Statistics for Medical Science

Analysis of categorical data

Mi Mi Ko, PhD
Korea Institute of Oriental Medicine

2025-12-04

Syllabus

Week		Description
Thu 14:00-17:00		
9	(2025-10-30)	Design of clinical studies
10	(2025-11-06)	Estimating and hypothesis testing
11	(2025-11-13)	No class (make-up will be scheduled: 12/12)
12	(2025-11-20)	Analysis of continuous data (1)
13	(2025-11-27)	Analysis of continuous data (2)
14	**(2025-12-04)**	**Analysis of categorical data**
15	**(2025-12-11)**	**Systematic Review & Meta-analysis**
16	**(2025-12-12)**	**Sample size calculation (14:00-)** 
17	**(2025-12-XX)**	**Final exam** (The details will be announced later)

Adequate Statistical Method

Purpose of analysis	Numerical data	Categorical data	Non-parametric test
Two groups	**t-test (Student t-test)**	**χ2-test (Pearson)** χ2-test with Yate's correction **χ^2-test for trend** **Fisher's exact test**	**Mann-Whitney(U) test**
More than Three groups	**One way ANOVA (F-test)** Two way ANOVA		**Kruskal-Wallis test** Friedman's Two way ANOVA
Pair (related groups)	**Paired t-test**	**McNemar's χ2-test**	**Wilcoxon signed-rank test**
correlation	**Pearson correlation**		**Spearman rank correlation**
Regression	Multiple linear regression	Multiple logistic regression	

Analysis of continuous data

- Continuous data : Comparison of Mean
- **Comparison of mean between the TWO group: (student) *t*-test (Mann-Whitney test)**

  Comparison of mean paired group : paired *t*-test

  (Wilcoxon signed-rank test)
- **Comparison of mean between above the THREE group: ANOVA (F-test) (Kruskal-Wallis test)**

  (multiple comparison)
- **Results : mean± SD, test statistic (t, F), p-value**

ANOVA (analysis of variance)

Levels of factor (group)	1	2	…	a
Data	y_{11} y_{12} . . . y_{1n_1}	y_{21} y_{22} . . . $y_{2\,n_2}$	…	y_{a1} y_{a2} . . . $y_{a\,n_a}$
MEAN $(\overline{y})$	$\overline{y}_1$	$\overline{y}_2$	…	$\overline{y}_a$
SD (S^2)	${S_1}^2$	${S_2}^2$	…	${S_a}^2$

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

H_1 : not H_0

ANOVA (analysis of variance)

❖ **Assumptions**

◆ The groups are defined by the *levels* of a single factor

(e.g. different ethnic backgrounds).

◆ In the population of interest, the variable is Normally distributed in each

group and the variance in each group is the same.

+ **Independence**

: The observations must be independent.

◆ We have a reasonable sample size so that we can check these assumptions.

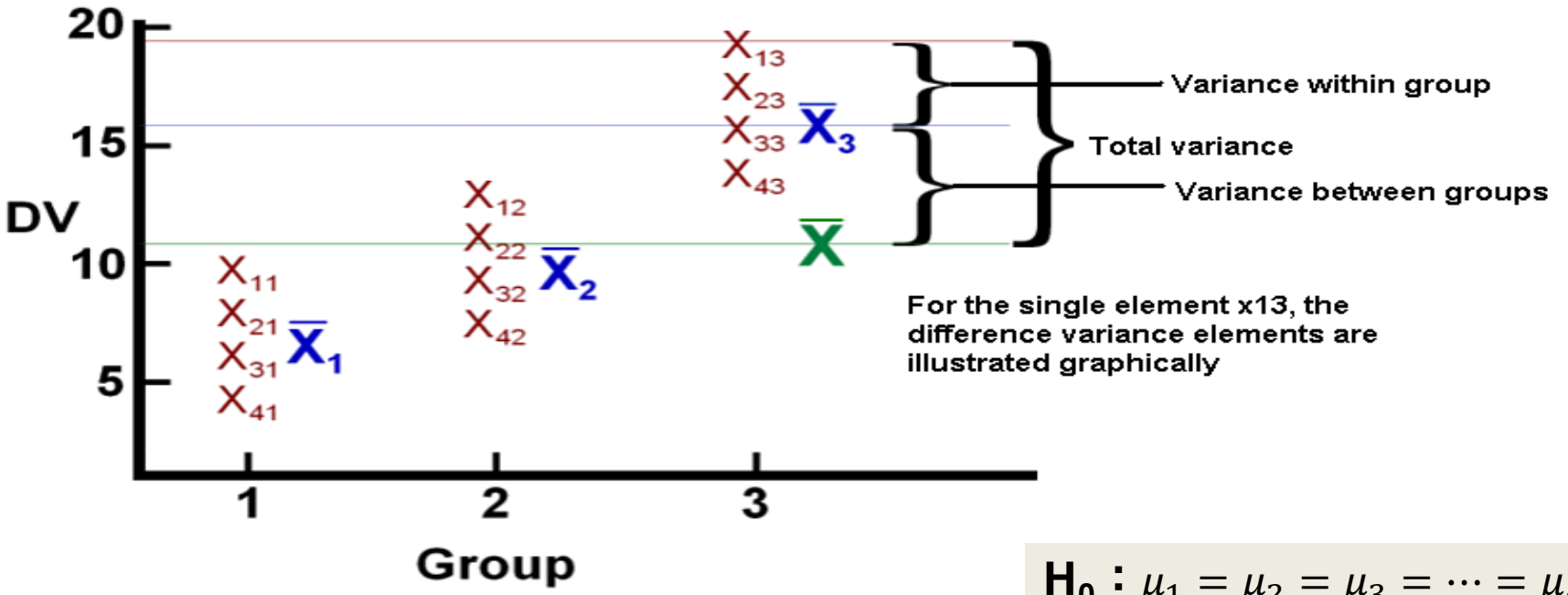
ANOVA (analysis of variance)

❖ **Assumptions**

- ◆ The groups are defined by the *levels* of a single factor (e.g. different ethnic backgrounds).
- ◆ In the population of interest, the variable is Normally distributed in each group and the variance in each group is the same.
- ◆ We have a reasonable sample size so that we can check these assumptions.

