Chapter 10

Hypothesis testing

⇒ for categorical data ⇐

Data

qualitative
نوعية
"Categorical"

quantitative
كمية

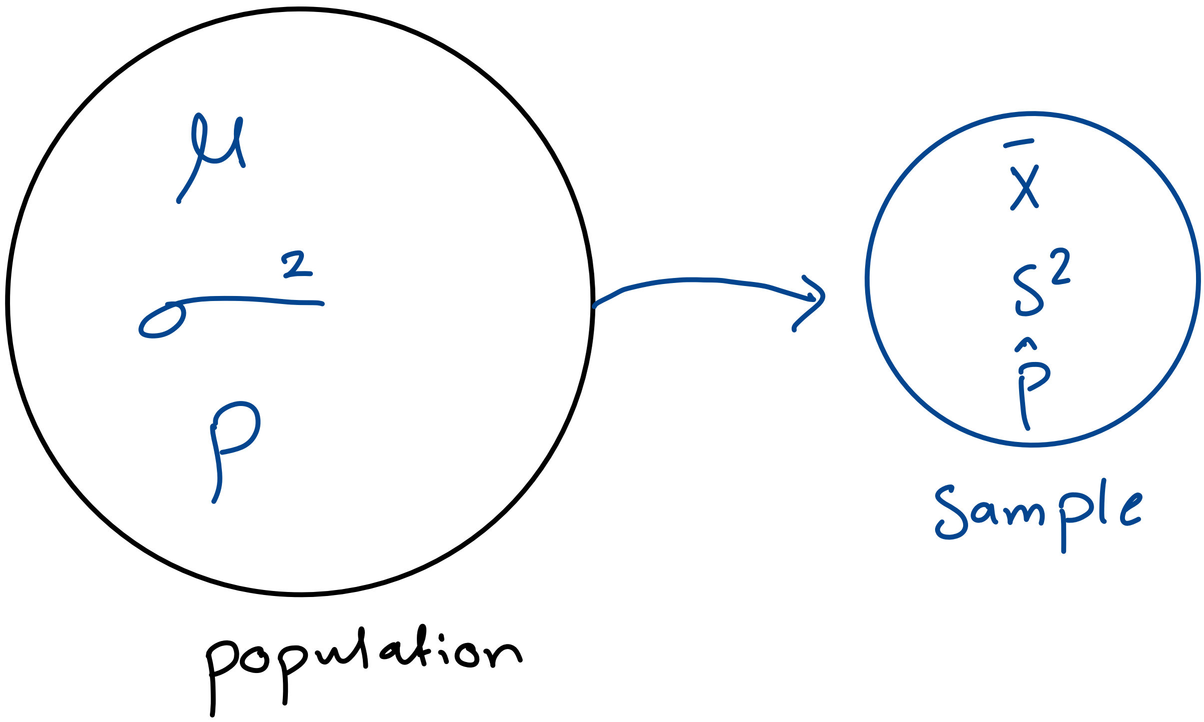
Examples

Categorical data:

⇒ Blood Group (A, B, O, AB)

⇒ Sick / not sick

⇒ Male / Female



* Test of hypothesis between 2 Sample proportions $(\hat{P}_1 - \hat{P}_2)$

normal method $\begin{pmatrix} n_1 P_1 q_1 > 5 \\ m P_2 q_2 > 5 \end{pmatrix}$

Contigency table

① $$Z = \frac{(\hat{P}_1 - \hat{P}_2) - 0}{\sqrt{p^* q^* \left(\frac{1}{n} + \frac{1}{m}\right)}}$$

② $$Z_{Corr} = \frac{|\hat{P}_1 - \hat{P}_2| - \left(\frac{1}{2n} + \frac{1}{2m}\right)}{\sqrt{p^* q^* \left(\frac{1}{n} + \frac{1}{m}\right)}}$$

① The normal method $\begin{pmatrix} np_1q_1 > 5 \\ mp_2q_2 > 5 \end{pmatrix}$

$$H_0: P_1 = P_2 \quad \text{Vs.} \quad H_1: P_1 \neq P_2$$

test stat

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - \cancel{(P_1 - P_2)}^{\,0}}{\sqrt{p^* q^* \left(\frac{1}{n} + \frac{1}{m}\right)}}$$

p^* : pooled proportion $= \dfrac{x + y}{n + m}$

$\boxed{x : n * \hat{P}_1}$ $\boxed{y = m * \hat{P}_2}$

$\boxed{q^* = 1 - p^*}$

$$Z_{Corr} = \frac{|\hat{P}_1 - \hat{P}_2| - \left(\frac{1}{2n} + \frac{1}{2m}\right)}{\sqrt{p^* q^* \left(\frac{1}{n} + \frac{1}{m}\right)}}$$

Example Two types of medication for hives are being tested to determine if there is a difference in the proportions of adult patient Reactions. Twenty out of a random sample of 200 adults given medication "A" still had hives 30 minutes after taking the medication. Twelve out of another Random sample of 200 adults given medication "B" still had hives 30 mins after taking the medication. Test using 1% significance level when: $\boxed{\alpha = 0.01}$

① no Continuity Correction applied

$\hat{P}_1 = \frac{x}{n} = \frac{20}{200} = 0.1$

$\hat{P}_2 = \frac{y}{m} = \frac{12}{200}$

$H_0 : P_1 = P_2 \quad Vs. \quad H_1 : P_1 \neq P_2$

test stat

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - 0}{\sqrt{p^* q^* * \left(\frac{1}{n} + \frac{1}{m}\right)}}$$

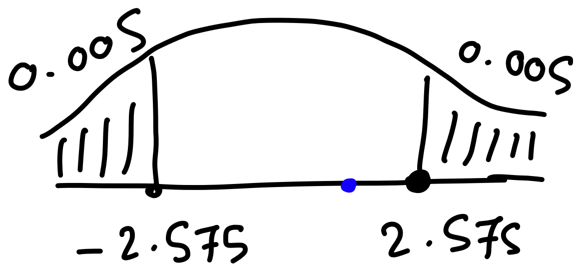
$p^* = \frac{x+y}{n+m} = \frac{20+12}{200+200} = 0.08$

$= 0.06$

$$= \frac{0.1 - 0.06}{\sqrt{0.08 * 0.92 * \left(\frac{1}{200} + \frac{1}{200}\right)}} = 1.47$$

$\alpha = 0.01$

$\boxed{\frac{\alpha}{2} = 0.005}$



so we accept H_0 and Reject H_1

② Applied Z_{corr} on your answer

$$H_0: P_1 = P_2 \quad Vs. \quad H_1: P_1 \neq P_2$$

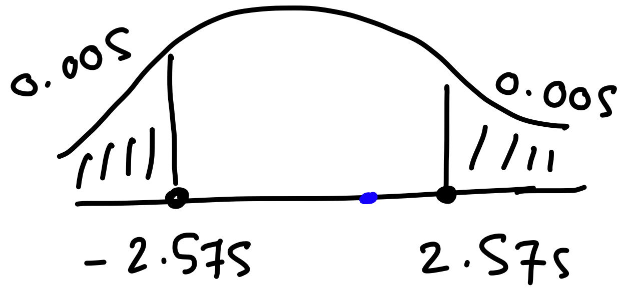
test stat

$$Z_{corr} = \frac{|\hat{P}_1 - \hat{P}_2| - \left(\frac{1}{2n} + \frac{1}{2m}\right)}{\sqrt{p^* q^* \left(\frac{1}{n} + \frac{1}{m}\right)}}$$

$$= \frac{|0.1 - 0.06| - \left(\frac{1}{200*2} + \frac{1}{200*2}\right)}{\sqrt{0.08 * 0.92 * \left(\frac{1}{200} + \frac{1}{200}\right)}} = 1.29$$

$\alpha = 0.01$

$\frac{\alpha}{2} = 0.005$



So we accept H_0 and Reject H_1

كتاب

Example A study looked at the effect of OC use on heart disease in women (40-44 y/o). The Research found that among (5000) Current OC users at baseline, 13 women developed Myocardial infarction (MI) over 3 years period whereas among (10 000) non-OC users 7 developed an MI over a 3-years period. Asses the statistical significance of the Results. (use Corrected)

$H_0: P_1 = P_2 \quad Vs. \quad H_1: P_1 \neq P_2$

<u>test stat</u>

$$Z_{Corr} = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n} + \frac{1}{2m}\right)}{\sqrt{p^* q^* \left(\frac{1}{n} + \frac{1}{m}\right)}}$$

$$= \frac{|0.0026 - 0.0007| - \left(\frac{1}{2*5000} + \frac{1}{2*10\,000}\right)}{\sqrt{0.00133 * 0.99867 \left(\frac{1}{5000} * \frac{1}{10000}\right)}}$$

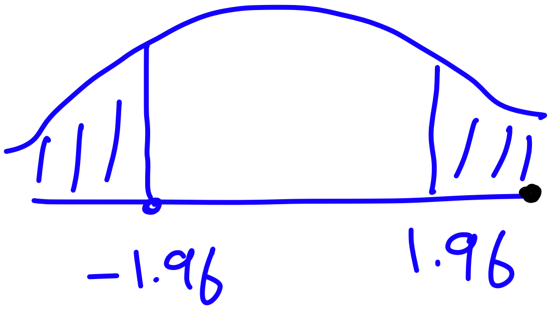
$$= 2.77$$

$$\hat{P}_1 = \frac{13}{5000} = 0.0026 \quad , \quad \hat{P}_2 = \frac{7}{10\,000} = 0.0007$$

$$p^* = \frac{x+y}{n+m} = \frac{13+7}{5000+10000} = 0.00133$$

$\alpha = 0.05$

$\frac{\alpha}{2} = 0.025$

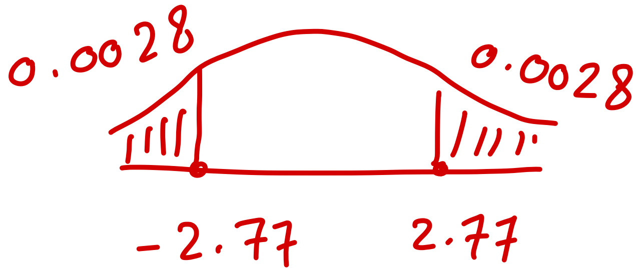


$\therefore$ we Reject H_0 and accept H_1

P-value $= 2 * 0.0028$

$= 0.0056$

$\Rightarrow$ highly significant



Guidelines for Judging the Significance of a *p*-Value

If $.01 \le p < .05$, then the results are *significant*.
If $.001 \le p < .01$, then the results are *highly significant*.
If $p < .001$, then the results are *very highly significant*.
If $p > .05$, then the results are considered *not statistically significant* (sometimes denoted by NS).
However, if $.05 < p < .10$, then a trend toward statistical significance is sometimes noted.

