}$ = 102.86 -2.145(4.6855) = 92.81

Step5: Calculate effect size(Using cohen's d)

Cohen's d ($\hat{d}$) = $\frac{|\overline{X} - \mu|}{s}$

If, 0 < $\hat{d}$ < 0.2 , it is small effect

0.21 < $\hat{d}$ < 0.8 , it is medium effect

$\hat{d}$ > 0.80 , it is large effect

For given scenario, $\hat{d}$ = |(102.86 – 230) | /18.133 = 7.0115
Therefore, it is a large effect.

R code:

```r
score <- c(78,120,89,92,135,86,99,108,111,97,140,89,86,102,111)
mean_score <- mean(score)
std_score <- sd(score)


#T test fro single sample

t.test(score,conf.int=.95, mu=230)
# p value is 1.64e-13 which is very less than 0.05 , therefore we can
# reject null hypothesis and can say that there is a signifcant difference
# in sample and population mean.

#We can see confidence interval is (92.8244 , 112.90888)


#Cohen's d
delta <- abs(mean_score-230)/std_score
delta
#As cohen's d is 7.01 which is greater than 0.8 . Therefore, it is large effect.
```

```
> score <- c(78,120,89,92,135,86,99,108,111,97,140,89,86,102,111)
> mean_score <- mean(score)
> std_score <- sd(score)
> t.test(score,conf.int=.95, mu=230)

        One Sample t-test

data:  score
t = -27.153, df = 14, p-value = 1.648e-13
alternative hypothesis: true mean is not equal to 230
95 percent confidence interval:
  92.82445 112.90888
sample estimates:
mean of x
 102.8667

> #Cohen's d
> delta <- abs(mean_score-230)/std_score
> delta
[1] 7.010813
>
```

T test for two groups:

There are two different designs for T test with two groups:

1) Dependent samples(Correlated groups)
2) Independent samples

Correlated group design:

This design involves 2 different observations from the same group of subjects or two separate sets of subjects that have a meaningful connection. Here, sample size will be same for both groups.

T-test for correlated groups will be same as one sample. The only difference is that it will consider difference scores for all calculations.

Therefore, t = $\frac{\overline{D}-\mu_D}{s_D/\sqrt{N}}$

where, D id difference score

$\overline{D}$ is mean of sample difference score

μ_D is mean of population of difference scores

s_D is standard deviation of the sample difference scores

N is number of difference scores

For example: We have to see if there is any difference between scores before and after one activity or is there any effect of that activity on scores. We have scores for 10 students before and after that activity.

Subject	Score_before_activity	Score_after_activity
1	72	79
2	64	71
3	57	60
4	65	68
5	77	71
6	68	66
7	78	78
8	81	84
9	73	76
10	60	63

Step1:
Null hypothesis- There is no difference between scores before and after the activity. $\mu1 = \mu2$
alternate hypothesis- There is significant difference between scores before and after the activity. $\mu1 \neq \mu2$

Step2:

Subject	Score_before_activity	Score_after_activity	Difference(D)	$(D - \overline{D})^2$
1	72	79	7	24.01
2	64	71	7	24.01
3	57	60	3	0.81

4	65	68	3	0.81
5	77	71	-6	65.61
6	68	66	-2	16.81
7	78	78	0	4.41
8	81	84	3	0.81
9	73	76	3	0.81
10	60	63	3	0.81
			21	138.9

N = 10 , $\overline{D} = \frac{\sum D}{N}$ = 21/ 10 =2.1

Step3:

Assume sampling distribution of mean difference score is normal with mean=0 , therefore, $\mu_{\overline{D}} = \mu_D = 0$ and $s_{\overline{D}} = {}^{s_D}/_{\sqrt{N}}$

$$S_D = \sqrt{\frac{SS_D}{N-1}} = \sqrt{\frac{\sum\left(D - \overline{D}\right)^2}{N-1}} = \sqrt{\frac{138.9}{9}} = \sqrt{15.43} = 3.928$$

t = $\frac{\overline{D}-\mu_D}{{}^{s_D}/_{\sqrt{N}}}$ = (2.1 – 0)/ (3.928/3.162) = 1.69

Step4:

