情報学基礎B

統計 4

Hypothesis Testing, χ^2 test

Correlation, Regression

Hypothesis and its testing (仮説とそれの検定)

Hypothesis is a belief of characteristics of a population. **Hypothesis testing** is to test the belief/claim from sample evidence. **Null Hypothesis**, denoted as H_0 (H-naught), is the statement assumed to be true until evidence (from samples) proves otherwise. **Alternate Hypothesis**, denoted as H_1 (H-one) is a claim to be tested. We will try to find evidence for the alternative hypothesis.
Let us give some examples to explain the notion of Hypothesis testing.

Example 1: Suppose a light bulb company claims that their products' mean lifetime is 500hours ($\mu = 500$hrs). The consumer forum claims that it is less than 500 hours ($\mu < 500$hrs.). In this case, $H_0 : \mu = 500$ is the null hypothesis, and $H_1 : \mu < 500$ is the alternate hypothesis.

We shall elaborate this example with some real numbers. Let the standard deviation of the bulb lifetime is known to be $\sigma = 42$ hours. The consumer forum collected samples of size 49 (i.e., the lifetime of 49 bulbs), and the average lifetime was found to be 482 hours. The question is, if the true average is 500 hrs., what is the probability that the average of 49 samples will be 482 hrs. or less? We know that the standard deviation of the averages of 49 samples would be $\sigma/\sqrt{n}$, i.e., $42/7 = 6$. We calculate the z value corresponding to 482 for the distribution of the average of 49 samples. Here, $z = (482 - 500)/6 = -3$. Now, from "Standard Normal Distribution" table we know that the probability is 0.0013, i.e., 0.13%. This probability is rather very low, and we can doubt, in fact, reject the hypothesis (claim) by the bulb manufacturing company. If the average of the 49 samples be 494 hours, the probability that the average could be 494 or less would be calculated as ($z = 494 - 500/6 = 1$ and therefore probability 15.87%) quite high. In that case, we could not reject the null hypothesis.
This probability, when to accept or reject the Null hypothesis, is set aribitrarily depending on the application. We will discuss this issue later.

Example 2: A Japanese airlines claims that the flight between Sendai to Fukuoka takes an average of 100 mins. A person, who flies this route often, have collected the following 9 sample journey times (in mins):
117, 95, 109, 103, 111, 91, 100, 99, 106
He claims that this journey takes, on an average, longer than 100 mins. In this example, therefore, $H_0 : \mu = 100$, and $H_1 : \mu > 100$.
Does the 9 trip times follow normal distribution? If yes, is the sample mean *significantly* different from the null hypothesis? By significantly different we mean that, assuming the null hypothesis to be correct, the probability of this sample mean is less than 5%.
At least, how long should be the average flight time of 9 flights be, so that we can reject the null hypothesis. As before, we will reject the null hypothesis if its probability is less than 5%.

Example 3: An 煎餅 manufacturer claims that the standard deviation of a 350gms. packet of 煎餅 is 6 gms. Now a group of students claims that the average weight of a packet is less. In other words, $H_0 : \mu = 350$, and $H_1 : \mu < 350$. They bought 40 such packets and found that the

average to be 348 gms. Can the Null hypotheses $H_1 : \mu < 500$ be rejected on the basis of this sample data?
The above problem when viewed from the manufacturer's side is a different story. The 煎餅 manufacturer wants to ascertain that the weight of a packet should lie within ±3gms. from the claimed weight of 350 gms. If it varies too much, there could be claims from the customer. The 煎餅 manufacturer needs to verify whether this assumption (Null Hypothesis) is true or not. In this case, the alternate hypothesis is $H_1 : \mu \neq 350$. Can we reject the null hypothesis when the average weight of 40 packets is 348 gms.? As before, we may reject the null hypothesis if the probability of the sample mean is less than 5%.