0.001 • 0.01 0.05 0.10

Example Police officers in new york City can stop a driver who is not wearing their seat belt. In Boston, police officers can issue citations to driver for not wearing their seat belts only if the driver has been stopped for another violation Data from Random samples of femal in 2002 is summarized as the following:

City	Drivers	wearing seat belts
Boston	117	68
New york	220	183

Is there compelling evidence to conclude a difference in Rate of drivers wear their seat belts in Boston as compared to new york? (Assume Continuity Correction is applied, use $\alpha = 0.05$)

$$H_0: P_1 = P_2 \quad Vs \quad H_1: P_1 \neq P_2$$

Test stat

$$Z_{Corr} = \frac{|\hat{P}_1 - \hat{P}_2| - \left(\frac{1}{2n} + \frac{1}{2m}\right)}{\sqrt{p^* q^* \left(\frac{1}{n} + \frac{1}{m}\right)}}$$

$$= \frac{|0.58 - 0.83| - \left(\frac{1}{2*117} + \frac{1}{2*220}\right)}{\sqrt{0.74 * 0.26 * \left(\frac{1}{117} + \frac{1}{220}\right)}}$$

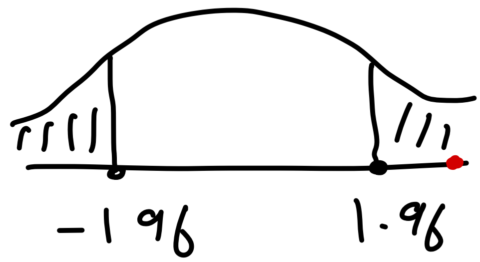
$$= 4.85$$

$$\hat{P}_1 = \frac{68}{117} = 0.58 \quad , \quad \hat{P}_2 = \frac{183}{220} = 0.83$$

$$p^* = \frac{x+y}{n+m} = \frac{68+183}{117+220} = 0.74$$

$\alpha = 0.05$

$\frac{\alpha}{2} = 0.025$



$\therefore$ we Reject H_0 and accept H_1

② The contingency table

Ⓐ 2 X 2 table

Example observed table

	Rt-hand	Lt-hand	total	
Males	43 O_{11}	9 O_{12}	52	Row margin
Females	44 O_{21}	4 O_{22}	48	Row margin
total	87	13	100	
	Column margin	Column margin	grand total	

Row ciP / Column

عمود

Expected table

	Rt-hand	Lt-hand	total
males	$\frac{52*87}{100} = 45.24$ E_{11}	$\frac{52*13}{100} = 6.76$ E_{12}	52
Females	$\frac{48*87}{100} = 41.76$ E_{21}	$\frac{48*13}{100} = 6.24$ E_{22}	48
total	87	13	100

NOTE

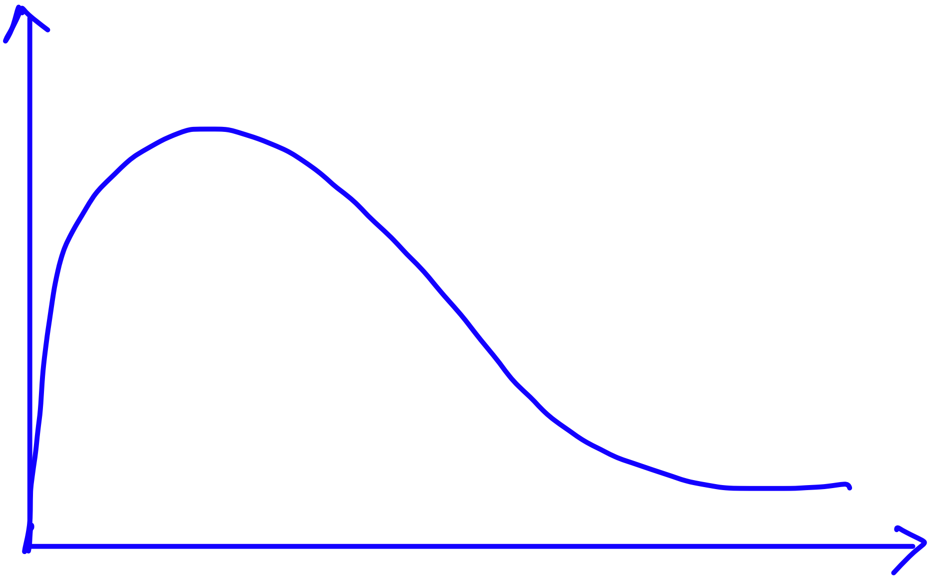
$$\boxed{E = \frac{R * C}{\text{grand total}}}$$

$H_0: P_1 = P_2$ Vs. $H_1: P_1 \neq P_2$

test stat

χ^2

* Chi-squared



⇒ Skewed to the Right

⇒ All values are positive

⇒ degree of freedom

$$d.f = (R-1)(C-1)$$

<u>NOTE</u> : Always d.f in 2 x 2 contingency

table is equal to 1

---

* General notes

1. Always the test is Right
 failed test

2. test statistics

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(O_{11}-E_{11})^2}{E} + \frac{(O_{12}-E_{12})^2}{E} + \frac{(O_{21}-E_{21})^2}{E} + \frac{(O_{22}-E_{22})^2}{E}$$

$$\chi^2_{Corr} = \sum \frac{\left(|O-E| - \frac{1}{2}\right)^2}{E}$$

3. Always the Expected values

are more than 5

Example The following table lists Results from an experiment designed to test the ability of dogs to use their extraordinary sense of smell to detect malaria in samples of Children's socks. The accompanying information shows the following:

	Malaria was present	Malaria wasn't present	total
Dog was correct	123	131	254
Dog was wrong	52	14	66
Total	175	145	320

Identify the test statistics and

the p-value, and then state the conclusion about the null hypothesis.

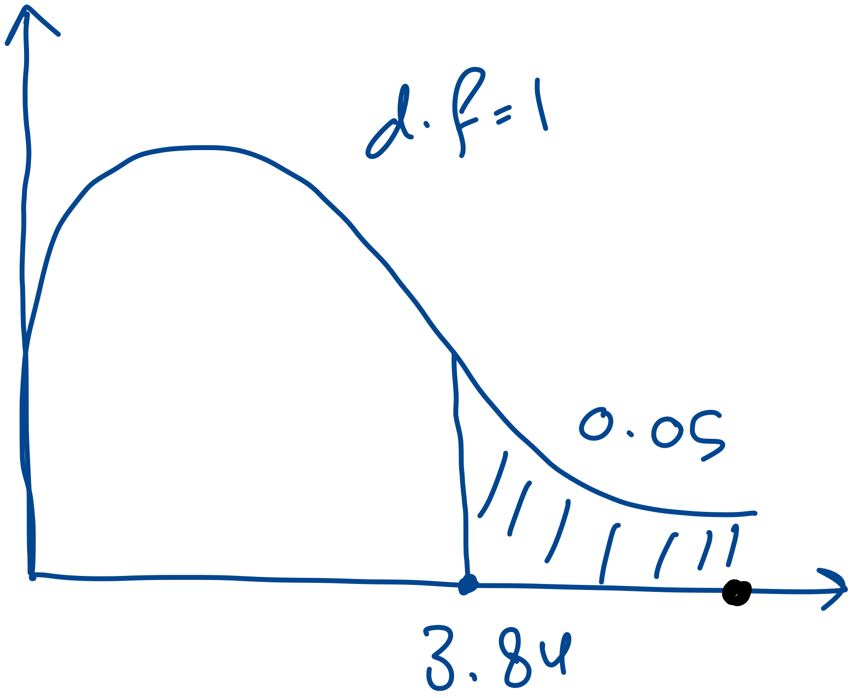
$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

	Malaria present	Malaria wasn't present
Dog was correct	$\frac{254 * 175}{320}$ 138.91	$\frac{254 * 145}{320}$ 115.09
Dog was wrong	$\frac{66 * 175}{320}$ 36.09	$\frac{66 * 145}{320}$ 29.91

$$= \frac{(123 - 138.91)^2}{138.91} + \frac{(131 - 115.09)^2}{115.09}$$

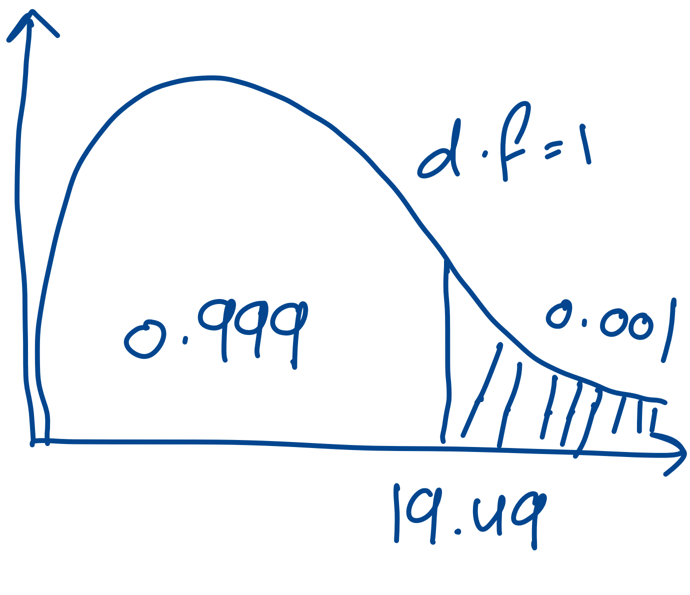
$$+ \frac{(52 - 36.09)^2}{36.09} + \frac{(14 - 29.91)^2}{29.91} = 19.49$$

$\alpha = 0.05$



So we Reject H_0 and accept H_1

~~ظل~~ P_value



P_value = 0.001

$$\chi^2_{Corr} = \frac{(|123 - 138.91| - \frac{1}{2})^2}{138.91} + \frac{(|131 - 115.09| - \frac{1}{2})^2}{115.09}$$

$$+ \frac{(|52 - 36.09| - \frac{1}{2})^2}{36.09} + \frac{(|14 - 29.91| - \frac{1}{2})^2}{29.91}$$

$$= 19.59$$

Example: Suppose we want to know if the Rate of smoking in males is different from Females in a sample of 203 Jordanian the observed values set as the following:

(use $\alpha = 0.001$)

	Smoker	Non Smoker	total
Males	72	44	116
Females	34	53	87
total	106	97	203

	Smoker	Non Smoker	total
Males	60.57	55.43	116
Females	45.43	41.57	87
total	106	97	203

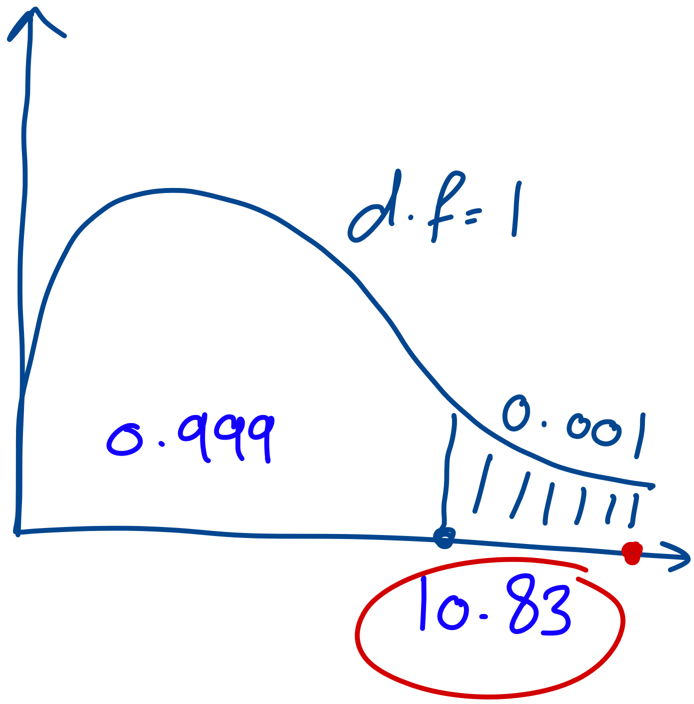
$$\chi^2_{Corr} = \sum \frac{\left(\left|O-E\right| - \frac{1}{2}\right)^2}{E}$$

$$= \frac{\left(\left|72-60.57\right| - \frac{1}{2}\right)^2}{60.57} + \frac{\left(\left|44-55.43\right| - \frac{1}{2}\right)^2}{55.43}$$

$$+ \frac{\left(\left|34-45.43\right| - \frac{1}{2}\right)^2}{45.43} + \frac{\left(\left|53-41.57\right| - \frac{1}{2}\right)^2}{41.57}$$

$$= \underline{12.74}$$

∴ we Reject H_0 and accept H_1



Critical value

مثال

Example Compute the expected table for the breast cancer data shown in the following table:

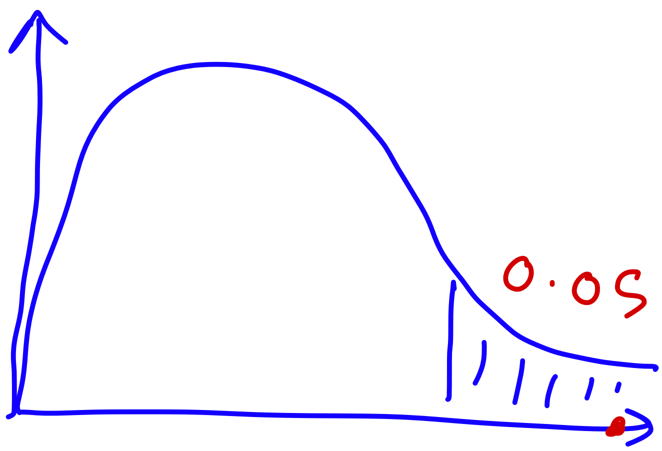
	≥ 30	< 29	
case	683	2537	3220
control	1498	8747	10245
	2181	11284	13465

Expected

	≥ 30	< 29	
Case	521.6	2698.4	3220
Control	1659.4	8585.6	10245
	2181	11284	13465

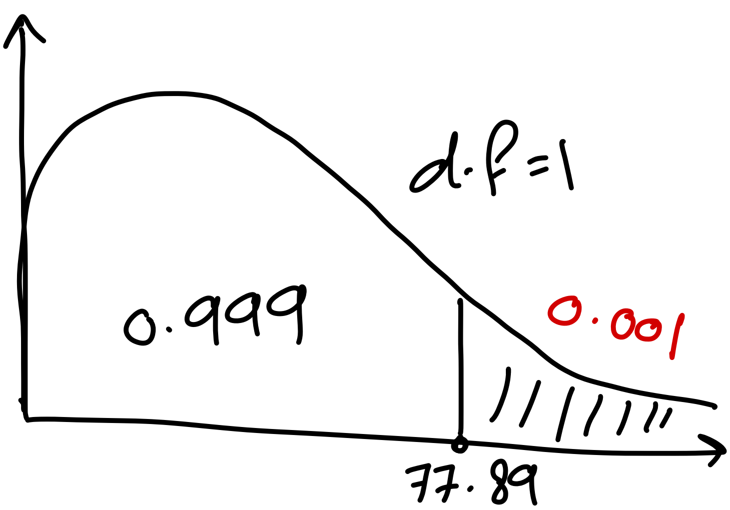
$$\chi^2_{Corr} = 77.89$$

∴ we Reject H_0 and accept H_1



P-value = 0.001

⇒ highly significant



كتاب

Example Assess the OC-MI data for statistical significance, using contingency table approach?