ANOVA (analysis of variance)

Partition of Sum of Squares



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

H_1 : not H_0

$$\underbrace{\sum_{i,j}\left(y_{ij} - \bar{\bar{y}}\right)^2}_{\text{Total (SST)}} = \underbrace{\sum_{i,j}\left(y_{ij} - \bar{y}_i\right)^2}_{\text{within (SSW)}} + \underbrace{\left(\bar{y}_i - \bar{\bar{y}}\right)^2}_{\text{between (SSB)}}$$

$$F = \frac{\text{(Between-group variation)}}{\text{(Within-group variation)}}$$

ANOVA (analysis of variance)

	SS	df	MS	F^*
Model	SSB	k-1	MSB=SSB/k-1	MSB/MSW
Error	SSW	n-k	MSW=SSW/n-k	
Total	SST	n-1		

$$F^* = \frac{MSB}{MSW} \sim F_{k-1,\ N-k}$$

F^* , P-value < 0.05 , Reject H_0

Ex.

ANOVA

beats per minute

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2387.685	2	1193.843	14.079	.000
Within Groups	3561.309	42	84.793		
Total	5948.994	44			

- ***F* statistic** is the ratio of the *MSB* to *MSW*

$$F_{stat} = \frac{MSB}{MSW} = \frac{1193.843}{84.793} = 14.08$$

Multiple comparison (post-hoc analysis)

- **Consider a comparison of three groups. There are three possible *t* tests when considering three groups:**

  (1) H_0: $\mu_1 = \mu_2$ versus H_a: $\mu_1 \neq \mu_2$

  (2) H_0: $\mu_1 = \mu_3$ versus H_a: $\mu_1 \neq \mu_3$

  (3) H_0: $\mu_2 = \mu_3$ versus H_a: $\mu_2 \neq \mu_3$

- ***However*, we do *not* perform separate *t* tests without modification → this would identify too many random differences**

Multiple comparison (post-hoc analysis)

The more comparisons you make, the greater the family−wise error rate. This table demonstrates the magnitude of the problem

Table 13.5 Family-wise error rates for multiple hypothesis tests.

No. of groups (k)	2	3	4	5	6	7	8	9	10
No. of pair-wise comparisons (c)	1	3	6	10	15	21	28	36	45
Pr(at least one P-value < 0.05)[1]	0.050	0.143	0.265	0.401	0.537	0.659	0.762	0.842	0.901

[1] Pr(at least one P-value less than 0.05) $= 1 -$ Pr(no P-value less than 0.05) $= 1 - 0.95^c$

Mitigating the Problem of Multiple Comparisons

Two-step approach:

1. **Test for overall significance using a technique called "Analysis of Variance"**

2. **Do *post hoc* comparison on individual groups**

Example 2

Quality-of-life scores, measured using the SF-36 questionnaire, were obtained in three groups of individuals: those with severe haemophilia, those with mild/moderate haemophilia, and normal controls. Each group comprised a sample of 20 individuals. Scores on the physical functioning scale (PFS), which can take values from 0 to 100, were compared in the three groups. As visual inspection of Fig. 22.1 showed that the data were not Normally distributed, we performed a **Kruskal–Wallis** test.

$$W = \frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(n+1)$$

Group	Severe	Mild/moderate	Controls
Sample size, n	20	20	20
Median (95% CI)	47.5 (30 to 80)	87.5 (75 to 95)	100 (90 to 100)
Range	0–100	0–100	0–100

Figure 22.1 Dot plot showing physical functioning scores (from the SF-36 questionnaire) in individuals with severe and mild/moderate haemophilia and in normal controls. The horizontal bars are the medians.

1 H_0: each group has the same distribution of PFS scores in the population

H_1: at least one of the groups has a different distribution of PFS scores in the population.

2 The data are shown in Fig. 22.1.

3 Sum of ranks in severe haemophilia group = 372
Sum of ranks in mild/moderate haemophilia group = 599
Sum of ranks in normal control group = 859.

$$H = \frac{12}{60(60+1)}\left(\frac{372^2}{20} + \frac{599^2}{20} + \frac{859^2}{20}\right) - 3(60+1) = 19.47$$

4 We refer H to Appendix A3: $P < 0.001$.

5 There is substantial evidence to reject the null hypothesis that the distribution of PFS scores is the same in the three groups. Pairwise comparisons were carried out using Wilcoxon rank sum tests, adjusting the P-values for the number of tests performed using the Bonferroni correction (Chapter 18). The individuals with severe and mild/moderate haemophilia both had significantly lower PFS scores than the controls ($P = 0.0003$ and $P = 0.03$, respectively) but the distributions of the scores in the haemophilia groups were not significantly different from each other ($P = 0.09$).

Data kindly provided by Dr A. Miners, Department of Primary Care and Population Sciences, Royal Free and University College Medical School, London, UK, and Dr C. Jenkinson, Health Services Research Unit, University of Oxford, Oxford, UK.

Analysis of categorical data

- Categorical data : proportion(s), %
- A single proportion : A single proportion (or one-sample) test
- Two proportions : **Chi-squared test** (χ^2-test)

  Related(paired) groups : **McNemar's test**
- more than two categories : Chi-squared test (χ^2-test)

  Chi-squared test for trend
- **Fisher's exact test** (E<5)

A single proportion test

❖ Problem

◆ We have **a single sample** of *n* individuals;

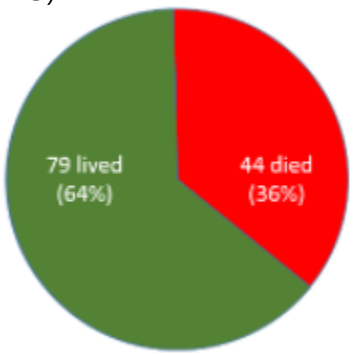
each individual either 'possesses' a characteristic of interest

(e.g. is male, is pregnant, has died) or

does not possess that characteristic

(e.g. is female, is not pregnant, is still alive).

ex. Mortality from COVID-19 in AA clinic
(Total N=123)



◆ A useful summary of the data is provided by

the **proportion** of individuals with the characteristic.