Null hypothesis is always an equality expression. There are three ways to set up alternate hypothesis:

1. **two-tailed test:** when H_0: parameter = some value
   H_1: parameter ≠ some value
2. **left-tailed test:** when H_0: parameter = some value
   H_1: parameter < some value
3. **right-tailed test:** when H_0: parameter = some value
   H_1: parameter > some value

The above example 1 is a **left-tailed** one, as the alternate hypothesis claims that, $H_1 : \mu < 500$. Example 2 is a **right-tailed** one, as the passenger's claim is $H_1 : \mu > 100$. The last part of Example 3, where the owner of the factory suspects the precision of the packaging machine, the alternate hypothesis is $H_1 : \mu \neq 350$. This is a case of **two-tailed** problem.

Type I and Type II Errors as outcome of Hypothesis Testing
Depending on what H_0 and H_1 claim, and how the collected samples support them, we have four outcomes out of which two are erroneous as listed below.

1. We reject H_0 when in fact H_1 is true. The decision is correct.
2. We do not reject H_0, when in fact H_0 is true. The decision is correct.
3. We reject H_0 when in fact H_0 is true. This is a **Type I error**.
4. We do not reject H_0 when in fact H_0 is not true. This is a **Type II error**.

The **level of significance** is the probability of making an error (usually Type I error is referred here). It is an user defined quantity, and its value is decided depending on the severity of consequence (say, regarding cost) of making an error. We use the symbol α to denote *the probability of making a Type-I error* (i.e., Rejecting H_0 when H_0 is true). The symbol β is used to denote *the probability of making a Type-II error* (i.e., Not rejecting H_0 when H_1 is true). If the consequence of making a Type-I error is severe, the value of α is set low, say 0.01. If the consequence is not severe, it is set to higher values like 0.10. By making α small, we increase the probability of β high. In general, we deal with α only. The researcher, whose duty is to find which hypothesis is true, set the value of α before collecting data.
For example, let us consider the last part of example 3 above. This is a *two-tailed* case. The management is suspicious about the precision of the old packaging machine, which is still in working condition. Suppose that a new packaging machine is very very costly. Thus, making a

Type I error is costly, because then it would lead to unnecessary purchase of a costly machine. Consequently, α is to be set to a low value. *On the other hand*, suppose that the machine is not costly. But, the legal consequences of different packet weight, that could be raised by the consumer forum, is severe. Then, the management would surely give more importance to *Type II* error. In that case, α is set to higher value to reduce β, the probability of *Type II* error. In conclusion, whether it is*Type I error* or *Type II error*, the level of significance is controlled by α. When the cost of *Type I error* is severe, $\alpha$is made small. When the cost of *Type II error* is severe, $\alpha$is made large

We elaborate the meaning of the term *level of significance* by setting it to different values while solving the above examples 1 to 3. Below, I will just write down the steps to solve these problems for your ready reference:

1. First ascertain the null hypothesis H_0 and the alternate hypothesis H_1.
2. Next check whether it is a *right-tailed*, *left-tailed* or a *two-tailed* problem.
3. Next step is to fix the level of significance i.e., the value of α. When making *Type I error* (i.e. *Rejecting H_0 when H_0 is true*) is costly, α is set to a small value (say 1%). When the cost of *Type II error* is severe, α is made large (say 10%). If nothing is mentioned or known about the severity of making error, we assume a value $\alpha = 0.05$ (i.e., 5%).
4. In most of the problems, the value of standard deviation σ is given. The information about collected samples could come in two different forms. (a) If it is mentioned that the collected samples follow normal distribution and the average is given, well and good. (b) Otherwise, a list of raw sample values are given and nothing is mentioned about the distribution. Then we need to perform the following:
   (i) Ensure that the samples follow normal distribution by Normal-probability plot (see Text 2).
   (ii) Find $\bar{x}$, the average of the given n samples.
5. Find
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$
   For *left-tailed* case, z would be negative, *right-tailed* case z would be positive, and for *both-tailed* case it could be negative or positive.
6. From *Standard Normal Distribution* table, find the probability P corresponding to that z value. The following situations may arise:
   (i) For *left-tailed* problem,
   If $P \leq \alpha$, i.e., the mean is too low from what is claimed in Null Hypothesis, we reject H_0;
   else if, $P > \alpha$, we will not reject H_0 (such low values is probable).
   (ii) For *right-tailed* problem,
   If $(1 - P) \leq \alpha$, i.e., the mean is too high from what is claimed in Null Hypothesis, we reject H_0;
   else if, $(1 - P) > \alpha$, we will not reject H_0 (such high value is probable).
   (iii) For *both-tailed* problem, when $\bar{x} < \mu$ (as μ defined in $H_0 f$), it is treated similar to *left-tailed* case as in item (i), except that α is replaced by $\alpha/2$.