2 × 2 contingency table for the OC–MI data in Example 10.6

	MI incidence over 3 years		
OC-use group	Yes	No	Total
Current OC users	13	4987	5000
Never-OC users	7	9993	10,000
Total	20	14,980	15,000

الحل

2 × 2 contingency table for the OC–MI data in Example 10.6

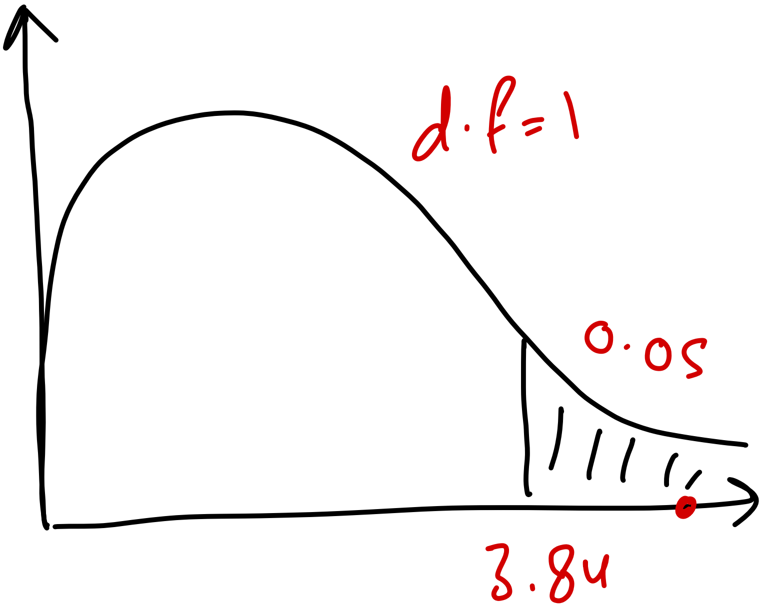
	MI incidence over 3 years		
OC-use group	Yes	No	Total
Current OC users	6.7	4993.3	5000
Never-OC users	13.3	9986.7	10,000
Total	20	14,980	15,000

$$\chi^2_{Corr} = \frac{\left(|13-6.7| - \frac{1}{2}\right)^2}{6.7} + \frac{\left(|4987 - 4993.3| - \frac{1}{2}\right)^2}{4993.3}$$

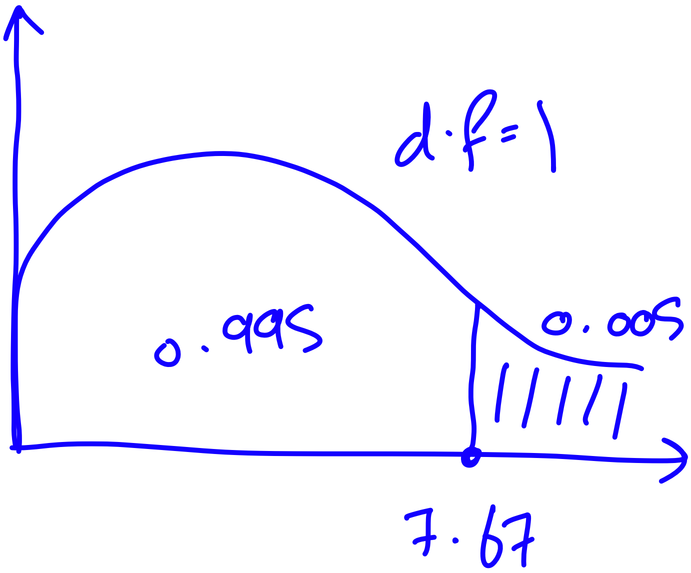
$$+\frac{\left(|7-13.3|-\frac{1}{2}\right)^2}{13.3}+\frac{\left(|9993-9986.7|-\frac{1}{2}\right)^2}{9986.7}$$

$$= 7.67$$

$\therefore$ we Reject H_0 and accept H_1



P-value = 0.005



Guidelines for Judging the Significance of a p-Value

If $.01 \le p < .05$, then the results are *significant*.
- If $.001 \le p < .01$, then the results are *highly significant*.
If $p < .001$, then the results are *very highly significant*.
If $p > .05$, then the results are considered *not statistically significant* (sometimes denoted by NS).
However, if $.05 < p < .10$, then a trend toward statistical significance is sometimes noted.

$0.001 \leftarrow 0.005 \rightarrow 0.01$

$\Rightarrow$ highly significant

NOTES

① The purpose of Contingency table is to summarize a large set of data

② χ^2_{Corr} is called Yates - corrected chi squared.

③ Always the Expected values are more than 5

④ The contingency table is often used to determine if the two variable have an association

⑤ Ho : if they are independent

H_1 : if they are dependent

---

* R x C Contingency table

	X	y	Z	K
a				
B				
C				

1. $$\boxed{E = \frac{R * C}{Total}}$$

2. test stat : Always χ^2 العادية

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

3. degree of freedom in RXC

table is calculated as the following:

$$(R-1) * (C-1)$$

(4) The Conditions of the table:

(A) No cell has an exepected < 1

(B) No more than $\frac{1}{5}$ of the cells have expected value less than 5

Example Assess the statistical significance in 300 persons, giving the following:

Table of Observed Values

Qualification / Marital Status	Middle School	High School	Bachelor's	Master's	Ph.D	Total
Never married	18	36	21	9	6	90
Married	12	36	45	36	21	150
Divorced	6	9	9	3	3	30
Widowed	3	9	9	6	3	30
Total	39	90	84	54	33	300

H_0: Marital status independent from qualification

H_1: " " dependent " "

31

Expected

Qualification / Marital Status	Middle School	High School	Bachelor's	Master's	Ph.D
Never Married	$\frac{90 \times 39}{300} = 11.7$	$\frac{90 \times 90}{300} = 27$	25.2	16.2	9.9
Married	19.5	45	42	27	16.5
Divorced	3.9	9	8.4	5.4	3.3
Widowed	3.9	9	8.4	5.4	3.3

Test stat

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

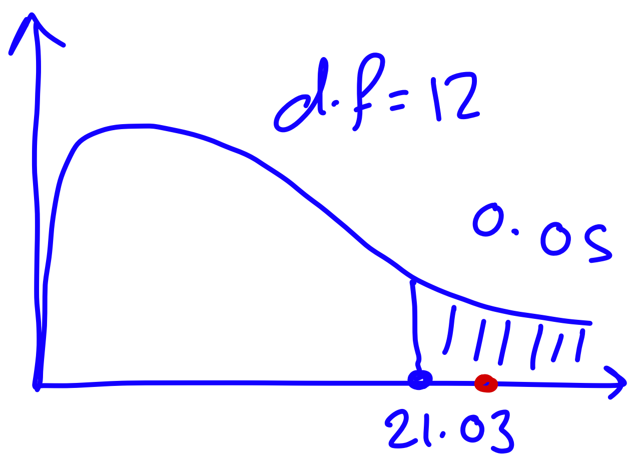
$$= \frac{(18-11.7)^2}{11.7} + \dots\dots + \frac{(3-3.3)^2}{3.3}$$

$$= 23.57$$

$$d.f = (4-1)(5-1)$$

$$= 3 * 4$$

$$= 12$$



∴ we Reject H_0 and accept H_1

حتي

Example Assess the statistical significance of the data between 2 variables, the age of first birth and the prevelance of breast cancer.

TABLE 10.16 Data from the international study in Example 10.4 investigating the possible association between age at first birth and case–control status

Case–control status	Age at first birth <20	20–24	25–29	30–34	≥35	Total
Case	320	1206	1011	463	220	3220
Control	1422	4432	2893	1092	406	10,245
Total	1742	5638	3904	1555	626	13,465
% cases	.184	.214	.259	.298	.351	.239

Source: Based on *WHO Bulletin*, *43*, 209–221, 1970.

Test stat

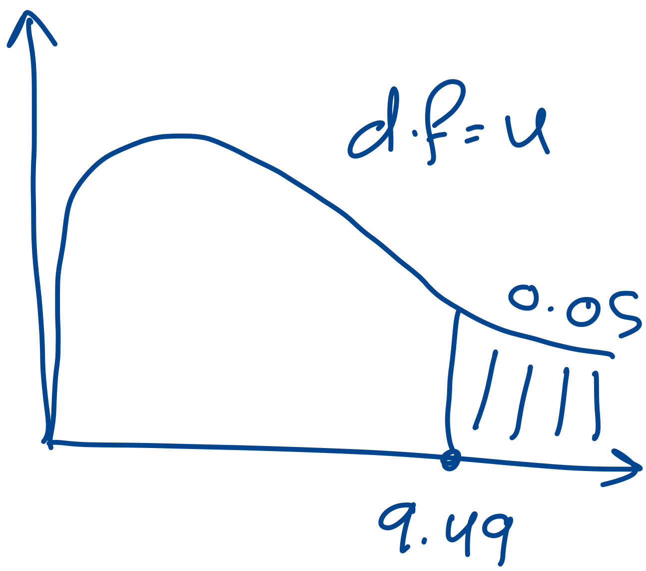
$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(320 - 416.6)^2}{416.6} + \cdots\cdots + \frac{(406 - 476.3)^2}{476.3}$$

$= \underline{130.3}$

$d.f = (2-1)(5-1)$

$= 1 * 4 = 4$

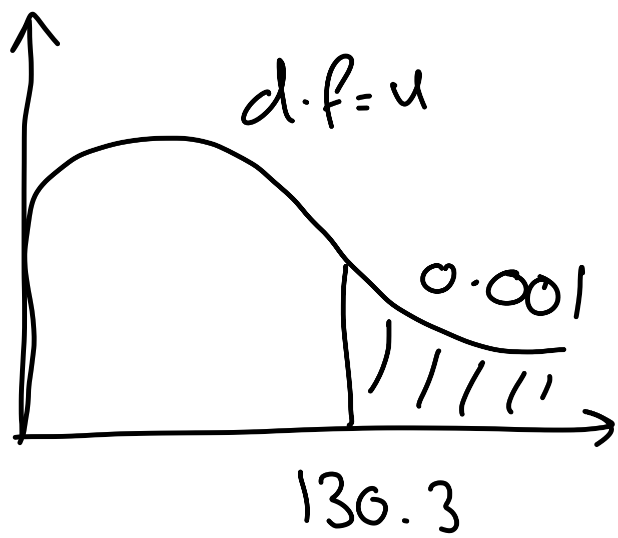


∴ we Reject H_0 and accept H_1

(There is an association between the first birth age and breast cancer)

P-value = 0.001

⇒ highly significant



Example Determine to the 5% Significance level whether school and grade are dependent.