◆ We are interested in determining

whether the true proportion in the population of interest takes a particular value.

A single proportion test

❖ Assumption

- ◆ Our sample of individuals is selected from the population of interest.
- ◆ Each individual either has or does not have the particular characteristic.

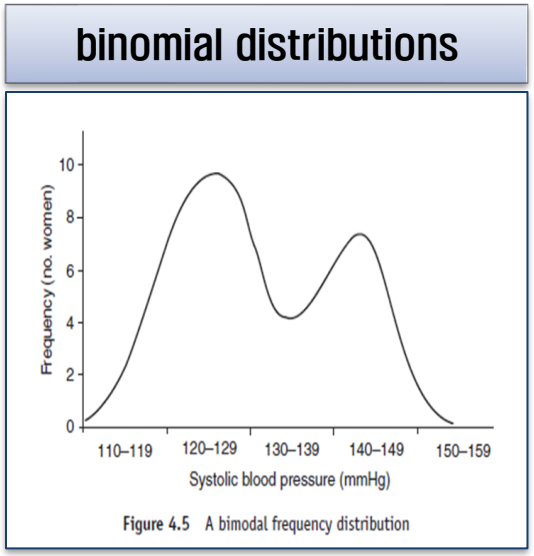
❖ Notation

- ◆ ***r*** individuals in our sample of size ***n*** have the characteristic.
- ◆ The estimated proportion with the characteristic is $p = r/n$.
- ◆ The proportion of individuals with the characteristic in the population is π.
- ◆ We are interested in determining whether π takes a particular value, π_1.

A single proportion test

❖ **Rationale**

◆ The number of individuals
with the characteristic follows
the Binomial distribution,
but this can be approximated by the **Normal distribution**,
providing **np** and ***np(1 - p)*** are each greater than **5**.



◆ Then ***p*** is approximately **Normally distributed** with **an estimated mean = *p*** and

an estimated $S.D. = \sqrt{\frac{p(1-p)}{n}}$

◆ Therefore, our **test statistic**, which is based on ***p***, also follows the **Normal distribution**.

A single proportion test

1. **Define the null and alternative hypotheses under study**
   - *H*0: the population proportion, π, is equal to a particular value, π_1 ($\pi = \pi_1$)
   - *H*1: the population proportion, π, is not equal to π_1 ($\pi \neq \pi_1$)
2. **Collect relevant data from a sample of individuals**
3. **Calculate the value of the test statistic specific to *H*0**

$$Z = \frac{|p - \pi_1| - \frac{1}{2n}}{\sqrt{\frac{\pi_1(1-\pi_1)}{n}}}$$

   which follows a Normal distribution.

   - The $\frac{1}{2n}$ in the numerator is a **continuity correction**: it is included to make an allowance for the fact that we are approximating the discrete Binomial distribution by the continuous Normal distribution.
4. **Compare the value of the test statistic to values from a known probability distribution**
   - Refer *z* to **z table**.
5. **Interpret the *P*-value and results**
   - Interpret the *P*-value and calculate a confidence interval for the true population proportion, p.

     The 95% confidence interval for π is: $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$

We can use this confidence interval to assess the clinical or biological importance of the results. A wide confidence interval is an indication that our estimate has poor precision.

Example

Human herpes-virus 8 (HHV-8) has been linked to Kaposi's sarcoma, primary effusion lymphoma and certain types of multicentric Castleman's disease. It has been suggested that HHV-8 can be transmitted sexually. In order to assess the relationships between sexual behaviour and HHV-8 infection, the prevalence of antibodies to HHV-8 was determined in a group of 271 homo/bisexual men attending a London sexually transmitted disease clinic. In the blood donor population in the UK, the seroprevalence of HHV-8 has been documented to be 2.7%. Initially, the seroprevalence from this study was compared to 2.7% using a **single proportion** test.

1 H_0: the seroprevalence of HHV-8 in the population of homo/bisexual men equals 2.7% $\pi_1 = 0.027$

H_1: the seroprevalence of HHV-8 in the population of homo/bisexual men does not equal 2.7%. $\pi_1 \neq 0.027$

2 Sample size, $n = 271$; number who are seropositive to HHV8, $r = 50$

Seroprevalence, $p = 50/271 = 0.185$ (i.e. 18.5%). $z = \dfrac{|p-\pi_1|-\frac{1}{2n}}{\sqrt{\frac{\pi_1(1-\pi_1)}{n}}}$

3 Test statistic is $z = \dfrac{|0.185-0.027|-\frac{1}{2\times 271}}{\sqrt{\frac{0.027(1-0.027)}{271}}} = 15.86$

4 We refer z to Appendix A1: $P < 0.0001$.

5 There is substantial evidence that the seroprevalence of HHV-8 in homo/bisexual men attending sexually transmitted disease clinics in the UK is higher than that in the blood donor population. The 95% confidence interval for the seroprevalence of HHV-8 in the population of homo/bisexual men is 13.9% to 23.1%, calculated as

$$\left\{0.185 \pm 1.96 \times \sqrt{\frac{0.185\times(1-0.185)}{271}}\right\} \times 100\%.$$

Data kindly provided by Drs N.A. Smith, D. Barlow, and B.S. Peters, Department of Genitourinary Medicine, Guy's and St Thomas' NHS Trust, London, and Dr J. Best, Department of Virology, Guy's, King's College and St Thomas's School of Medicine, King's College, London, UK.