Similarly, when $\bar{x} > \mu$ (as μ defined in H_0, it is treated similar to *right-tailed* case as in item (2), except that α is replaced by $\alpha/2$.

Example:

Price of Gasoline: In Iwate Ken, it is claimed that the average gasoline price is 156 Yen. When it is sampled at 20 different gasoline stands, the prices were as follows:

152, 161, 155, 158, 166, 151, 158, 155, 158, 161,

158, 159, 157, 153, 159, 156, 161, 157, 165, 153

(i) Find whether these 20 samples form normal distribution or not.

(ii) Assuming $\sigma = 0.05$, test the hypothesis at $\alpha = 0.1$ level of significance.

Age of Bride: In 1985, the average age of marriage for a woman was 25 years. It is claimed that the average age has increased in recent years. What is the Null hypothesis and alternate hypothesis? When the data collected from the city office of 20 recently performed marriages, brides' ages were as follows:

40, 23, 30, 24, 31, 29, 28, 24, 35, 34,

24, 21, 46, 29, 31, 29, 29, 21, 33, 39

(i) Find whether these 20 samples form normal distribution or not.

(ii) Assuming $\sigma = 6.2$, test the hypothesis at $\alpha = 0.05$ level of significance.

Hypothesis tests using the χ^2 distribution(カイ二乗分布を使った仮定検定)

Suppose the claim is that a random number generator generates digits from 0 to 9 randomly. This is the Null-hypothesis and we are going to prove it. A set of digits is a sequence of random digits (乱数) if every position in the sequence is equally likely to be occupied by any one of the digits being used, and positions are filled independently.

Example:
A run of 1000 digits had the following frequencies of $0, 1, 2, \ldots 9$. Are they random?
(0〜9 までの数字を 1000 個取り出した。これはランダムですか?)

表 1: Frequency of different digits

Digit, O_r	0	1	2	3	4	5	6	7	8	9	Total
Frequency f_r	106	88	97	101	92	103	96	112	114	91	$N = 1000$

If these digits are a sequence of random digits, the null hypothesis is that $Pr(r) = \frac{1}{10}$ for each r from 0 to 9. The frequencies of $r = 0, 1, \ldots, 9$ are all equal to $Pr(r) = N/10$, which is 100.

表 2: Expected frequency

Digit, r	0	1	2	3	4	5	6	7	8	9	Total
Observed Frequency f_r	106	88	97	101	92	103	96	112	114	91	$N = 1000$
Expected frequency E_r	100	100	100	100	100	100	100	100	100	100	$N = 1000$

The statistic

$$\mathbf{X}^2 = \sum_{r=0}^{9} \frac{(O_r - E_r)^2}{E_r}$$

is the basis of this test. In this instance,

$$\begin{aligned} X^2 &= \tfrac{(106-100)^2}{100} + \tfrac{(88-100)^2}{100} + \tfrac{(97-100)^2}{100} + \tfrac{(101-100)^2}{100} + \tfrac{(92-100)^2}{100} \\ &\quad + \tfrac{(103-100)^2}{100} + \tfrac{(96-100)^2}{100} + \tfrac{(112-100)^2}{100} + \tfrac{(114-100)^2}{100} + \tfrac{(91-100)^2}{100} \\ &= (36 + 144 + 9 + 1 + 64 + 9 + 16 + 144 + 196 + 81)/100 \\ &= 7.00 \end{aligned}$$

Obviously we shall obtain a different value of X^2 each time we take a fresh sample of 1000 digits: X^2 has a sampling distribution which does not follow any of the distribution we have already studied.