		Grade			Totals
		A	*B*	*C*	
School	*X*	18	12	20	**50**
	Y	26	12	32	**70**
Totals		**44**	**24**	**52**	**120**

Ho: School is independent from the Grade

H_1: School is dependent on grade

Expected

		Grade			Totals
		A	*B*	*C*	
School	*X*	$\frac{50\times 44}{120} = 18.33$	$\frac{50\times 24}{120} = 10$	$\frac{50\times 52}{120} = 21.67$	**50**
	Y	$\frac{70\times 44}{120} = 25.67$	$\frac{70\times 24}{120} = 14$	$\frac{70\times 52}{120} = 30.33$	**70**
Totals		**44**	**24**	**52**	**120**

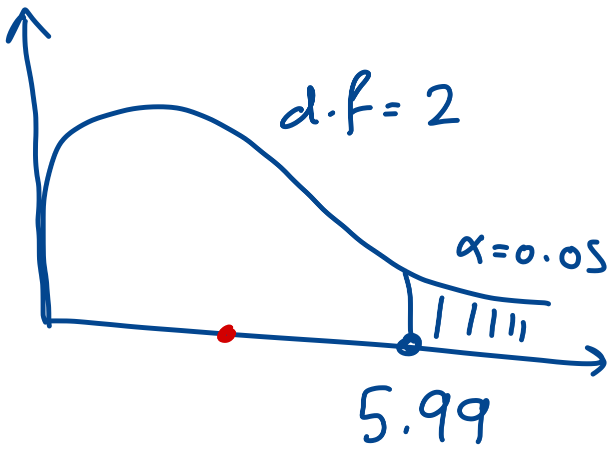
$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(18-18.33)^2}{18.33} + \frac{(12-10)^2}{10} + \frac{(20-21.67)^2}{21.67}$$

$$+\frac{(26-25.67)^2}{25.67}+\frac{(12-14)^2}{14}$$

$$+\frac{(32-30.33)^2}{30.33}=0.916$$

$\therefore$ we accept H_0
and reject H_1



---

* Goodness of fit test
(chi-squared)

$\Leftarrow$ إختبار حسن الملائمة

⇒ Approximation of discrete Random variable to continous Random variable

$$D \longrightarrow n$$

$\Rightarrow P(X < 16)$ [Discrete]

$\Rightarrow P(X \le 15)$ (- يمين)

$\Rightarrow P(X \le 15.5)$ [Continuity correction]

NOTE

1. $P(X \le a) \Rightarrow P(X \le a + 0.5)$
2. $P(X \ge a) \Rightarrow P(X \ge a - 0.5)$
3. $P(a \le X \le b) \Rightarrow P(a - 0.5 \le X \le b + 0.5)$

Examples

① $P(X > 18)$ [Discrete]

$= P(X \geq 19)$

$= P(X \geq 18.5)$

② $P(18 < X < 26)$

$= P(19 \leq X \leq 25)$

$= P(18.5 \leq X \leq 25.5)$

③ $P(18 \leq X < 26)$

$= P(18 \leq X \leq 25)$

$= P(17.5 \leq X \leq 25.5)$

4 $P(18 < X \leq 25)$

$= P(19 \leq X \leq 25)$

$= P(18.5 \leq X \leq 25.5)$

Example If the $\mu = 20$, $\sigma^2 = 16$

Find:

1 $P(X < 26)$

$= P(X \leq 25) \Rightarrow P(X \leq 25.5)$

$\Rightarrow P\left(Z \leq \frac{25.5 - 20}{4}\right)$

$= P(Z \leq 1.38) = 0.9162$

2 $P(18 < X \leq 26)$

$= P(19 \leq X \leq 26) \Rightarrow P(18.5 \leq X \leq 26.5)$

$$= P(X \le 26.5) - P(X \le 18.5)$$

$$= P\left(Z \le \frac{26.5 - 20}{4}\right) - P\left(Z \le \frac{18.5 - 20}{4}\right)$$

$$\vdots \qquad\qquad \vdots$$

EXAMPLE 10.46

Hypertension Diastolic blood-pressure measurements were collected at home in a community-wide screening program of 14,736 adults ages 30–69 in East Boston, Massachusetts, as part of a nationwide study to detect and treat hypertensive people [6]. The people in the study were each screened in the home, with two measurements taken during one visit. A frequency distribution of the mean diastolic blood pressure is given in Table 10.20 in 10-mm Hg intervals.

We would like to assume these measurements came from an underlying normal distribution because standard methods of statistical inference could then be applied on these data as presented in this text. How can the validity of this assumption be tested?

TABLE 10.20 Frequency distribution of mean diastolic blood pressure for adults 30–69 years old in a community-wide screening program in East Boston, Massachusetts

Group (mm Hg)	Observed frequency	Expected frequency	Group	Observed frequency	Expected frequency
<50	57	69.0	≥80, <90	4604	4538.6
≥50, <60	330	502.5	≥90, <100	2119	2545.9
≥60, <70	2132	2018.4	≥100, <110	659	740.4
≥70, <80	4584	4200.9	≥110	251	120.2
			Total	14,736	14,736

$\bar{x} = 80.68$

$s = 12$

$$= P(X < 50)$$

$$= P(X \le 49)$$

$$= P(X \le 49.5) \Rightarrow P\left(Z \le \frac{49.5 - 80.68}{12}\right)$$

$$= P(Z \le -2.60)$$

$$= 0.0047$$

$$0.0047 * 14736 \simeq 69$$

$$\Rightarrow P(50 \le X < 60)$$

$$= P(50 \le X \le 59)$$

$$\Rightarrow P(49.5 \le X \le 59.5)$$

$$\Rightarrow P(X \le 59.5) - P(X \le 49.5)$$

$$= P\left(Z \le \frac{59.5 - 80.68}{12}\right) - P\left(Z \le \frac{49.5 - 80.68}{12}\right)$$

$$= P(Z \le -1.77) - P(Z \le -2.598)$$

$$= 0.0337$$

$$0.0337 * 14736 \simeq 502.5$$

Test stat

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(57-69)^2}{69} + \cdots\cdots + \frac{(251-120.2)^2}{120.2}$$

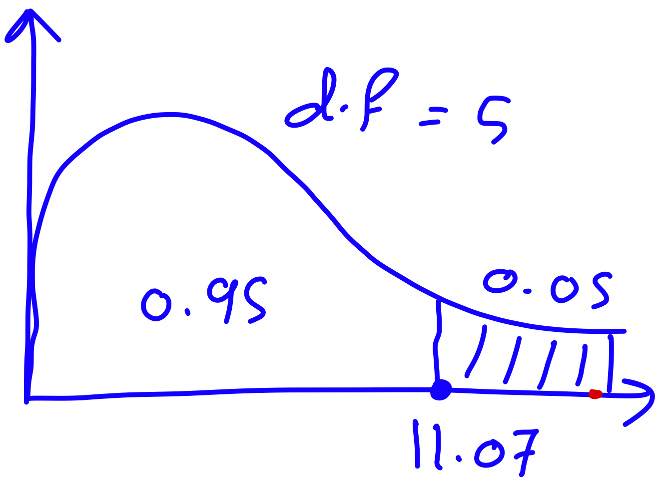
$$= 326.2$$

$\alpha = 0.05$

$$\boxed{D.f = g - k - 1}$$

$= 8 - 2 - 1$

$= 5$



$\therefore$ we Reject H_0 and accept H_1

⇒ normal method doesn't provide an adequate fit to the data.

NOTES

① we study the fit of the test to a data

② Expected
- Ⓐ equality
- Ⓑ Continuity correction
- Ⓒ probability
- Ⓓ Probability * grand total

③ $$\chi^2 = \sum \frac{(O - E)^2}{E}$$

④ $D.f = g - k - 1$

Example The mean weights of a sample of 200 patients is 52 KGs and the standard deviation is 3 KGs.

weight	$w < 45$	$45 \le w < 50$	$50 \le w < 55$	$55 \le w < 60$	$w \geq 60$
frequency	12	44	82	53	9

we would like to assume that these measurments came from the normal distribution. How can the validity of this assumption be tested?

weight	$\cancel{w} x < 45$	$45 \le x < 50$	$50 \le w < 55$	$55 \le w < 60$	$w \geq 60$
frequency	12	44	82	53	9
Expected	1.24	39.42	118.7	39.42	1.24

$$P(x < 45)$$

$$= P(x \le 44)$$

$$= P(x \le 44.5) \Rightarrow P\left(Z \le \frac{44.5 - 52}{3}\right)$$

$$= P(Z \le -2.5)$$

$$= 0.0062$$

$$0.0062 * 200 = 1.24$$

$$\Rightarrow P(45 \le X < 50)$$

$$\Rightarrow P(45 \le X \le 49)$$

$$\Rightarrow P(44.5 \le X \le 49.5)$$

$$= P(X \le 49.5) - P(X \le 44.5)$$

$$= P\left(Z \le \frac{49.5-52}{3}\right) - P\left(Z \le \frac{44.5-52}{3}\right)$$

$$= P(Z \le -0.83) - P(Z \le -2.5)$$

$$= 0.2033 \quad - \quad 0.0062$$

$$= 0.1971$$

$$0.1971 * 200 = 39.42$$

$$x^2 = \sum \frac{(O-E)^2}{E}$$

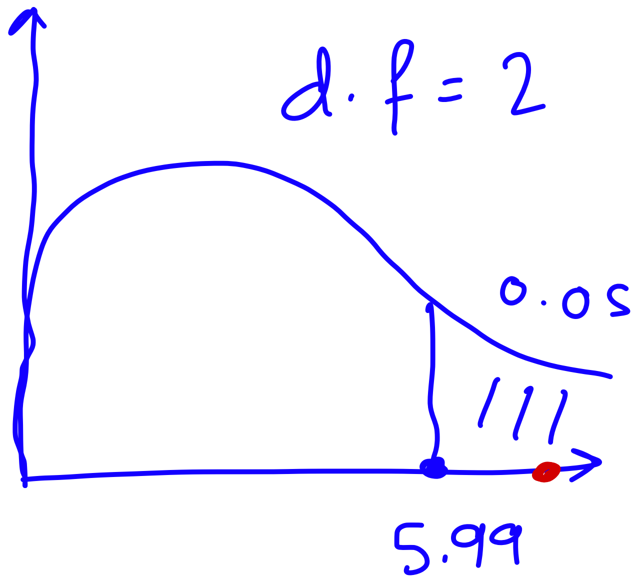
$$= \frac{(12-1.24)^2}{1.24} + \text{------} + \frac{(9-1.24)^2}{1.24}$$

$$= 158.49$$

$$d.f = g-k-1$$

$$= 5-2-1$$

$$= 2$$



So we Reject H_0 and accept H_1

Chapter 11

Regression and Correlation Method

⇒ for quantitative data ⇐

1. Scatter plot
2. Correlation coefficient
3. Hypothesis testing
4. Confidence interval

---

* Correlation ⇐

(علاقة)

Graphical
رسومات

numerical
بالأرقام

① Graphical (Correlation)

" Scatter plot "