Table A1 Standard Normal distribution.

z	2-tailed P-value
0.0	1.000
0.1	0.920
0.2	0.841
0.3	0.764
0.4	0.689
0.5	0.617
0.6	0.549
0.7	0.484
0.8	0.424
0.9	0.368
1.0	0.317
1.1	0.271
1.2	0.230
1.3	0.194
1.4	0.162
1.5	0.134
1.6	0.110
1.7	0.089
1.8	0.072
1.9	0.057
2.0	0.046
2.1	0.036
2.2	0.028
2.3	0.021
2.4	0.016
2.5	0.012
2.6	0.009
2.7	0.007
2.8	0.005
2.9	0.004
3.0	0.003
3.1	0.002
3.2	0.001
3.3	0.001
3.4	0.001
3.5	0.000

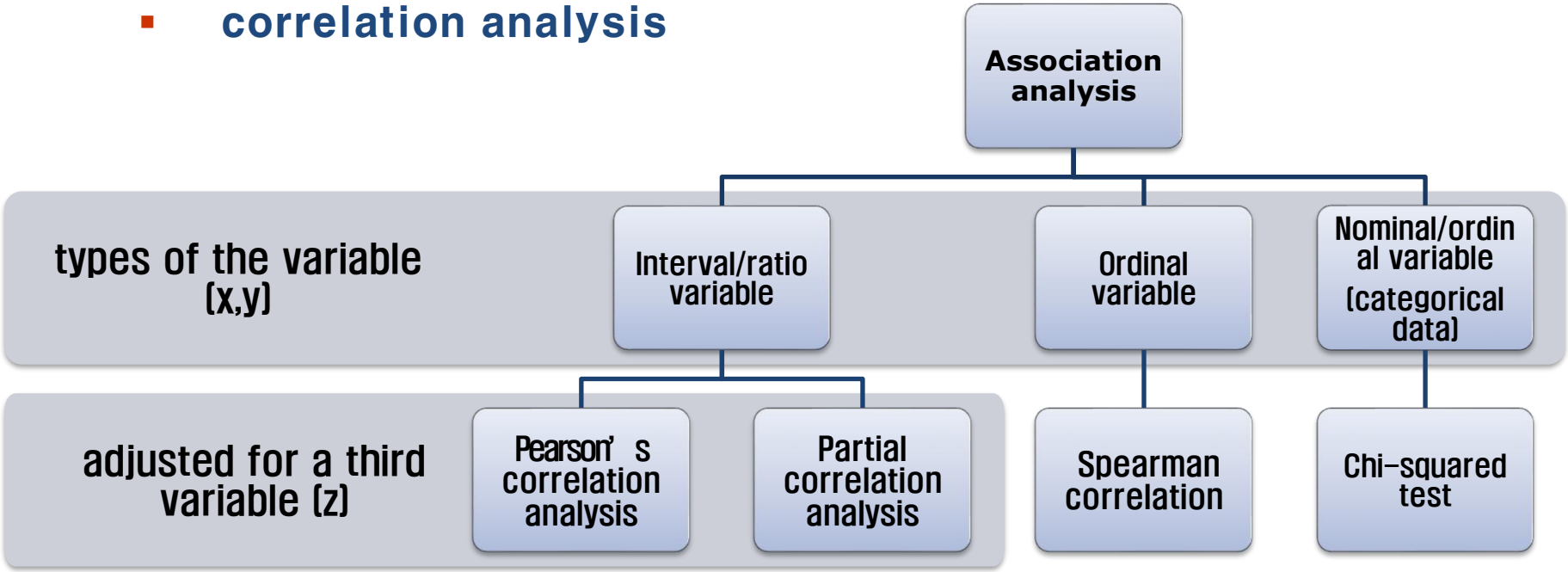
Derived using Microsoft Excel Version 5.0.

Association analysis

❖ method to determine whether two variables are independent of each other or whether there is any relationship between them

◆ **Types of the variable**

- **cross tabulation(tab) analysis (Chi-squared test)**
- **correlation analysis**



Two proportions test: Chi-squared test (χ^2-test)

❖ Problem

◆ We have two **independent** groups of individuals

(e.g. homosexual men with and without a history of gonorrhoea).

- We want to know if the proportions of individuals with a characteristic

(e.g. infected with human herpesvirus-8, HHV-8) are the same in the two groups.

➔ **Chi-squared test**

◆ We have two **related** groups, e.g. individuals may be matched, or measured twice in different circumstances (say, before and after treatment).

- We want to know if the proportions with a characteristic

(e.g. raised test result) are the same in the two groups. ➔ **McNemar's test**

Chi-squared test (χ^2-test): Terminology

- The data are obtained, initially, as **frequencies**,
  i.e. the numbers with and without the characteristic in each sample.
- A table in which the entries are frequencies is called a **contingency table**; when this table has two rows and two columns it is called a **2 × 2 table**.
- Table 24.1 shows the **observed** frequencies in the four cells corresponding to each row/column combination, the four **marginal totals** (the frequency in a specific row or column, e.g. *a* + *b*), and the **overall total**, *n.*

Table 24.1 Observed frequencies.

Characteristic	Group 1	Group 2	Total
Present	a	b	$a+b$
Absent	c	d	$c+d$
Total	$n_1=a+c$	$n_2=b+d$	$n=a+b+c+d$