The χ^2 distribution and its degree of freedom(カイ二乗分布と自由度)

Fig. 1 shows the general shape of χ^2 distributions. The mean of $\chi^2_{(\nu)}$ is ν, and the variance is $2\,\nu$. These distributions only become approximately symmetrical when ν is very large. The χ^2 family of distributions is usually tabulated as in Table A4. From the table, $\chi^2_{(3)}$ is greater than 7.81 with probability 0.05, $\chi^2_{(10)}$ is greater then 23.21 with probability 0.01, $\chi^2_{(5)}$ is greater than 12.83 with probability 0.025, and $\chi^2_{(5)}$ is greater than 0.83 with probability 0.975. As the diagram beneath Table A4 shows, we tabulate the values of $\chi^2_{(\nu)}$ which have to the right of them the areas stated in the column headings. From Table A4, it is also clear that when ν is large, the probability of large value of χ^2 is not negligible.

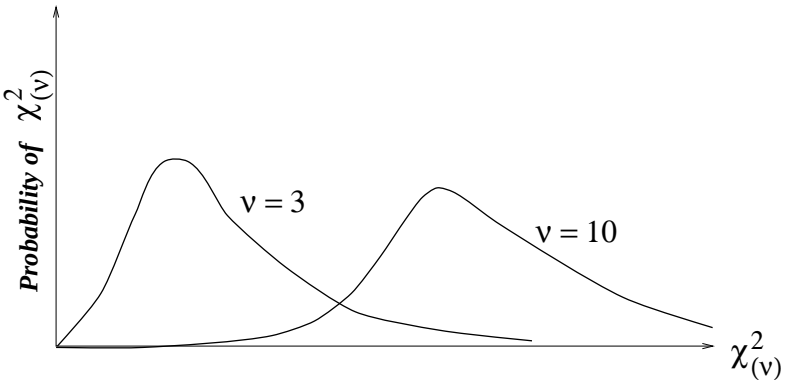

図 1: The χ^2 distribution

Example:
Find constants a, b, c, d, e, f, g, h such that: (次の定数 a, b, c, d, e, f, g, h を求めよ)

(i) $Pr(\chi^2_{(1)} > a) = 0.05$;

(ii) $Pr(\chi^2_{(2)} > 5.99) = b$;

(iii) $Pr(\chi^2_{(c)} > 59.70) = 0.001$;

(iv) $Pr(\chi^2_{(4)} > d) = 0.025$;

(v) $Pr(23.68 < \chi^2_{(14)} < 29.14) = e$;

(vi) $Pr(f < \chi^2_{(6)} < g) = 0.025$; ($f$, g are not unique)

(vii) $Pr(16.79 < \chi^2_{(h)} < 46.98) = 0.95$;

The rule for finding degrees of freedom in χ^2 tests
(カイ二乗検定中に自由度を求めるルール)
In order to find ν, the degrees of freedom of the χ^2 distribution which approximates X^2, we first count the number of cells in the table, that is the number of pairs (O_r, E_r) a variable for comparison. This is 10 in our example. Then we ask what restrictions or constraints we have placed on the set $\{E_r\}$. In the present case we require that the total of the members of the set $\{E_r\}$ should equal the observed total N, i.e., $\sum_{r=0}^{9} E_r = N = 1000$. This is a linear constraint on the members of the set $\{E_r\}$, in other words a linear equation which they must satisfy. In any χ^2 test, this constraint applies. In this example, it is the only one constraint.

$$Degree\ of\ freedom\ =\ number\ of\ cells - number\ of\ linear\ constrains\ on\ the\ expected\ frequencies$$

For the above example, $df = \nu = (10 - 1) = 9$

Example:
χ^2 will be approximately distributed as $\chi^2_{(9)}$ if the null hypothesis is true. The computed value of χ^2 was 7.00 . By reference to Table A4 we see that such a value is by no mean an unlikely one for a $\chi^2_{(9)}$ random variable, and we cannot reject the null hypothesis. On the evidence of this test, it is reasonable to regard the observed set of 1000 digits as random.