كتاب

Estregon

Estriol

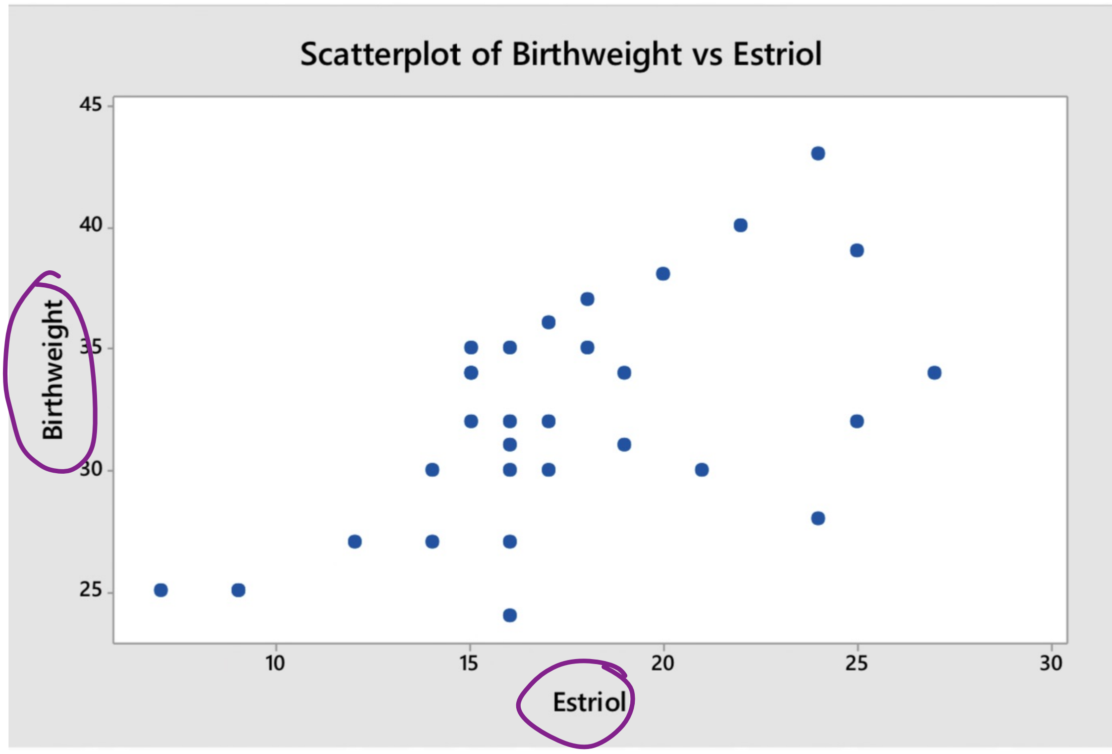
⇒ uterus

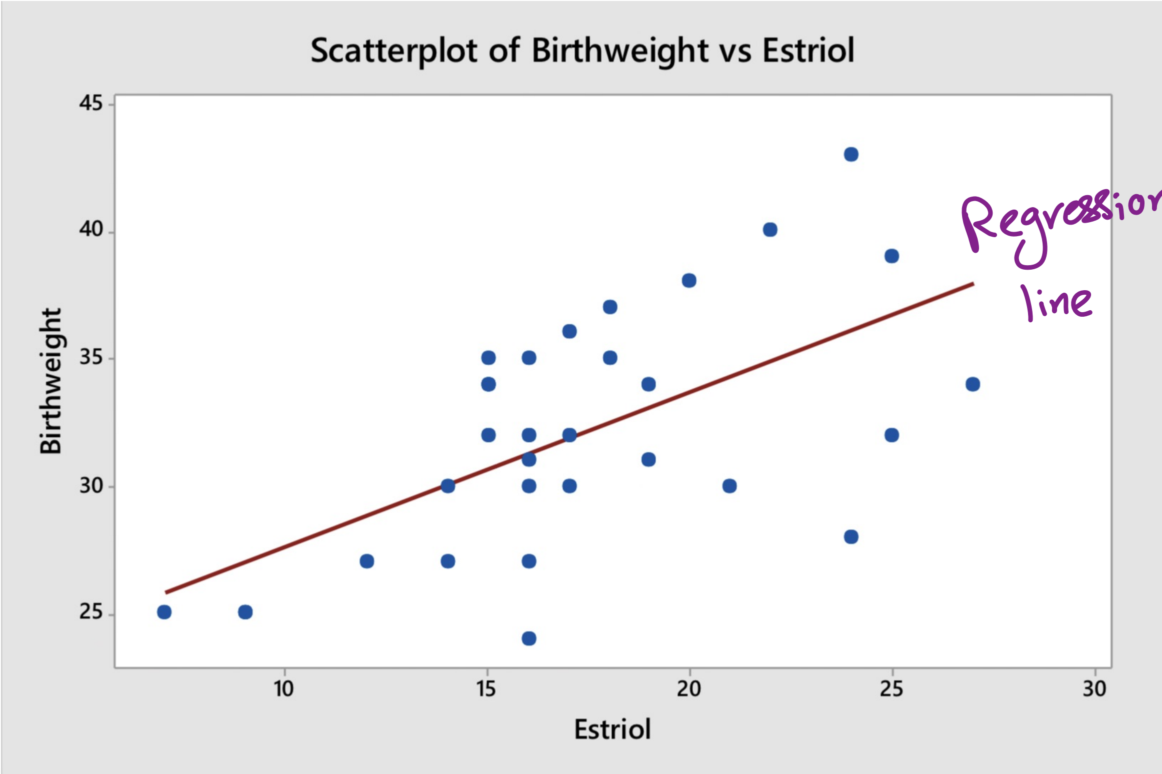
⇒ fetus growing

TABLE 11.1 Sample data from the Greene-Touchstone study relating birthweight and estriol level in pregnant women near term

i	Estriol (mg/24 hr) x_i	Birthweight (g/100) y_i	i	Estriol (mg/24 hr) x_i	Birthweight (g/100) y_i
1	7	25	17	17	32
2	9	25	18	25	32
3	9	25	19	27	34
4	12	27	20	15	34
5	14	27	21	15	34
6	16	27	22	15	35
7	16	24	23	16	35
8	14	30	24	19	34
9	16	30	25	18	35
10	16	31	26	17	36
11	17	30	27	18	37
12	19	31	28	20	38
13	21	30	29	22	40
14	24	28	30	25	39
15	15	32	31	24	43
16	16	32			

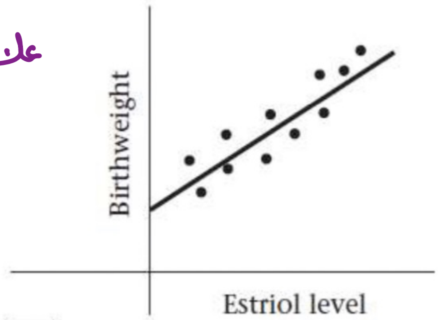
Source: Based on the *American Journal of Obstetrics and Gynecology, 85*(1), 1–9, 1963.





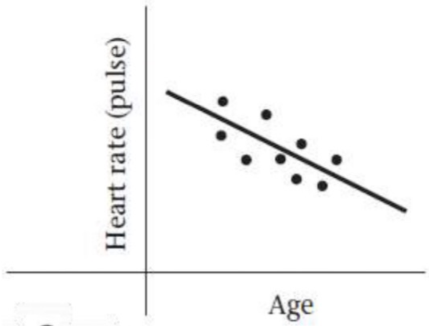
* حالات الـ Scatter plot

علاقة طردية



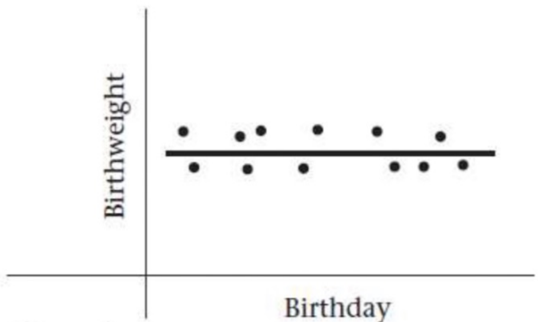
$\rho > 0$

علاقة عكسية



$\rho < 0$

لا يوجد علاقة



$\rho = 0$

NOTE ρ is called the Correlation Coefficient and its value between -1,1

-1 0 1

negative Relationship

No Relationship

positive Relationship

<u>NOTE</u> $\rho = 1$ [perfect positive Relationship]

$\rho = -1$ [perfect negative Relationship]

2 Numerical Method

* Covariance

$$Cov(x, y) = E\left((x-\mu_x)(y-\mu_y)\right)$$

$$= E(xy) - \mu_x \mu_y$$

negative
علاقة عكسية

0
لا يوجد
علاقة

positive
علاقة
طردية

<u>NOTE</u> units limit the use of covariance

$x = KG$ $\quad$ $y = mmHg$

$$E\left((x-\mu_x)(y-\mu_y)\right)$$

* Correlation Coefficient

ρ : population correlation coefficient

$$\rho = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

r : Sample correlation coefficient

" Pearson's correlation coefficient "

$$r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}}$$

$$S^2 = \frac{\sum x^2}{n-1} - \frac{(\sum x)^2}{n(n-1)}$$

$$L_{xy} = \sum xy - \frac{\sum x * \sum y}{n}$$

$$L_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$L_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

NOTES

1. L_{xx} and L_{yy} never ever be negative

2. r will be unchanged by a change in the unit of x, y

Example The Data shown in the table below obtained in a study of age (x) in years and systolic blood pressure (y) in mmHg for Random sample of six patients selected from the emergency of

JUH in a given day:

Age	Systolic Blood pressure
43	128
48	120
56	135
61	143
67	141
70	152

Calculate the value of the Correlation Coefficient for data? and give a Conclusion?

x Age	y SBP	x^2	y^2	xy
43	128	1849	16384	5504
48	120	2304	14400	5760
56	135	3136	18225	7560
61	143	3721	30449	8723
67	141	4489	19881	9447
70	152	4900	23104	10640

Sum $\quad \sum x \quad \sum y \quad \sum x^2 \quad \sum y^2 \quad \sum xy$

$$r = \frac{L_{xy}}{\sqrt{L_{xx} \; L_{yy}}} \qquad (-1 \; , \; 1)$$

$$L_{xy} = \sum xy - \frac{\sum x * \sum y}{n}$$

$$= 47634 - \frac{345 * 819}{6}$$

$$= 541.5$$

$$L_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$= 20399 - \frac{(345)^2}{6} = 561.5$$

$$L_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$= 112443 - \frac{(819)^2}{6} = 649.5$$

$$r = \frac{541.5}{\sqrt{561.5 * 649.5}} = 0.8966$$

So There is a strong correlation between the age and SBP.

Example Calculate the Correlation Coefficient of the given data:

x	12	15	18	21	27
y	2	4	6	8	12

حل

x	12	15	18	21	27
y	2	4	6	8	12
x^2	144	225	324	441	729
y^2	4	16	36	64	144
xy	24	60	108	168	324

$\sum x = 93$, $\sum y = 32$, $\sum xy = 684$

$\sum x^2 = 1863$, $\sum y^2 = 264$

$$r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}} = \frac{88.8}{\sqrt{133.2 * 59.2}} = 1$$

$$L_{xy} = \sum xy - \frac{\sum x * \sum y}{n}$$

$$= 670 - \frac{93 * 32}{5} = 88.8$$

$$L_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$= 1863 - \frac{(93)^2}{5} = 133.2$$

$$L_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$= 264 - \frac{(32)^2}{5} = 59.2$$

∞ Perfect positive Relationship

Example Calculate the Correlation Coefficient of the given data:

x	50	51	52	53	54
y	3.1	3.2	3.3	3.4	3.5

Ans. $r = 1$ [perfect positive Relationship]

NOTE

$$L_{xx} = (n-1) * S_x^2$$

$$S_x^2 = \frac{L_{xx}}{n-1}, \quad S_y^2 = \frac{L_{yy}}{n-1}$$

$$S_{xy} = \frac{L_{xy}}{n-1}$$

(Sample covariance)

$$\boxed{r = \frac{S_{xy}\,\cancel{(n-1)}}{\sqrt{\frac{S_x^2}{\cancel{(n-1)}} * \frac{S_y^2}{\cancel{(n-1)}}}} = \frac{S_{xy}}{S_x * S_y}}$$

---

* Statistical inference for correlation coefficent

Hypothesis testing

Confidence interval

* Hypothesis testing for Correlation coeffecient

$H_0: \rho = 0$

$H_1: \rho \neq 0$

$H_0: \rho = \rho_0$

$H_1: \rho \neq \rho_0$

1 $\quad H_0: \rho = 0 \quad$ Vs. $\quad H_1: \rho \neq 0$

<u>test stat</u>

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

r : sample correlation coefficient

<u>NOTE</u> Always the $\boxed{d.f = n-2}$

كتاب

Example Suppose serum cholestrol levels in spouse pairs are measured to determine whether there is a correlation between cholesterol levels in spouses.

Specifically, we wish to test:

$$H_0: \rho = 0 \quad \text{vs.} \quad H_1: \rho \neq 0$$

suppose that $r = 0.897$ based on $n = 6$ spouse pairs. Is there enough evidence to warrent Rejecting H_0?