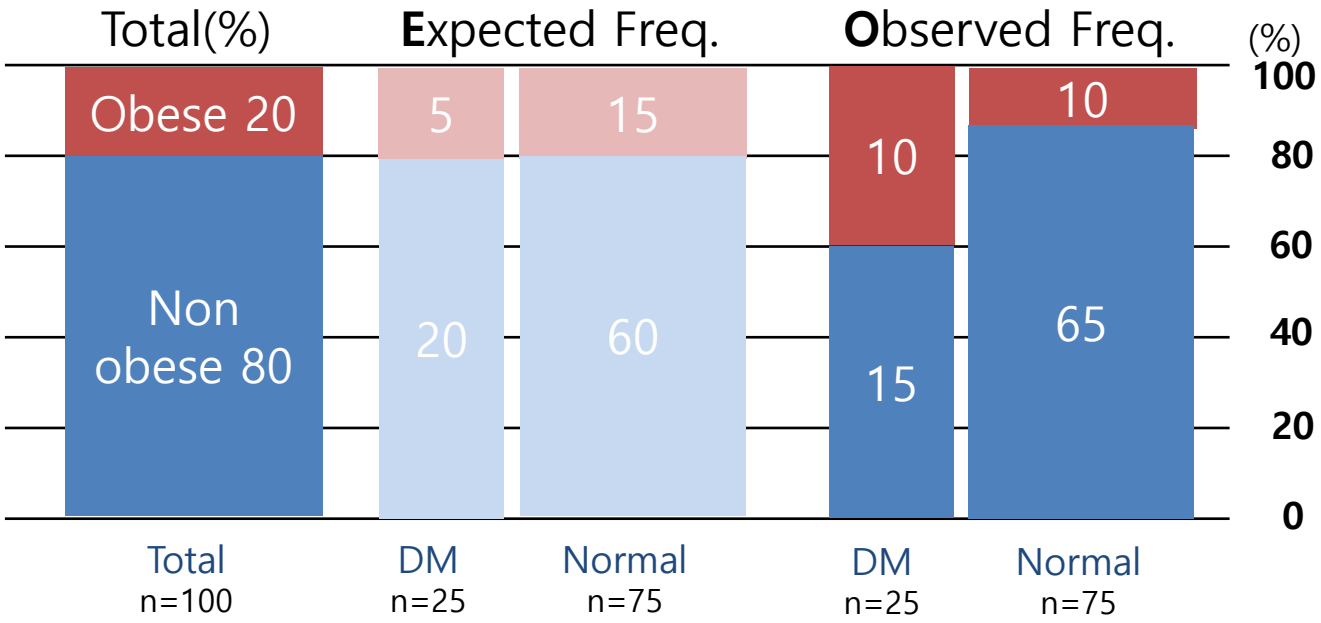
- We can calculate the frequency that we would expect in each of the four cells of the table if H_0 were true (the **expected frequencies**).

Chi-squared test (χ^2-test): Assumptions

- We have samples of sizes *n1* and *n2* from two independent groups of individuals.
- We are interested in
  whether the proportions of individuals who possess the characteristic are the same in the two groups .
- Each individual is represented only once in the study.
  - The rows (and columns) of the table are mutually exclusive, implying that each individual can belong in only one row and only one column.
- The usual, albeit conservative, approach requires that the
  expected frequency in each of the four cells is at least five (E>5).

Chi-squared test (χ^2-test): Terminology

	DM	Normal	Total
Obese	10 (40.0%)	10 (13.3%)	20 (20%)
Non obese	15 (60.0%)	65 (86.7%)	80 (80%)
Total	25 (100%)	75 (100%)	100 (100%)



Chi-squared test (χ^2-test): Terminology

	DM	Normal	Total
Obese	10 [5]	10 [15]	20 (20%)
Non obese	15 [20]	65 [60]	80 (80%)
Total	25	75	100 (100%)

Chi-square (Pearson): $\chi^2 = \sum \frac{(O - E)^2}{E}$

- Observed vs. Expected, [ ]
- NO ASSOCIATION : observed freq. = expected freq.
- ASSOCIATION : observed freq. ≠ expected freq.

a significant chi-square statistic, there is strong evidence that an association exists

Chi-squared test (χ^2-test): Terminology

- Used for comparison on frequency
- test statistic χ^2 : Difference between observed and expected values

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

Table 24.1 Observed frequencies.

Characteristic	Group 1	Group 2	Total
Present	a	b	$a+b$
Absent	c	d	$c+d$
Total	$n_1 = a+c$	$n_2 = b+d$	$n = a+b+c+d$
Proportion with characteristic	$p_1 = \frac{a}{n_1}$	$p_2 = \frac{b}{n_2}$	$p = \frac{a+b}{n}$

Expected freq.	Group 1	Group 2	Total
Present	$n1 \times \frac{(a+b)}{n}$	$n2 \times \frac{(a+b)}{n}$	$a+b$
Absent	$n1 \times \frac{(c+d)}{n}$	$n2 \times \frac{(c+d)}{n}$	$c+d$
Total	$n1 = a+c$	$n2 = b+d$	$n = a+b+c+d$

Chi-squared test (χ^2-test): Terminology

	DM	Normal	Total
Obese	10 [5]	10 [15]	20 (20%)
Non obese	15 [20]	65 [60]	80 (80%)
Total	25	75	100 (100%)

$$\frac{(25 * 20)}{100}$$

$$\text{Expected frequency} = (\text{Column total}) \cdot \frac{(\text{Row total})}{\text{Grand total}}$$

$$\frac{(75 * 20)}{100}$$

	DM	Normal	Total
Obese	5	15	20
Non obese	20	60	80
Total	25	75	100

$$\frac{(25 * 80)}{100}$$

$$\frac{(75 * 80)}{100}$$

Chi-squared test (χ^2-test): Rationale

- If the proportions with the characteristic in the two groups are equal,
  we can estimate the overall proportion of individuals with the characteristic
  by ***p = (a + b)/n***;
  we **expect** *n1 × p* of them to be in Group 1 and *n2 × p* to be in Group 2.
- We evaluate expected numbers without the characteristic similarly.
- Therefore, each expected frequency is the product of the two relevant marginal totals divided by the overall total.
- A large discrepancy between the observed (O) and the corresponding expected (E) frequencies is an indication that the proportions in the two groups differ.
  - The test statistic is based on this discrepancy.