Testing the fit of data to a binomial distribution
(二項分布に合致するデータの検定)

We have collected a set of observation and wish to test whether they do conform to a binomial distribution. A survey was made of the numbers of boys among families having five children altogether. In 320 families, the number of boys R occurred with the following frequencies.

表 3:

Number of boys in a family	0	1	2	3	4	5	Total families	Total boys
Observed number of families O_r	8	40	88	110	56	18	$N = 320$	860
Expected frequencies E_r	10	50	100	100	50	10	$N = 320$	800

If births are all independent of one another, and the probability p of a male birth is the same from one family to another, R should be binomially distributed with parameters $n = 5$ and p. First let us suppose that $p = \frac{1}{2}$. The null hypothesis is now fully specified : R is binomial with parameters $n = 5$ and $p = \frac{1}{2}$. This gives the set of expected frequencies

$$E_r = N \times Pr(r) = N \frac{n!}{r!(n-r)!} \left(\frac{1}{2}\right)^5, \quad r = 0, 1, \ldots, 5$$

$$\mathbf{X}^2 = \sum_{r=0}^{5} \frac{(O_r - E_r)^2}{E_r} = 11.96$$

This statistic is based on six pairs (O_r, E_r) and the E_r values are subject to one linear constraint, so $\mathbf{X}^2$ is approximately χ^2 with $(6-1) = 5$ degrees of freedom. It is significant at the 5% level (the 5% point for $\chi^2_{(5)}$ is 11.07). So at this level we reject the null hypothesis. One alternative is that the binomial conditions do still hold, with $p \neq \frac{1}{2}$. If $p \neq \frac{1}{2}$, we must estimate p from the data. For a binomial distribution, the mean $\bar{r}$ is $p \times n$. Thus $\bar{r}/n$ will estimate p. We find $\bar{r} = \frac{no.\ of\ boys}{no.\ of\ families} = \frac{860}{320} = \frac{43}{16}$. The estimate of p is then $\frac{1}{5} \times \frac{43}{16} = 0.5375$. Let us calculate E_r with this p value. Then

$$N\,Pr(r) = N \begin{pmatrix} 5 \\ r \end{pmatrix} p^r (1-p)^{5-r} = N \frac{5!}{r!(5-r)!} (0.5375)^r (0.4625)^{(5-r)}, \quad r = 0, 1, \ldots, 5$$

表 4:

r	0	1	2	3	4	5	Total	Total boys
O_r	8	40	88	110	56	18	$N = 320$	860
E_r	6.8	39.3	91.5	106.3	61.8	14.4	(320.1)	860

$$\mathbf{X}^2 = \frac{(8-8.6)^2}{6.8} + \frac{(40-39.3)^2}{39.3} + \frac{(88-91.5)^2}{91.5} + \frac{(110-106.3)^2}{106.3} + \frac{(56-61.8)^2}{61.8} + \frac{(18-14.4)^2}{14.4} = 1.93$$

The expected values were calculated subject to two constrains this time: as usual, $\sum_r E_r = N$, but also this time the mean value of r calculated using the expected frequencies has to be equal to the mean using the observed frequencies, because this was the equation that we used to estimate p ($\sum_r r \times O_r = \sum_r r \times E_r$). Thus $\mathbf{X}^2$ will be distributed approximately as χ^2 with

$(6-2)=4$ degrees of freedom. The probability that X^2 takes value 1.93 is certainly significant and we shall not reject the null hypothesis that the data were binomially distributed. This result suggests that π is greater than $\frac{1}{2}$, but that otherwise the binomial conditions are reasonable.