(use $\alpha = 0.05$)

الحل $H_0: \rho = 0 \quad \text{vs.} \quad H_1: \rho \neq 0$

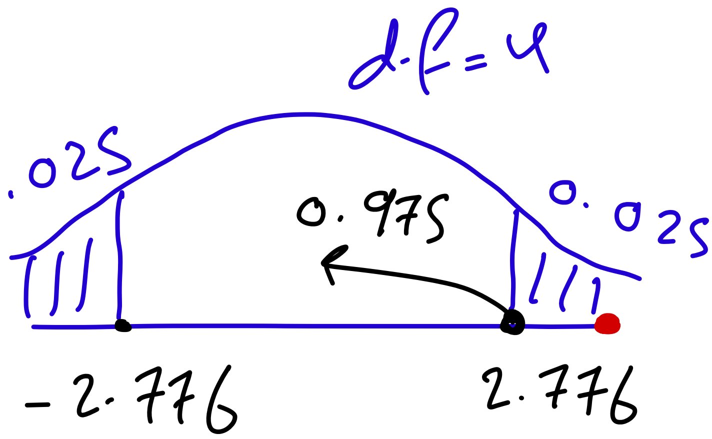
test stat

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$= \frac{0.897 * \sqrt{6-2}}{\sqrt{1-0.897^2}} = 4.056$$

$\alpha = 0.05$

$\frac{\alpha}{2} = 0.025$



∴ we Reject H_0 and accept H_1

"There is a positive Relationship" between cholesterol level spouses

Example Test to see if the correlation for hour studies on the exam and grade on the exam is statistically significant. Use $\alpha = 0.05$, $r = 0.825$, $n = 13$

$H_0 : \rho = 0 \quad$ Vs. $\quad H_1 : \rho \neq 0$

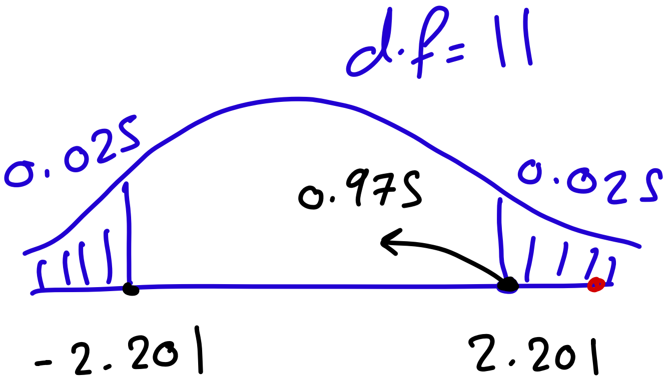
<u>test stat</u>

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.825 * \sqrt{13-2}}{\sqrt{1-0.825^2}}$$

$$= 5.271$$

$\alpha = 0.05$

$\frac{\alpha}{2} = 0.025$



∴ we Reject Ho and accept H_1

" There is a Correlation between"
the hour studies and the grade

2. $H_0 : \rho = \rho_0$ vs. $H_1 : \rho \neq \rho_0$

Test stat

$$\lambda = (Z - Z_0)\sqrt{n-3}$$

$$Z \text{ transformation} = \frac{1}{2} \ln \frac{(1+r)}{(1-r)}$$

$$Z_0 = \frac{1}{2} \ln \frac{(1+\rho)}{(1-\rho)}$$

NOTE "Fisher's Z transformation"

Example Suppose the Body weights of 100 father (x) and first born son (y) are measured and a sample correlation coefficient r of 0.38 is found. we might ask whether or not this sample correlation is compatible with an underlying correlation of 0.5 that might be expected on genetic grounds. Perform a test of significance, use $\alpha = 0.05$

Sol $H_0: \rho = 0.5$ Vs. $H_1: \rho \neq 0.5$

test stat

$$\lambda = (Z - Z_0)\sqrt{n-3}$$

$$Z = \frac{1}{2} \ln \frac{(1+r)}{(1-r)} = \frac{1}{2} * \ln\left(\frac{1+0.38}{1-0.38}\right)$$

$$= 0.4$$

$$Z_0 = \frac{1}{2}\ln\frac{(1+\rho)}{(1-\rho)} = \frac{1}{2}\ln\left(\frac{1+0.5}{1-0.5}\right)$$

$$= 0.549$$

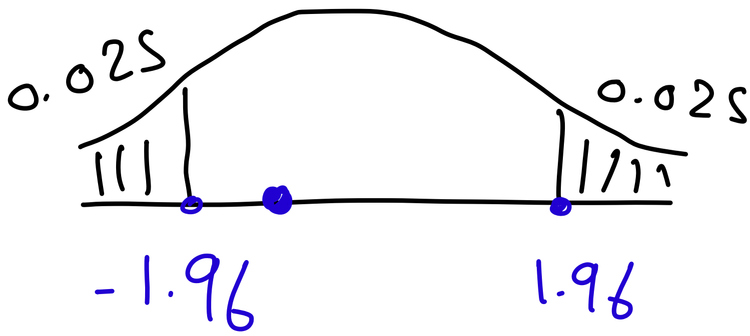
$$\lambda = (Z - Z_0)\sqrt{n-3}$$

$$= (0.4 - 0.549)\sqrt{100-3}$$

$$= -1.47$$

$\alpha = 0.05$

$\frac{\alpha}{2} = 0.025$



∴ we accept H_0 and Reject H_1

$\Rightarrow$ Yes, $r = 0.38$ is compatible with correlation of $\rho = 0.5$

Example vancomycin is an antibiotic used to treat C. difficile bacteria that cause pseudomembranous colitis. A study was done on a sample of 120 patients showed a sample correlation coefficient of 0.775 between the dose of vancomycin and the percentage of bacteria in the colon. test whether it suitable to the underlying correlation of 0.7? (use $\alpha = 0.05$)?

Q31 $H_0: \rho = 0.7 \quad vs. \quad H_1: \rho \neq 0.7$

<u>test stat</u>

$$\lambda = (Z - Z_0)\sqrt{n-3}$$

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

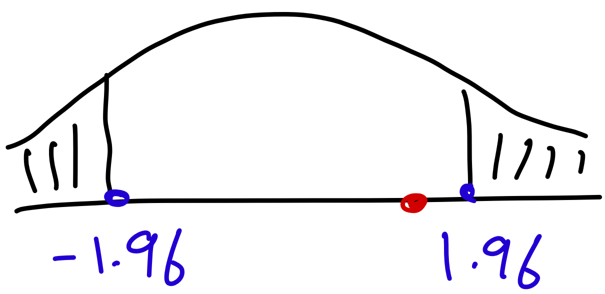
$$= \frac{1}{2} * \ln\left(\frac{1+0.775}{1-0.775}\right) = 1.032$$

$$Z_0 = \frac{1}{2}\ln\left(\frac{1+0.7}{1-0.7}\right) = 0.87$$

$$\lambda = (1.032 - 0.87) * \sqrt{120-3}$$

$$= 1.75$$

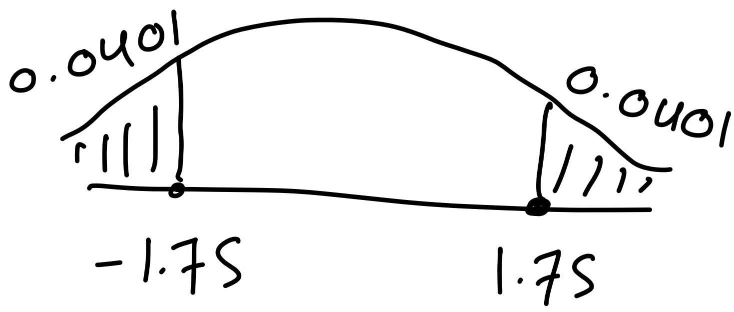
$\alpha = 0.05$



So we accept H_0 and reject H_1

$$P\text{-value} = 2 * 0.0401$$

$$= 0.0802$$



NOTE you can find the Fisher's Z transformation by this table:

TABLE 12 Fisher's z transformation

r	z	r	z	r	z	r	z	r	z
.00	.000								
.01	.010	.21	.213	.41	.436	.61	.709	.81	1.127
.02	.020	.22	.224	.42	.448	.62	.725	.82	1.157
.03	.030	.23	.234	.43	.460	.63	.741	.83	1.188
.04	.040	.24	.245	.44	.472	.64	.758	.84	1.221
.05	.050	.25	.255	.45	.485	.65	.775	.85	1.256
.06	.060	.26	.266	.46	.497	.66	.793	.86	1.293
.07	.070	.27	.277	.47	.510	.67	.811	.87	1.333
.08	.080	.28	.288	.48	.523	.68	.829	.88	1.376
.09	.090	.29	.299	.49	.536	.69	.848	.89	1.422
.10	.100	.30	.310	.50	.549	.70	.867	.90	1.472
.11	.110	.31	.321	.51	.563	.71	.887	.91	1.528
.12	.121	.32	.332	.52	.576	.72	.908	.92	1.589
.13	.131	.33	.343	.53	.590	.73	.929	.93	1.658
.14	.141	.34	.354	.54	.604	.74	.950	.94	1.738
.15	.151	.35	.365	.55	.618	.75	.973	.95	1.832
.16	.161	.36	.377	.56	.633	.76	.996	.96	1.946
.17	.172	.37	.388	.57	.648	.77	1.020	.97	2.092
.18	.182	.38	.400	.58	.662	.78	1.045	.98	2.298
.19	.192	.39	.412	.59	.678	.79	1.071	.99	2.647
.20	.203	.40	.424	.60	.693	.80	1.099		

* Confidence interval for correlation

Fisher's Z Transformation (Z_1, Z_2)

ρ (ρ_1, ρ_2)

NOTES

① $(1-\alpha) = CI$

② for Fisher's Z transformation (Z_1, Z_2)

$$Z_1 = Z - \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n-3}}$$

$$Z_2 = Z + \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n-3}}$$

$$\boxed{Z \pm \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n-3}}}$$

$$Z = \frac{1}{2} \ln \frac{(1+r)}{(1-r)}$$

③ for population correlation coefficient (ρ)

$$(\rho_1 , \rho_2)$$

$$\rho_1 = \frac{e^{2Z_1} - 1}{e^{2Z_1} + 1}$$

$$\rho_2 = \frac{e^{2Z_2} - 1}{e^{2Z_2} + 1}$$

كتاب

Example Suppose the Body weights of 100 father (x) and first born son (y) have a sample correlation coefficient of $r = 0.38$, find 0.95 confidence interval for the underlying correlation?

الحل

$r = 0.38$

$CI = 0.95$

$n = 100$

$$Z \pm \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n-3}}$$

$$Z = \frac{1}{2} \ln \frac{(1+r)}{(1-r)}$$

$$= \frac{1}{2} * \ln\left(\frac{1+0.38}{1-0.38}\right)$$

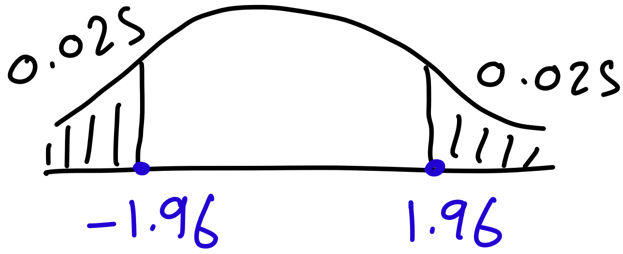
$$= 0.4$$

$$1 - \alpha = 0.95$$

$$\alpha = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

$$Z_{\frac{\alpha}{2}} = \pm 1.96$$



$$\left(z - \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n-3}} \, , \; z + \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n-3}} \right)$$

$$\left(0.4 - \frac{1.96}{\sqrt{97}} \, , \; 0.4 + \frac{1.96}{\sqrt{97}} \right)$$

$$\left(\underset{Z_1}{0.201} \, , \; \underset{Z_2}{0.599} \right)$$

for ρ $\quad (\rho_1 \, , \; \rho_2)$

$$\rho_1 = \frac{e^{2z_1} - 1}{e^{2z_1} + 1}$$

$$= \frac{e^{2*0.201} - 1}{e^{2*0.201} + 1} = 0.198$$

$$\rho_2 = \frac{e^{2z_2} - 1}{e^{2z_2} + 1}$$

$$= \frac{e^{2*0.599} - 1}{e^{2*0.599} + 1} = 0.536$$

$$(0.198,\ 0.536)$$

Example: Suppose we want to estimate the correlation coefficient between height and weight of residents in a certain country. We select a random sample of 60 residents and find the following information:

- Sample size $n = 60$
- Sample correlation coefficient $r = 0.56$

Find a 95% confidence interval for the correlation?

$r = 0.56$

$n = 60$

$$z \pm \frac{z_{\frac{\alpha}{2}}}{\sqrt{n-3}}$$

$CI = 0.95 \mid z = \frac{1}{2} \ln \frac{1+r}{1-r}$

$$= \frac{1}{2} * \ln \frac{1+0.56}{1-0.56}$$

$$= 0.633$$

$$1 - \alpha = 0.95$$

$$\alpha = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

$$\boxed{z_{\frac{\alpha}{2}} = \pm 1.96}$$

$$\left(z - \frac{z_{\frac{\alpha}{2}}}{\sqrt{n-3}} , z + \frac{z_{\frac{\alpha}{2}}}{\sqrt{n-3}} \right)$$

$$\left(0.633 - \frac{1.96}{\sqrt{57}} , 0.633 + \frac{1.96}{\sqrt{57}} \right)$$

$$(0.373,\ 0.892)$$

$$\quad z_1 \qquad\qquad z_2$$

$\Rightarrow$ for ρ $\quad (\rho_1, \rho_2)$

$$\rho_1 = \frac{e^{2z_1} - 1}{e^{2z_1} + 1}$$

$$= \frac{e^{2*0.373} - 1}{e^{2*0.373} + 1} = 0.3568$$

$$\rho_2 = \frac{e^{2*z_2} - 1}{e^{2z_2} + 1} = 0.7126$$

$$(0.3568,\ 0.7126)$$

Summary for Ch. 11

"Statistical inference"

① Hypothesis testing

Ⓐ $H_0 : \rho = 0$

vs.

$H_1 : \rho \neq 0$

$$\boxed{\underline{\text{Test stat}} \quad t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}}$$

Ⓑ $H_0 : \rho = \rho_0$

Vs.