Chi-squared test (χ^2-test)

1. Define the null and alternative hypotheses under study
   - ◆ H_0: the proportions of individuals with the characteristic are equal in the two groups in the population. ($\pi_1 = \pi_2$)
   - ◆ H_1: these population proportions are not equal. ($\pi_1 \neq \pi_2$)
2. Collect relevant data from samples of individuals
3. Calculate the value of the test statistic specific to H_0

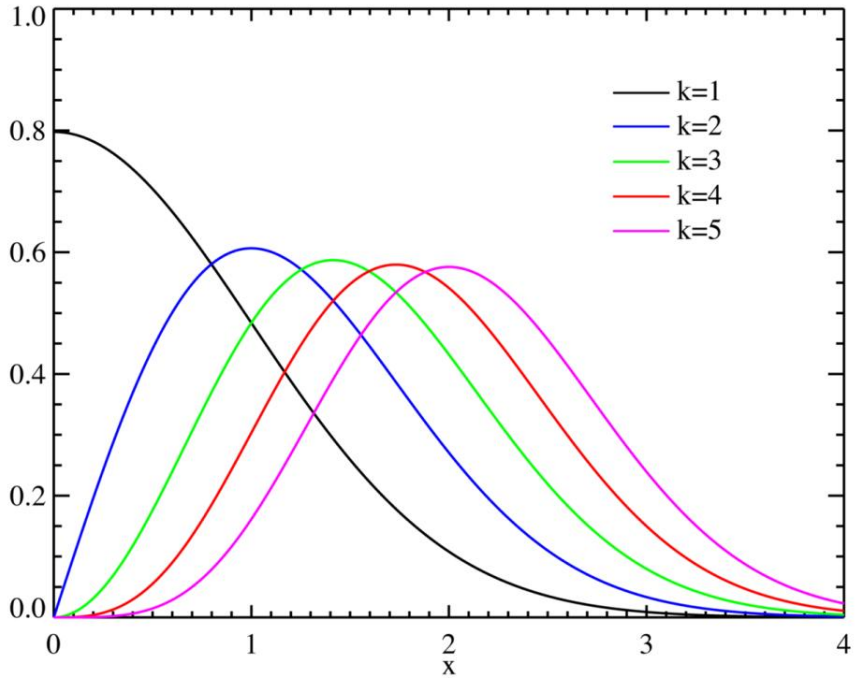
$$\chi^2 = \sum \frac{\left(|O - E| - \frac{1}{2}\right)^2}{E}$$

- ◆ where O and E are the observed and expected frequencies, respectively, in each of the four cells of the table (2*2 table).
  - The 1/2 in the numerator is the continuity correction.
  - The test statistic follows the **Chi-squared distribution** with **1 degree of freedom**.

Chi-squared test (χ^2-test)

4. Compare the value of the test statistic to values from a known probability distribution

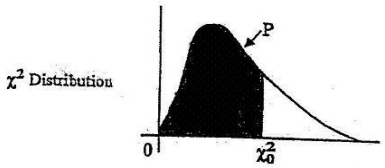
   - Refer χ^2 to χ^2 table.



5. Interpret the P-value and results

   - Interpret the *P*-value and calculate the confidence interval for the difference in the true population proportions.
   - The 95% confidence interval is given by:

$$(p_1 - p_2) \pm 1.96 \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$



The table below gives the value x_0^2 for which $P[x^2 < x_0^2] =$ P for a given number of degrees of freedom and a given value of P.

Degrees of Freedom	Values of P									
	0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.01	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997

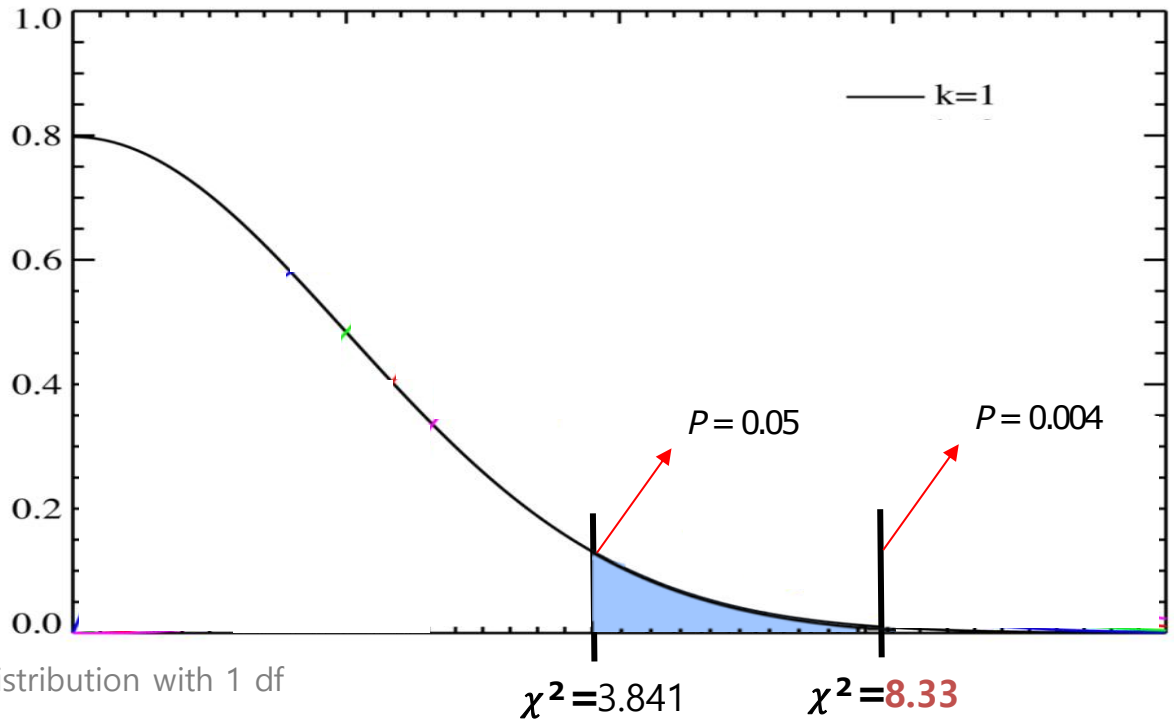
Table A3 Chi-squared distribution.