Confidence interval:

$$\mu_{upper} = \overline{D} + s_{\overline{D}}\, t_{crit}$$

$$\mu_{upper} = \overline{D} + s_{\overline{D}}\, t_{crit}$$

Cohen's d $(\hat{d}) = \frac{|\overline{D}|}{s_D}$

R code:

```r
# T test for correlated samples

scores_before_activity <- c(72,64,57,65,77,68,78,81,73,60)
scores_after_activity <- c(79,71,60,68,71,66,78,84,76,63)
data1 = data.frame(scores_before_activity, scores_after_activity)

data1

t.test(data1$scores_after_activity, data1$scores_before_activity,
       paired =TRUE, conf.level = 0.95)
# As p value is 0.1252which is greater than 0.05, we can not reject
# null hypothesis.
```

```
> scores_before_activity <- c(72,64,57,65,77,68,78,81,73,60)
> scores_after_activity <- c(79,71,60,68,71,66,78,84,76,63)
> data1 = data.frame(scores_before_activity, scores_after_activity)
> data1
   scores_before_activity scores_after_activity
1                      72                    79
2                      64                    71
3                      57                    60
4                      65                    68
5                      77                    71
6                      68                    66
7                      78                    78
8                      81                    84
9                      73                    76
10                     60                    63
> t.test(data1$scores_after_activity, data1$scores_before_activity,
+        paired =TRUE, conf.level = 0.95)

	Paired t-test

data:  data1$scores_after_activity and data1$scores_before_activity
t = 1.6904, df = 9, p-value = 0.1252
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7102997  4.9102997
sample estimates:
mean of the differences 
                    2.1 

> 
```

Independent group design:

This design involves two unrelated(independent) groups randomly sampled from their respective populations. Here, samples do not have to be the same size. So basically, we want to know whether the two population means are equal.

For example: Suppose that students are assigned to either a maths class(n1=4) or computer class(n2=5) and calculated their scores. We want to know whether computer class has any impact on score.
For maths class: scores = [67, 58, 72, 69]
For computer class: scores = [73,69,75,71,84]

Step1: Specify null hypothesis
H0 – Computer class does not affect the scores. ($\mu_1 = \mu_2$)

Step2: Calculate pooled variance
Maths:

X = Score	$(X - \overline{X})^2$
67	0.25
58	72.25
72	30.25
69	6.25
	SS_1 = 109

$\overline{X}$ = 66.5

Computer:

X = Score	$(X - \overline{X})^2$
73	1.96
69	29.16
75	0.36
71	11.56
84	92.16
	SS_2 =135.2

$\overline{X}$ = 74.4

Degree of freedom (df) = (n1-1)+(n2-1) = n1+n2-2 =4+5-2 = 7
Therefore, pooled variance (s_w^2) = $(SS_1 + SS_2)$/df = 34.88

Step3: Calculate t statistics

$$t = \frac{(\overline{X_1} - \overline{X_2})}{\sqrt{s_W^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{(66.5-74.4)}{\sqrt{34.88\left(\frac{1}{4}+\frac{1}{5}\right)}} = -\frac{7.9}{3.96} = -1.995$$

Step4: Calculate effect size and confidence interval.

$$\hat{d} = \frac{|(\overline{X_1} - \overline{X_2})|}{s_W}$$

Confidence interval:

$$(\overline{X_1} - \overline{X_2}) \pm S_{(\overline{X_1} - \overline{X_2})} t_{crit}$$

R code:

```
#T test for independent samples

scores <- c(67, 58, 72, 69,73,69,75,71,84)
group <- c(1,1,1,1,2,2,2,2,2)

data2 = data.frame(group, scores)
data2


t.test(scores~ group, data =data2, var.equal=TRUE)

# As we can see from results, p value = 0.0864 which is greater than
# Therefore, we can not reject null hypothesis and we will say class
# doesnot affect the scores.
```

```
> scores <- c(67, 58, 72, 69,73,69,75,71,84)
> group <- c(1,1,1,1,2,2,2,2,2)
> data2 = data.frame(group, scores)
> data2
  group scores
1     1     67
2     1     58
3     1     72
4     1     69
5     2     73
6     2     69
7     2     75
8     2     71
9     2     84
> t.test(scores~ group, data =data2, var.equal=TRUE)

        Two Sample t-test

data:  scores by group
t = -1.9939, df = 7, p-value = 0.0864
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -17.268978   1.468978
sample estimates:
mean in group 1 mean in group 2 
           66.5            74.4 

> |
```

T-test only work from 1 group and two groups.

If we have more than two groups, it seems that we can run multiple t tests to compare the means from different groups and if we found that at least one of the t ttests was significant, we can say there is group effect. But the multiple t tests inflate Type1 error for the test of overall group effect(overall α).

If tests are independent, we can calculate exact overall α as:

P(at least one significant result) = 1 – p(no significant results) = $1 - (1 - \alpha_1)(1 - \alpha_2)\ldots.(1 - \alpha_k)$

For example: 3 groups involve 3 t tests. If the tests are independent and α = 0.5 is used for each test, then overall alpha is 1-0.95*0.95*0.95 = 0.143

But tests are often dependent and it is not easy to calculate the exact overall α. Therefore if we have more than two groups, we will go for ANOVA test.

For R code:
https://github.com/palak-j/hypothesis-testing--T-tests