$H_1 : \rho \neq \rho_0$

$$\underline{\text{Test stat}} \quad \lambda = (z - z_0) * \sqrt{n-3}$$

② Confidence interval

Ⓐ for Fisher's z transformation

$$\boxed{z \pm \frac{z_{\frac{\alpha}{2}}}{\sqrt{n-3}}}$$

Ⓑ for ρ

$$\left(\frac{e^{2z_1} - 1}{e^{2z_1} + 1} , \frac{e^{2z_2} - 1}{e^{2z_2} + 1} \right)$$

Chapter 12

Multi Sample inference

⇒ $\mu_1, \mu_2, \mu_3, \mu_4 \ldots\ldots \mu_k$

⇒ we will study the inference between 3 or more means

ANOVA

تحليل تشتت
" Analysis of variance ⇐

⇒ Extension of T-test

⇒ Testing of more than 2 sample means.

Example

A	B	C
29	28	25
30	29	28
31	27	29
31	30	27
29	29	29

variance within group

variance between groups

NOTE

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

vs.

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$$

at least one pairs are not equal

$$F = \frac{\text{variance between}}{\text{variance within}}$$

> 1 Reject H_0

≤ 1 accept H_1

* One way ANOVA fixed effects model:

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

y_{ij} : i : group

j : number of observation

* $y_{2,3}$: المشاهدة رقم 3 في المجموعة 2

μ : overall mean

e_{ij} : error about mean

$\alpha_i = \bar{y}_i - \mu$

---

* Hypothesis testing of Multisample using one way ANOVA modal:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$

Vs

$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \cdots \mu_k$

NOTE we accept Ho only if all

the μ are equal, if one of the μ differ, we will accept H_1

طريقة أخرى

H_0 : all $\alpha_i = 0$ Vs. $H_1 : \alpha_i \neq 0$ at least one

test stat

$$F = \frac{MS_B}{MS_w}$$

MS_B : Mean Square between

$$= \frac{SS_B}{K-1}$$

K : groups

$$SS_B = \sum n_i (\bar{y}_i)^2 - \frac{(\sum n_i \bar{y}_i)^2}{N}$$

$$MS_w = \frac{SS_w}{N-k}$$

N : total sample size

$$SS_w = \sum (n_i - 1) S_i^2$$

NOTE Degree of freedom

$\Rightarrow$ for $MS_B = k - 1$ (بسط)

$\Rightarrow$ for $MS_w = N - k$ (مقام)

Example

	1	2	3
	1	2	2
	2	4	3
	5	2	4
$\bar{y}$	2.67	2.67	3
S^2	4.33	1.33	1

Test whether the mean differ significantly among 3 groups? (use $\alpha = 0.05$)

$H_0 : \mu_1 = \mu_2 = \mu_3$ $\qquad$ $H_0 : \alpha_i = 0$ (All)

vs. $\qquad$ vs.

$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$ $\qquad$ $H_1 : \alpha_i \neq 0$ (At least one)

<u>test stat</u>

$$F = \frac{MS_B}{MS_w}$$

$$\Rightarrow MSB = \frac{SS_B}{k-1} = \frac{0.22}{3-1} = 0.11$$

$$SS_B = \sum n_i (\bar{y}_i)^2 - \frac{\left(\sum n_i \bar{y}_i\right)^2}{N}$$

$$= 69.7734 - \frac{(25.02)^2}{9} = 0.22$$

$$\underline{\sum n_i(\bar{y}_i)^2}$$

$$3*(2.67)^2 + 3*(2.67)^2 + 3*(3)^2$$

$$= 69.7734$$

$$\underline{\sum n_i \bar{y}_i}$$

$$3*2.67 + 3*2.67 + 3*3$$

$$= 25.02$$

$$\Rightarrow MS_w = \frac{SS_w}{N-K} = \frac{13.32}{9-3} = 2.22$$

$$SS_w = \sum (n_i - 1) S_i^2$$

HINT

$$S^2 = \frac{\sum x^2}{n-1} - \frac{(\sum x)^2}{n(n-1)}$$

$= (3-1)\,4.33 + (3-1) * 1.33 + (3-1) * 1$

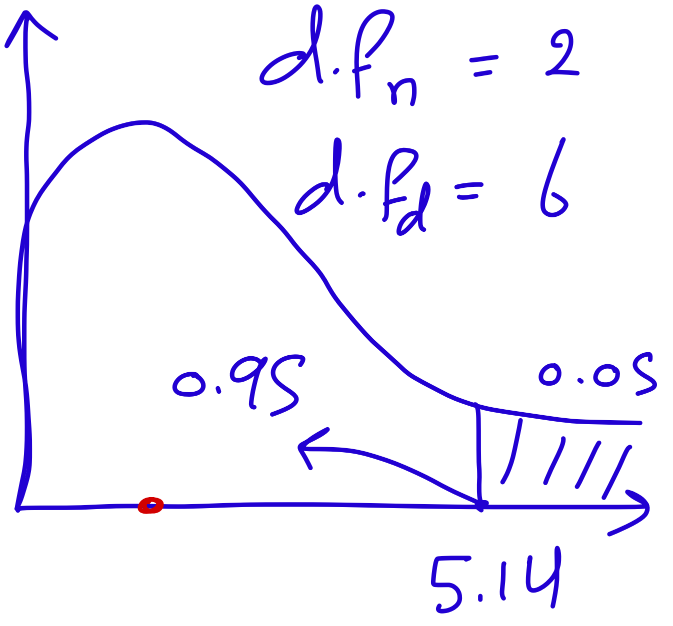
$= 13.32$

$$F = \frac{MS_B}{MS_w}$$

$$= \frac{0.11}{2.22} = 0.0495$$

$D.f_n = k - 1$

$D.f_d = N - K$



$\therefore$ we accept H_0 and Reject H_1

Example Suppose we want to know whether or not three different exam prep programs lead to different mean scores on a certain exam. To test this, we recruite 30 students to participate in a study and split them into three groups, shown with students marks after 3 weeks of prep:

Group 1	Group 2	Group 3
85	91	79
86	92	78
88	93	88
75	85	94
78	87	92
94	84	85
98	82	83
79	88	85
71	95	82
80	96	81

Source	SS	df	MS	F	P
between	192.2	2	96.1	2.358	0.11385
Within	1100.6	27	40.8		
Total	1292.8	29			

$H_0: \alpha_i = 0$ Vs. $H_1: \alpha_i \neq 0$

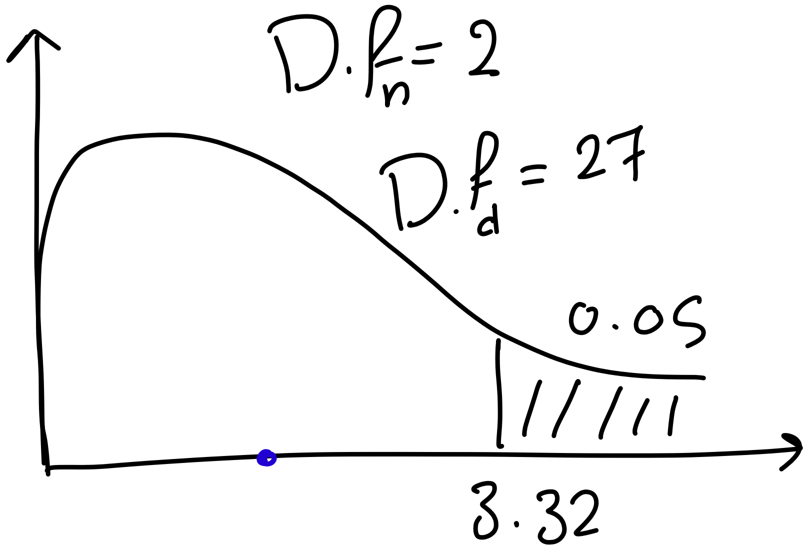
$H_0: \mu_1 = \mu_2 = \mu_3$ Vs. $H_1: \mu_1 \neq \mu_2 \neq \mu_3$

<u>test stat</u>

$$F = \frac{MS_B}{MS_w} = \frac{96.1}{40.8} = 2.35$$

$D.f_n = K - 1$

$D.f_d = N - K$



∴ we accept H_0 and Reject H_1

⇒ fail to reject, There is insufficient evidence to say that there is a statitically significant difference between the mean exam scores of three groups

---

Example The times required by three Surgeons to perform appendectomy were Recorded on five randomly selected occasions, Here are the times, to the nearest minute.

Maher	Haider	Tareq
8	8	10
10	9	9
9	9	10
11	8	11
10	10	9

Test if the mean time Recorded for each Surgery is different between Surgeons

Source		df	SS	MS = SS/df	F-statistic	p-value
Treatments	B	2	2.8	1.4	1.5556	p-value > 0.10
Error	w	12	10.8	0.9		
Total		14	13.6			

$H_0: \overset{\text{All}}{\alpha_i = 0}$ Vs. $H_1: \alpha_i \neq 0$ (At least one)

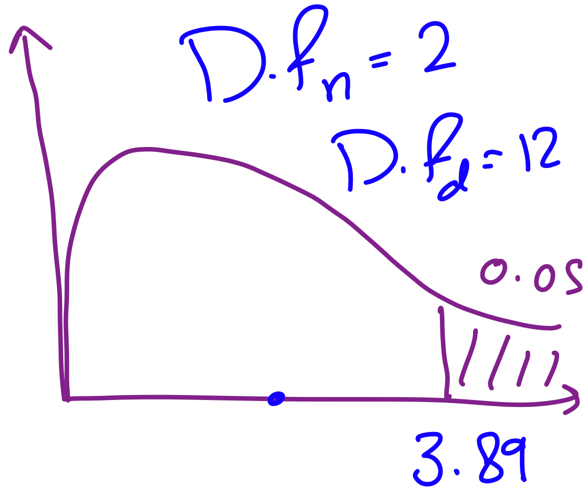
$H_0: \mu_1 = \mu_2 = \mu_3$ Vs. $H_1: \mu_1 \neq \mu_2 \neq \mu_3$

test stat

$$F = \frac{MS_B}{MS_w} = \frac{1.4}{0.9} = 1.55$$

$D.f_n = K - 1$

$D.f_d = N - K$



∴ we accept H_0 and reject H_1

Example Fill in the missing entries of the partially completed one-way ANOVA

	Source	df	SS	MS = SS/df	F-statistic
B	Treatments	3	2.124	0.708	0.75
w	Error	20	18.880	0.944	
	Total	23	21.004		

$$MS_B = \frac{SS_B}{d.f}$$

$$0.708 = \frac{2.124}{d.f}$$

$$d.f = \frac{2.124}{0.708} = 3$$

$$\Rightarrow F = \frac{MS_B}{MS_w}$$

$$0.75 = \frac{0.708}{MS_w} \Rightarrow MS_w = 0.944$$

$$\Rightarrow MS_w = \frac{SS_w}{N-k}$$

$$0.944 = \frac{SS_w}{20} = SS_w = 18.880$$

مثال

Example Test whether the mean FEF scores differ significantly among the six groups in the following table (use $\alpha = 0.05$)

TABLE 12.1 FEF data for smoking and nonsmoking males

Group number, *i*	Group name	Mean FEF (L/s)	*sd* FEF (L/s)	n_i
1	NS	3.78	0.79	200
2	PS	3.30	0.77	200
3	NI	3.32	0.86	50
4	LS	3.23	0.78	200
5	MS	2.73	0.81	200
6	HS	2.59	0.82	200

Source: Based on *The New England Journal of Medicine*, *302*(13), 720–723, 1980.

$H_0: \mu_1 = \mu_2 = \dots = \mu_6$

Vs.

