Mid 2

Chi-sq test of Goodness of fit

⇒ Observed counts are different from expected counts

→ How well the obs data fit the expected dist

H_0:- There is no inconsistency b/w obs and expect counts

H_A: There is inconsistency b/w obs and expect counts

General formula of test statistic is

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

Chi-square (χ^2) statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$ where k = total no. of cells

Why

→ Any standardized diff that is squared will now be positive

→ Diff that already looked unusual will become much larger after being squared

→ To determine χ^2 statistic is considered High or not we must describe its distribution

→ Chi-sq has one param called **degree of freedom (df)**, which influences the shape, center and spread of the dist

$$df = k-1 \quad \therefore k = \text{no. of cells}$$

p-value < 0.05 reject H_0

p-value > 0.05 ~~accept H_A~~ do not reject H_0

Condition for chi-sq

→ Independance:

Each case that contributes a count to the table must be independent of all other cases in the table.

→ Sample size:

Each particular scenerio must have atleast 5 expected cases

→ df > 1

Degree of freedom must be greater than 1

Example

Candidate	Obs	Reported
Ahmedinajad	338	63.29%
Moysari	136	34.10%
Minor Candidate	30	2.61%
Total	504	100%

H_0: The obs counts from the poll follow the same dist as reported votes

H_A: The obs counts from the poll do not follow the same dist as reported votes

Candidate	Obs	Exp	
a) Ahmedinajad	338	63.29%	504 * 0.6329 = 319
b) Moysari	136	34.10%	504 * 0.3410 = 172
c) Minor candidate	30	2.61%	504 * 0.261 = 13
Total	504		504

O	E	$(O-E)^2$	$(O-E)^2/E$
338	319	361	1.13
136	172	1296	7.53
30	13	289	22.23

$\Sigma (O-E)^2/E = 30.89$

$\chi^2_{df} = 2$

p-value $= P(\chi^2_{df=2} > 30.89)$

p-value $= 0$

$0 < 0.05$ therefore reject H_0.

Chi-sq test of Independance

Class	Grades	Popular	Sports
4	63	31	25
5	88	57	33
6	96	55	32

H_0: Grade do not vary by grade
H_A: Grade vary by grade

$$\chi^2_{df} = \sum_{i=1}^{k} (O-E)^2/E$$

$$\therefore df = (R-1)(C-1)$$

Rows Cols

Class	Grades	Popular	Sports	Total
4	63 61	31 35	25 23	119
5	88 91	55 52	33 33	176
6	96 95	54 54	32 34	183
Total	247	141	90	$\Sigma\Sigma$ 478

Expect Count = $\frac{(\text{Row total})(\text{col. Total})}{\text{table total}}$

Obs	Exp	$(O-E)^2$	$(O-E)^2/E$
63	61	4	0.07
31	35	16	0.46
25	23	4	0.17
88	91	9	0.10
55	52	9	0.17
33	33	0	0
96	95	4	0.04
55	54	1	0.02
32	34	4	0.12

$\Sigma\Sigma = 1.1688$

$df = (3-1)(3-1)$
$= 2 \times 2$
$= 4$

p-value = $\chi^2_{df=4} > 1.16$

p-value = 0.89 > 0.05

therefore we cannot reject p-value

One sample mean with t-distribution

Condition

→ Independance

rows are independant

→ Sample size / Skew

Sample distribution

must be skewed

⇒ We do not know σ and n is too small to assum s is reliable estimate for σ

⇒ As long as dist are independant and the population dist is not extremely skewed then large sample would ensure

→ Sample dist of the mean is nearly normal

→ The estimd of SE as $s/\sqrt{n}$ is reliable

When to use

⇒ When population standard deviation is unknown

⇒ Dist has a bell shaped, but its tails are thicker than the normal model

Note

Null value is zero because is the null hypothesis we set $\mu_{null} = 0$

18358

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

point estimate = $\bar{x}$ = mean

$$SE = \frac{s_{df}}{\sqrt{n}}$$ — Stad deviation

$df = n - 1$ ∴ n is sample size

Example

	x	$x - \bar{x}$	$(x - \bar{x})^2$
1	696		311875600
2	1104		297700516
3	1037		
4	1889		
5	1911		
6	2416		
7	2761		
8	4382		
9	1839		
10	321		
	$\sum x = 18358$		

← Ans 2

variance = Ans / 10

Stand Deviation = $\sqrt{\text{variance}}$

Z test statistics

⇒ Sample size of atleast 30, one can use the Z-distribution in place of t-dist

Pair t-test or Dependent sample t-test

When to use

=> When your data values are paired measurements. For example you may have before and after measurements

Conditions

=> Subjects must be independant measurement of one subject doesn't affect the other

=> Each of the paired measurements must be obtained from the same subject. for example before and after weight of a smoker should be of the same person

=> The measured differences are normally distributed

$$SE = sd / \sqrt{n}$$

or $\frac{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}}{(n-1)(n)}$ sample size

$$df = n-1$$

$$t = \frac{\text{Average Diff}}{SE}$$

$H_0 = Md_{diff} = 0$

$H_A = Md_{diff} \neq 0$

Example

Student	u_1 Exam 1 Score	u_2 Exam 2 Score	diff $X = u_1 - u_2$	ΣX^2
Bob	63	69	6	36
Ning	65	65	0	0
Tim	56	62	6	36
"	100	91	-9	81
"	88	78	-10	100
"	83	87	4	16
"	77	79	2	4
"	92	88	-4	16
"	90	85	-5	25
"	84	92	8	64
"	68	69	1	1
"	74	81	7	49
"	87	84	-3	9
"	64	74	11	121
"	71	84	13	169
"	88	82	-6	36
			$\Sigma X = 21$	$\Sigma X^2 = 763$

$\bar{X} = 21/16$

$= 1.31$

$n = 16$ (sample size)

$$SE = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{(n-1)n}} = \sqrt{\frac{763 - \frac{(21)^2}{16}}{(16-1)16}}$$

$= 1.95$

$$t = \frac{\bar{X}}{SE} = \frac{1.31}{1.75} = 0.750$$

$$df = n-1 = 16-1 = 15$$

p-value = $T_{df=15} > 0.750$

p-value 2.131

$0.750 < 2.131$ we cannot reject H_0

Difference in two means

$$T_{df} = \frac{\text{point estimate} - \text{Null value}}{\text{~~Stand~~ SE}}$$

point estimate = $\bar{X}_1 - \bar{X}_2$

$$SE = \sqrt{\frac{\Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n_1}}{(n_1-1)(n_1)} + \frac{\Sigma X_2^2 - \frac{(\Sigma X_2)^2}{n_2}}{(n_2-1)(n_2)}}$$

OR

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad df = \min(n_1-1, n_2-1)$$

=> To compare means of 2 groups we use a Z or T test

=> To compare means of 3+ groups we use ANOVA or F test

ANOVA

Conditions

=> data are a simple random sample from less than 10% of the population

=> data should be independent

=> Observation within each group should be nearly normal

$$F = \frac{\text{variability bet. groups}}{\text{variability within group}}$$

$df_g = k-1$ where k is the no. of groups

$df_T = n-1$, where n is the total sample size

error = $df_E = df_T - df_g$

$$SSG = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$$

$\bar{x}_i$ = avg of each group

$\bar{x}$ = overall mean

$$SST = \sum_{i=1}^{k} (\bar{x}_i - \bar{x})^2$$

$$SSE = SST - SSG$$

$$MSG = \frac{SSG}{df_g}$$

$$MSE = \frac{SSE}{df_E}$$

$$F = \frac{MSG}{MSE}$$