Example:
Four coins are thrown 160 times, and the distribution of the number of heads is observed to be
(4 つのコインを 160 回投げ、表のでる分布は次のようになる)

表 5: Fair coin tossing experiment

x number of heads	0	1	2	3	4
f frequency O_r	5	35	67	41	12
f frequency E_r	10	40	60	40	10

Find the expected frequencies if the coins are unbiased. Compare the observed and expected frequencies and apply the χ^2 test. Is there any evidence that the coins are biased?
(コインに偏りがない時の望ましい回数を求めよ。またそれをカイ 2 乗検定で表し、コインに偏りがあるかを示せ)

Problem
A bag contains a very large number of black marbles and white marbles. 8192 random samples of 6 marbles are drawn from the bag. The frequencies of the number of black marbles in these samples are tabulated below:
(バックの中に大量の黒色と白色のビー玉が入っている。6 個のビー玉を 8192 回引くとき、黒いビー玉を引いてくる個数を調べると以下の表のようになった)

表 6: Marble picking experiment

x Number of black marbles per sample	0	1	2	3	4	5	6	Total
f frequencies	3	42	255	1115	2505	2863	1409	8192

Test the hypothesis that the ratio of the numbers of black to white marbles in the bag is 3:1.
(バックの中の黒色と白色のビー玉の割合が 3:1 になるようにカイ 2 乗検定を行え)

Correlation(相関):
Univariate & Multivariate data(一変数データと多変数データ)

表 7: Mathematics and Physics scores

Student Name	A	B	C	D	E	F	G	H	J	K	L	Total	Mean
Mathematics mark, x	41	37	38	39	49	47	42	34	36	48	29	440	40
Physics mark, y	36	20	31	24	37	35	42	26	27	29	23	330	30

The above is the Physics and Maths score of 11 high school students. We are looking for any relation between these two sets of marks. For example, we ask if above-average marks in mathematics usually go with above-average marks in physics, and below-average mathematics marks usually go with below-average physics marks.
The dotted horizontal and vertical lines in Fig. 2(a) show the mean marks 40 for mathematics, 30 for physics. We see that almost all the points fall in either the first (top-right) or third (bottom-left) quadrant.
Two other sets of records we consider next. Mathematics marks X against pottery (陶芸) marks Y (Fig. 2(b)). The points appear spread at random all over the graph, there is no indication of a relation and we may infer that ability in mathematics is not associated with ability in pottery.

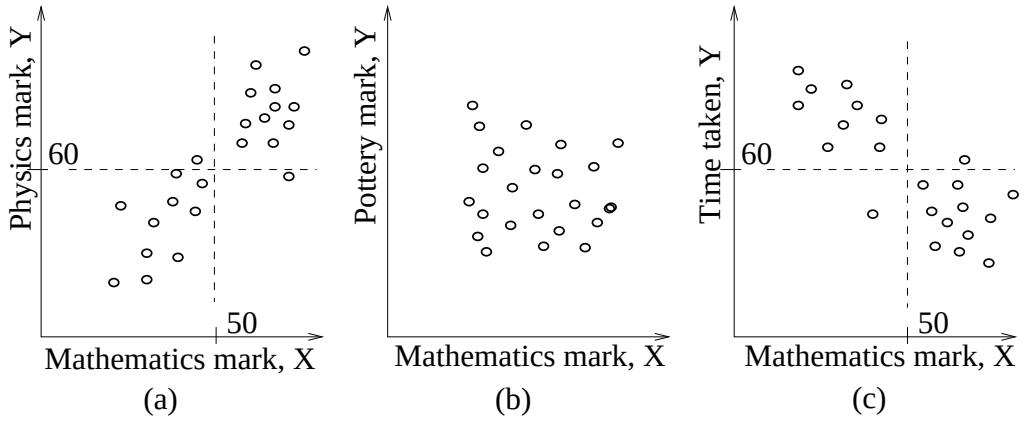

図 2: Scatter plot of marks

There does appear to be a relation between the mathematics mark and the time taken in answering the examination (Fig. 2(c)). The majority of points appear in either the second (top-left) or fourth (bottom-right) quadrant. The correlation between mathematics and physics marks is said to be positive (positive slope), while the relation between mathematics mark and time taken is called a negative correlation(negative slope). In the case of mathematics and pottery marks we would say there is no correlation.