	Two-tailed *P*-value			
df	0.10	0.05	0.01	0.001
1	2.706	3.841	6.635	10.827
2	4.605	5.991	9.210	13.815
3	6.251	7.815	11.345	16.266
4	7.779	9.488	13.277	18.466
5	9.236	11.070	15.086	20.515
6	10.645	12.592	16.812	22.457
7	12.017	14.067	18.475	24.321
8	13.362	15.507	20.090	26.124
9	14.684	16.919	21.666	27.877
10	15.987	18.307	23.209	29.588
11	17.275	19.675	24.725	31.264
12	18.549	21.026	26.217	32.909
13	19.812	22.362	27.688	34.527
14	21.064	23.685	29.141	36.124
15	22.307	24.996	30.578	37.698
16	23.542	26.296	32.000	39.252
17	24.769	27.587	33.409	40.791
18	25.989	28.869	34.805	42.312
19	27.204	30.144	36.191	43.819
20	28.412	31.410	37.566	45.314
21	29.615	32.671	38.932	46.796
22	30.813	33.924	40.289	48.268
23	32.007	35.172	41.638	49.728
24	33.196	36.415	42.980	51.179
25	34.382	37.652	44.314	52.619
26	35.563	38.885	45.642	54.051
27	36.741	40.113	46.963	55.475
28	37.916	41.337	48.278	56.892
29	39.087	42.557	49.588	58.301
30	40.256	43.773	50.892	59.702
40	51.805	55.758	63.691	73.403
50	63.167	67.505	76.154	86.660
60	74.397	79.082	88.379	99.608
70	85.527	90.531	100.43	112.32
80	96.578	101.88	112.33	124.84
90	107.57	113.15	124.12	137.21
100	118.50	124.34	135.81	149.45

Derived using Microsoft Excel Version 5.0.

	DM	Normal	Total
Obese	10 [5]	10 [15]	20 (20%)
Non obese	15 [20]	65 [60]	80 (80%)
Total	25	75	100 (100%)

Chi-square (Pearson): $\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(+5)^2}{5} + \frac{(-5)^2}{20} + \frac{(-5)^2}{15} + \frac{(-5)^2}{60} = 8.33$



χ^2 distribution with 1 df

Table A3 Chi-squared distribution.

	Two-tailed *P*-value			
df	0.10	0.05	0.01	0.001
1	2.706	3.841	6.635	10.827
2	4.605	5.991	9.210	13.815
3	6.251	7.815	11.345	16.266
4	7.779	9.488	13.277	18.466
5	9.236	11.070	15.086	20.515
6	10.645	12.592	16.812	22.457
7	12.017	14.067	18.475	24.321
8	13.362	15.507	20.090	26.124
9	14.684	16.919	21.666	27.877
10	15.987	18.307	23.209	29.588
11	17.275	19.675	24.725	31.264
12	18.549	21.026	26.217	32.909
13	19.812	22.362	27.688	34.527
14	21.064	23.685	29.141	36.124
15	22.307	24.996	30.578	37.698
16	23.542	26.296	32.000	39.252
17	24.769	27.587	33.409	40.791
18	25.989	28.869	34.805	42.312
19	27.204	30.144	36.191	43.819
20	28.412	31.410	37.566	45.314
21	29.615	32.671	38.932	46.796
22	30.813	33.924	40.289	48.268
23	32.007	35.172	41.638	49.728
24	33.196	36.415	42.980	51.179
25	34.382	37.652	44.314	52.619
26	35.563	38.885	45.642	54.051
27	36.741	40.113	46.963	55.475
28	37.916	41.337	48.278	56.892
29	39.087	42.557	49.588	58.301
30	40.256	43.773	50.892	59.702
40	51.805	55.758	63.691	73.403
50	63.167	67.505	76.154	86.660
60	74.397	79.082	88.379	99.608
70	85.527	90.531	100.43	112.32
80	96.578	101.88	112.33	124.84
90	107.57	113.15	124.12	137.21
100	118.50	124.34	135.81	149.45

Derived using Microsoft Excel Version 5.0.

Example

In order to assess the relationship between sexual risk factors and HHV-8 infection (study described in Chapter 23), the prevalence of seropositivity to HHV-8 was compared in homo/bisexual men who had a previous history of gonorrhoea, and those who had not previously had gonorrhoea, using the **Chi-squared test**. A typical computer output is shown in Appendix C.

1 H_0: the seroprevalence of HHV-8 is the same in those with and without a history of gonorrhoea in the population $\pi_1 = \pi_2$

H_1: the seroprevalence is not the same in the two groups in the population. $\pi_1 \neq \pi_2$

2 The observed frequencies are shown in the following contingency table: 14/43 (32.6%) and 36/228 (15.8%) of those with and without a previous history of gonorrhoea are seropositive for HHV-8, respectively.

3 The expected frequencies are shown in the four cells of the contingency table.

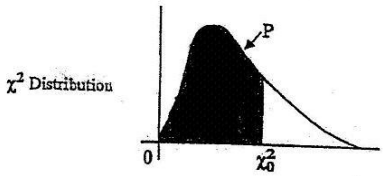
The test statistic is

$$\chi^2 = \left\{ \frac{(|14 - 7.93| - \frac{1}{2})^2}{7.93} + \frac{(|36 - 42.07| - \frac{1}{2})^2}{42.07} + \frac{(|29 - 35.07| - \frac{1}{2})^2}{35.07} + \frac{(|192 - 185.93| - \frac{1}{2})^2}{185.93} \right\} = 5.70$$

4 We refer χ^2 to Appendix A3 with 1 *df*: $0.01 < P < 0.05$ (computer output gives $P = 0.017$).

5 There is evidence of a real difference in the seroprevalence in the two groups in the population. We estimate this difference as 32.6% − 15.8% = 16.8%. The 95% CI for the true difference in the two percentages is 2.0% to 31.6%
i.e. $16.8 \pm 1.96 \times \sqrt{(\{32.6 \times 67.4\}/43 + \{15.8 \times 84.2\}/228)}$.

	Previous history of gonorrhoea				
	Yes		No		
HHV-8	Observed	Expected	Observed	Expected	Total observed
Seropositive	14	(43 × 50/271) = 7.93	36	(228 × 50/271) = 42.07	50
Seronegative	29	(43 × 221/271) = 35.07	192	(228 × 221/271) = 185.93	221
Total	43		228		271



The table below gives the value x_0^2 for which $P[x^2 < x_0^2] =$ P for a given number of degrees of freedom and a given value of P.