$H_1: \mu_1 \neq \mu_2 \neq \mu_3 = - - - - \neq \mu_6$

test stat

$$F = \frac{MS_B}{MS_w}$$

$$\Rightarrow MS_B = \frac{SS_B}{k-1} = \frac{184.38}{5} = 36.875$$

$$SS_B = \sum n_i (\bar{y}_i)^2 - \frac{\left(\sum n_i \bar{y}_i\right)^2}{N}$$

$\sum n_i (\bar{y}_i)^2$

$200 * (3.78)^2 + 200 * (3.30)^2 + 50 * (3.32)^2$
$+ 200 * (3.23)^2 + 200 * (2.73)^2 + 200 * (2.59)^2$

$= 10505.58$

$$\sum n_i \bar{y}_i$$

$200 * 3.78 + 200 * 3.30 + 50 * 3.32$

$+ 200 * 3.23 + 200 * 2.73 + 200 * 2.59$

$= 3292$

$$SS_B = 10505.85 - \frac{(3292)^2}{1050}$$

$= 184.38$

$$\Rightarrow MS_w = \frac{SS_w}{N-K} = \frac{663.87}{1044} = 0.636$$

$$SS_w = \sum (n_i - 1) s_i^2$$

$$(200-1) * 0.79^2 + 199 * 0.77^2 + 49 * 0.86^2$$

$$+ 199 * 0.78^2 + 199 * 0.81^2 + 199 * 0.82^2$$

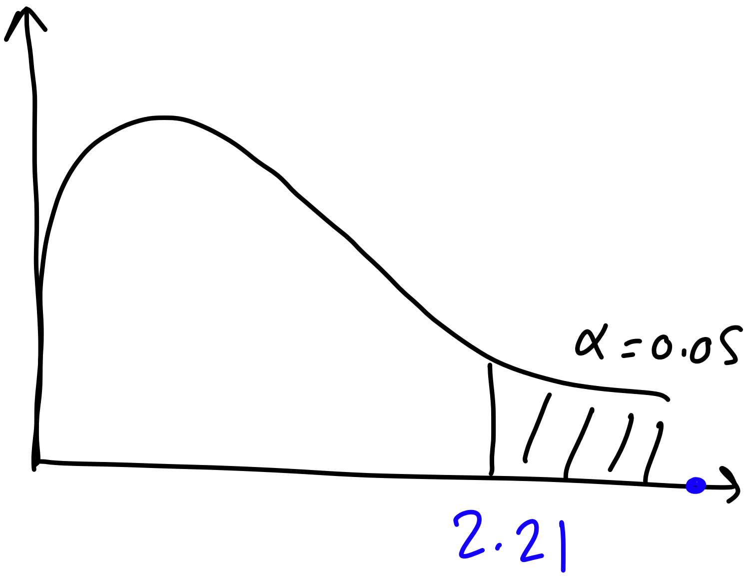
$$= 663.87$$

$$F = \frac{MS_B}{MS_w} = \frac{36.875}{0.636}$$

$$= 58$$

$D.f_n = 5$

$D.f_d = 1044$



∴ we Reject H_0 and accept H_1

TABLE 12.3 ANOVA table for FEF data in Table 12.1

	SS	*df*	MS	*F* statistic	*p*-value
Between	184.38	5	36.875	58.0	$p < .001$
Within	663.87	1044	0.636		
Total	848.25				

General NOTES About ANOVA:

① $y_{ij} - \bar{\bar{y}} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{\bar{y}})$

$$\Downarrow$$

$$\underbrace{\sum (y_{ij} - \bar{\bar{y}})^2}_{SS_T} = \underbrace{\sum (y_{ij} - \bar{y}_i)^2}_{SS_W} + \underbrace{\sum (\bar{y}_i - \bar{\bar{y}})^2}_{SS_B}$$

② $SS_T = SS_W + SS_B$

* Least significant difference Test (LSD)

$\Rightarrow$ used to see which means are not significantly equal the Rest of the means.

$\Rightarrow$ ~~الشرط~~ $\rightarrow$ Reject H_0 and accept H_1

$$H_0: \mu_1 = \mu_2 \quad Vs. \quad H_1: \mu_1 \neq \mu_2$$

<u>Test stat</u>

$$T = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sqrt{sp^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

$$\boxed{sp^2 = MS_w}$$

$$\boxed{d.f = N - K}$$

Example: Smoking in Book:

$$H_0: \mu_1 = \mu_2 \quad \text{vs.} \quad H_1: \mu_1 \neq \mu_2$$

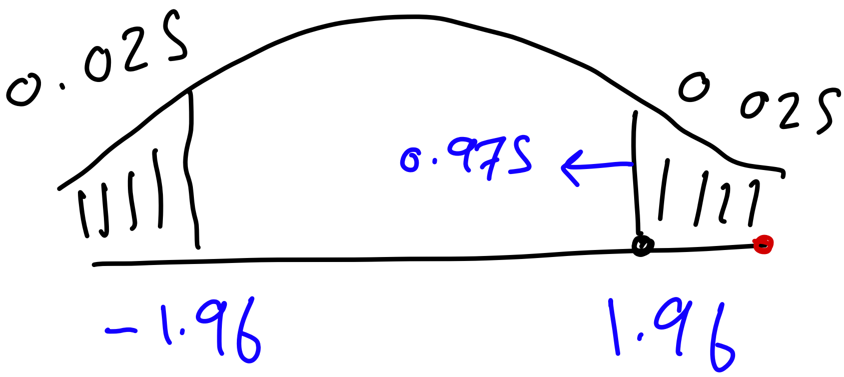
Test stat

$$T = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sqrt{s_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} = \frac{(3.78 - 3.30)}{\sqrt{0.636\left(\frac{1}{200} + \frac{1}{200}\right)}}$$

$$= 6.02$$

$$\boxed{s_p^2 = MS_w}$$

d.f = 1044



$\therefore$ we Reject H_0 and accept H_1

So, $\mu_1 \neq \mu_2$

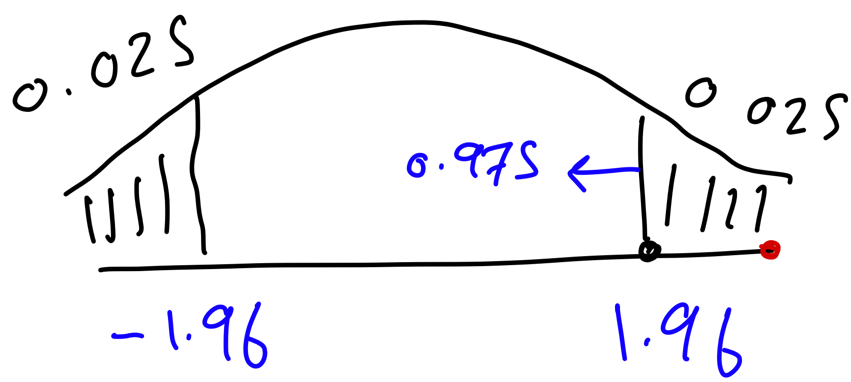
* $H_0: \mu_1 = \mu_3$ vs. $H_1: \mu_1 \neq \mu_3$

Test stat

$$T = \frac{(\bar{y}_1 - \bar{y}_3) - 0}{\sqrt{s_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

$$= \frac{(3.78 - 3.32)}{\sqrt{0.636 * \left(\frac{1}{200} + \frac{1}{50}\right)}} = 3.65$$

d.f. = 1044



$\therefore$ we Reject H_0 and accept H_1

So, $\mu_1 \neq \mu_3$

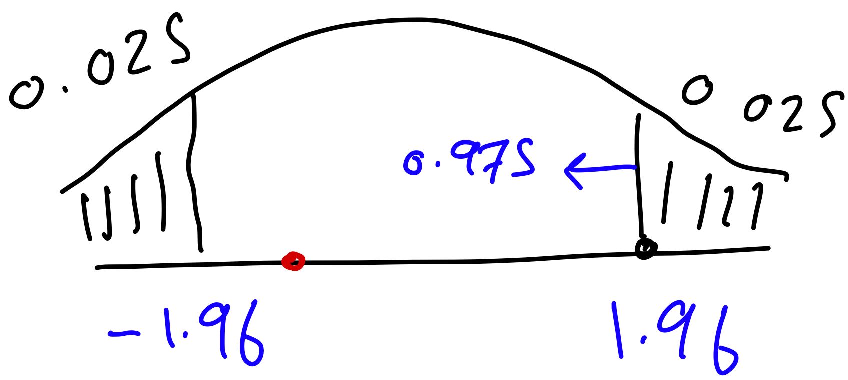
* $H_0 : \mu_2 = \mu_3$ vs. $H_1 : \mu_2 \neq \mu_3$

Test stat

$$T = \frac{(\bar{y}_2 - \bar{y}_3)}{\sqrt{s_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} = \frac{(3.3 - 3.32)}{\sqrt{0.636\left(\frac{1}{200} + \frac{1}{200}\right)}}$$

$$= -0.16$$

d.f = 1044



∴ we accept H_0 and Reject H_1

So, $\mu_2 = \mu_3$

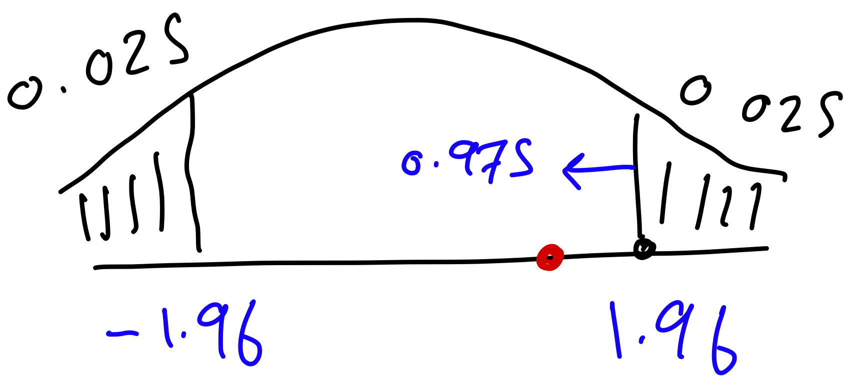
* $H_0: \mu_2 = \mu_4$ vs. $H_1: \mu_2 \neq \mu_4$

<u>Test stat</u>

$$T = \frac{(\bar{y}_2 - \bar{y}_4)}{\sqrt{sp^2\left(\frac{1}{n}+\frac{1}{m}\right)}} = \frac{(3.3 - 3.23)}{\sqrt{0.636\left(\frac{1}{200}+\frac{1}{200}\right)}}$$

$$= 0.88$$

d.f = 1044



So we accept H_0 and Reject H_1

So, $\mu_2 = \mu_4$

TABLE 12.4 Comparisons of specific pairs of groups for the FEF data in Table 12.1 (on page 552) using the LSD t test approach

$\alpha=0.05$

Groups compared	Test statistic	p-value
NS, PS	$t=\dfrac{3.78-3.30}{\sqrt{0.636\left(\dfrac{1}{200}+\dfrac{1}{200}\right)}}=\dfrac{0.48}{0.08}=6.02^{a}$	$<.001$
NS, NI	$t=\dfrac{3.78-3.32}{\sqrt{0.636\left(\dfrac{1}{200}+\dfrac{1}{50}\right)}}=\dfrac{0.46}{0.126}=3.65$	$<.001$
NS, LS	$t=\dfrac{3.78-3.23}{\sqrt{0.636\left(\dfrac{1}{200}+\dfrac{1}{200}\right)}}=\dfrac{0.55}{0.08}=6.90$	$<.001$
NS, MS	$t=\dfrac{3.78-2.73}{0.080}=\dfrac{1.05}{0.08}=13.17$	$<.001$
NS, HS	$t=\dfrac{3.78-2.59}{0.080}=\dfrac{1.19}{0.08}=14.92$	$<.001$
PS, NI	$t=\dfrac{3.30-3.32}{0.126}=\dfrac{-0.02}{0.126}=-0.16$	0.87
PS, LS	$t=\dfrac{3.30-3.23}{0.080}=\dfrac{0.07}{0.08}=0.88$	0.38
PS, MS	$t=\dfrac{3.30-2.73}{0.080}=\dfrac{0.57}{0.08}=7.15$	$<.001$
PS, HS	$t=\dfrac{3.30-2.59}{0.080}=\dfrac{0.71}{0.08}=8.90$	$<.001$
NI, LS	$t=\dfrac{3.32-3.23}{0.126}=\dfrac{0.09}{0.126}=0.71$	0.48
NI, MS	$t=\dfrac{3.32-2.73}{0.126}=\dfrac{0.59}{0.126}=4.68$	$<.001$
NI, HS	$t=\dfrac{3.32-2.59}{0.126}=\dfrac{0.73}{0.126}=5.79$	$<.001$
LS, MS	$t=\dfrac{3.23-2.73}{0.08}=\dfrac{0.50}{0.08}=6.27$	$<.001$
LS, HS	$t=\dfrac{3.23-2.59}{0.08}=\dfrac{0.64}{0.08}=8.03$	$<.001$
MS, HS	$t=\dfrac{2.73-2.59}{0.08}=\dfrac{0.14}{0.08}=1.76$	0.08

[a]All test statistics follow a t_{1044} distribution under H_0.

الأحمر ⇐ متساويين

NOTES

$\Rightarrow$ P-value $> \alpha$ (accept H_0)

P-value $\leq \alpha$ (Reject H_0)

$\Rightarrow MS_w = MS_E$