Example:
For each of the sets of data below, draw a scatter diagram and assess whether there appears to be correlation between the two measurements labeled X and Y.
(次に示す X と Y で表されている測定値の相関関係を表しているかどうかを、散布図を示し評価を行え)

1. An owner of greyhounds notes the dogs' weights when they enter a race and their finishing positions in the race.
(グレイハウンドのレース前の体重と、その時の順位の記録)

表 8: Dogs' weight vs. position in race

Dog's weights (lb),X	60	63	70	65	60	64	67	73	56	58	60	60	66	55
Finishing position, Y	2	6	2	4	6	5	4	2	3	2	1	3	3	3
Dog's weights (lb),X	61	60	64	53	68	65	55	60	60	65				
Finishing position, Y	1	2	3	1	3	2	2	3	7	2				

2. Vehicles and road deaths - latest available figures for each country
(交通事故に対する死者の数の国ごとの集計)

表 9: Country vs. Road accident deaths

Country	Vehicles per 100 population. X	Road deaths per 100,000 population. Y
Great Britain	31	14
Belgium	32	29
Denmark	30	22
France	47	32
West Germany	30	25
Irish Republic	19	20
Italy	36	21
Netherlands	40	22
Canada	47	30
USA	58	35

3 The weight and average daily food consumption were measured for 12 obese adolescent girls. (12 人の若い肥満女性の一日平均で食べる量と体重の関係)

表 10: Weight vs. food consumption

Weight(Kg), X	84	93	81	61	95	86	90	78	85	72	65	75
Food consumption, Y												
(hundred calories per day)	32	33	33	24	39	32	34	28	33	27	26	29

The correlation coefficient(相関係数)

First we shall redraw the scatter diagram. Now X-axis is deviation $d_x = x - \bar{x}$ of a mathematics mark from the mean, and Y-axis is the deviation $d_y = y - \bar{y}$ of a physics mark from the mean of all physics marks.

In quadrants 1 and 3 the product $d_y d_x$ is positive, since d_x and d_y are either both positive (quadrant 1) or both negative (quadrant 3). Thus in the case of positive correlation $d_y d_x$ will be positive, while in the case of negative correlation the products $d_x d_y$ will be negative. $\sum d_y d_x$ will be large in magnitude (either positive or negative in sign) when there is correlation, and small in magnitude where there is no correlation. The problem is that the correlation depends on the units in which X and Y are measured. It is possible to define a coefficient which is independent of both scales of measurement by dividing $\sum d_x d_y$ by $\sqrt{(\sum d_x^2)(\sum d_y^2)}$.

Definition: The correlation coefficient between n pairs of observations, whose values are (x_i, y_i) is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{i=1}^{n}(x_i - \bar{x})^2]\,[\sum_{i=1}^{n}(y_i - \bar{y})^2]}} = \frac{\sum d_x d_y}{\sqrt{(\sum d_x^2)(\sum d_y^2)}}$$

表 11: Deviations from mean marks in mathematics and physics for eleven students

Students' name	A	B	C	D	E	F	G	H	J	K	L	$\sum d_x^2$ etc.
Deviation of mathematics mark from mean d_x	1	-3	-2	-1	9	7	2	-6	-4	-8	-11	$\sum d_x^2 = 386$
Deviation of physics mark from the mean. d_y	6	-10	1	-6	7	5	12	-4	-3	-1	-7	$\sum d_y^2 = 466$
Product $d_x d_y$	6	30	-2	6	63	35	24	24	12	-8	77	$\sum d_x d_y = 267$

From table, We have $\sum d_x d_y = 267$. We find also that $\sum d_x^2 = 386$ and $\sum d_y^2 = 466$. So, for the example of mathematics and physics marks, $r = 267/\sqrt{386 \times 466} = 0.630$. This correlation coefficient r is often referred to as the *product-moment* correlation coefficient. r takes values only in the interval -1 to +1. It takes the values ± 1 when there is an exact straight line relation $Y = mx + c$. it is not helpful in detecting more general curved (non-linear) relationships like exponential or circular. When there are n pairs of $\langle x_i, y_i \rangle$ data,

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-\frac{1}{n}(\sum x_i)(\sum y_i)=\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}$$

Therefore an alternative formula for r, useful for calculation purposes, is

$$r=\frac{\sum_{i=1}^{n}x_iy_i-(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)/n}{\sqrt{\left[\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2/n\right]\left[\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2/n\right]}}$$

Example:

Ten people were asked a set of questions designed to measure their attitude to television as a news medium, and another set to measure their attitude to newspapers. A higher overall score shows greater satisfaction. The scores are shown in the following table. Calculate the correlation coefficient between the two scores. Draw a scatter diagram to illustrate them. (10人の人々に対し、報道機関としてのテレビの条件を測定するための質問と、新聞に対して同じ条件を測定するための質問を行った。高いスコアを示しているのは満足しているという意味である。二つのスコア間の相関係数を求め、散布図をかけ)

表 12: Attitude towards news source

persons' name	A	B	C	D	E	F	G	H	I	J
TV score, X	5	0	3	1	2	2	5	3	5	4
Newspaper score, Y	1	2	1	3	3	4	3	1	0	2

Regression Analysis (回帰分析)

The basic aim of regression analysis is to find the best-fit function for a set of data. By best fit we mean that the sum-squared error is minimized, i.e., the sum of the squares of differences between predicted value (using the regression function) and the actual data value is minimized. Of course, a complex non-linear regression function will always give the minimum sum-squared error. Yet, in most of the practical applications, we look for a linear fit. Moreover, if we segmentize the whole region of interest, we can always find a good linear function for different segments. There are many non-linear regression techniques and learning algorithms, including the well-established artificial neural network. But those are beyond this syllabus. Here we will discuss linear regression only.

Linear regression

Suppose we have the data for GDP and life-expectancy for several European countries as shown in Table. 13. We want to know whether there is any relation between *per capita GDP (x)* and the *life-expectancy (y)*. In other words, is it possible to predict *life expectancy* from *per capita GDP*? One way to look into this problem is to find out the relation/function between these two variables, if there is any. That is, to find Ψ, where *Life-expectancy*=Ψ(per-capita-GDP). In our discussion, we restrict Ψ to be linear and it should give minimum sum-squared error.

表 13: Per Capita GDP vs. Life Expectancy

Country	Per Capita GDP (000s)	Life Expectancy	Country	Per Capita GDP (000s)	Life Expectancy
Austria	21.4	77.48	Ireland	18.6	76.39
Belgium	23.2	77.53	Italy	21.5	78.51
Finland	20.0	77.32	Netherlands	22.0	78.15
France	22.7	78.63	Switzerland	23.8	78.99
Germany	20.8	77.17	U. Kingdom	21.2	77.37

One way to do that is to plot all these pair of data, namely (21.4, 77.48), (23.2, 77.53) etc., and see whether we can draw a straight line which more or less joins all the data points. This could be tricky. Depending on the relative unit-spacing (scaling of x-axis and y-axis) the linear relation may not be visible. If the y-axis is too compressed, it would appear that there is no correlation.

Equation of the least squares regression line: We can write the linear regression line (i.e., our linear regression function Ψ) as follows:

$$y = \Psi(x) = b_1\, x + b_0$$

where, b_0 and b_1 is to be determined such that the error,

$$Error = \sum_{i=1}^{n} (y_i - (b_1\, x_i + b_0))^2$$

is minimized. By differentiating, it is easy to find the optimum values of b_0 and b_1 as follows:

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\, \bar{x}$$

Example:

Find the linear regression line of son's height, where 9 available data are as follows:

表 14: Father and his eldest son's height

Father's height (x)	147	152	157	162	167	172	177	182	190
Son's height (y)	160	165	167	167	170	172	175	180	180