Degrees of Freedom	Values of P									
	0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.01	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997

Table A3 Chi-squared distribution.

	Two-tailed *P*-value			
df	0.10	0.05	0.01	0.001
1	2.706	3.841	6.635	10.827
2	4.605	5.991	9.210	13.815
3	6.251	7.815	11.345	16.266
4	7.779	9.488	13.277	18.466
5	9.236			
6	10.645			
7	12.017			
8	13.362			
9	14.684			
10	15.987			
11	17.275	19.675	24.725	31.264
12	18.549	21.026	26.217	32.909
13	19.812	22.362	27.688	34.527
14	21.064	23.685	29.141	36.124
15	22.307	24.996	30.578	37.698
16	23.542	26.296	32.000	39.252
17	24.769	27.587	33.409	40.791
18	25.989	28.869	34.805	42.312
19	27.204	30.144	36.191	43.819
20	28.412	31.410	37.566	45.314
21	29.615	32.671	38.932	46.796
22	30.813	33.924	40.289	48.268
23	32.007	35.172	41.638	49.728
24	33.196	36.415	42.980	51.179
25	34.382	37.652	44.314	52.619
26	35.563	38.885	45.642	54.051
27	36.741	40.113	46.963	55.475
28	37.916	41.337	48.278	56.892
29	39.087	42.557	49.588	58.301
30	40.256	43.773	50.892	59.702
40	51.805	55.758	63.691	73.403
50	63.167	67.505	76.154	86.660
60	74.397	79.082	88.379	99.608
70	85.527	90.531	100.43	112.32
80	96.578	101.88	112.33	124.84
90	107.57	113.15	124.12	137.21
100	118.50	124.34	135.81	149.45

Derived using Microsoft Excel Version 5.0.

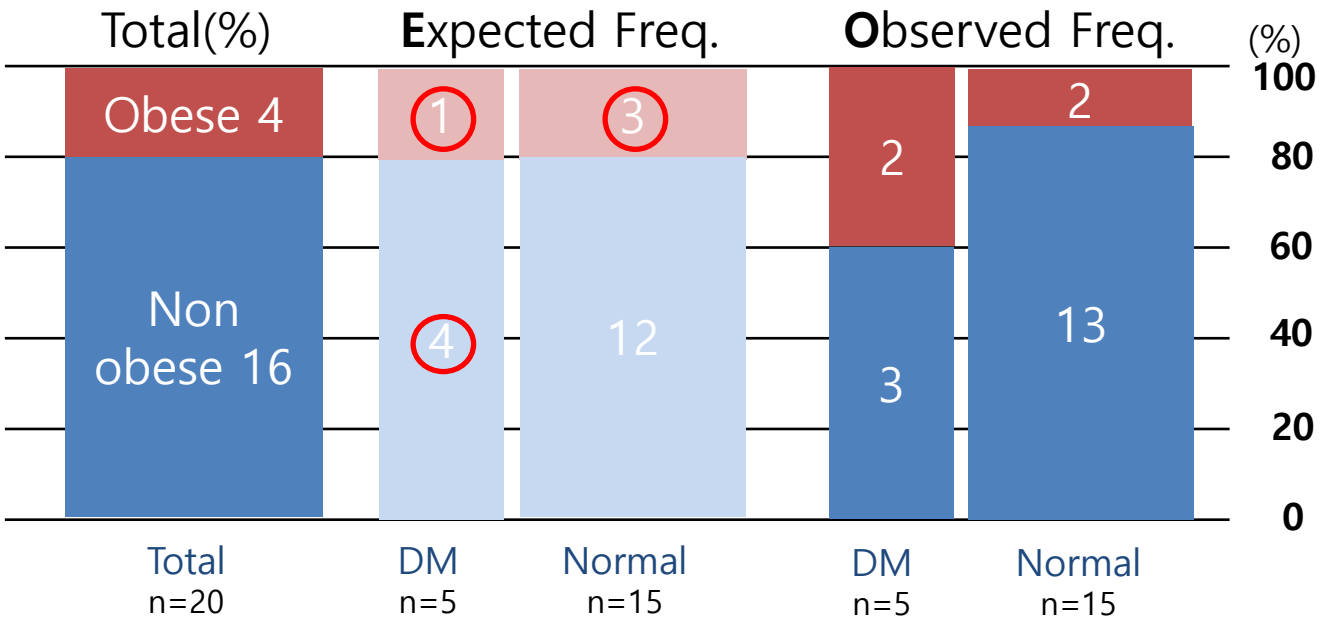
χ^2=5.70
0.01 < *P* < 0.05

If the assumptions are not satisfied

- If $E < 5$ in any one cell, we use **Fisher's exact test** to obtain a *P*-value that does not rely on the approximation to the Chi-squared distribution.
- This is best left to a computer program as the calculations are tedious to perform by hand.

Fisher's Exact Test

	DM	Normal	Total
Obese	2 (40.0%)	2 (13.3%)	4 (20%)
Non obese	3 (60.0%)	13 (86.7%)	16 (80%)
Total	5 (100%)	15 (100%)	20 (100%)



Fisher's Exact Test

	DM	Normal	Total
Obese	a	b	4
Non obese	c	d	16
Total	5	15	20

Obese n=4 (DM, Normal) = ① (4,0) ② (3,1) ③ (2,2) ④ (1,3) ⑤ (0,4)

Observed Freq.

Expected Freq.

①

	DM	Normal	Total
Obese	4	0	4
Non obese	1	15	16
Total	5	15	20

$$\frac{{}_5C_4 \times 1}{{}_{20}C_4} = 0.00103$$

②

	DM	Normal	Total
Obese	3	1	4
Non obese	2	14	16
Total	5	15	20

$$\frac{{}_5C_3 \times {}_{15}C_1}{{}_{20}C_4} = 0.03096$$

③

	DM	Normal	Total
Obese	2	2	4
Non obese	3	13	16
Total	5	15	20

$$\frac{{}_5C_2 \times {}_{15}C_2}{{}_{20}C_4} = 0.21672$$

④

	DM	Normal	Total
Obese	1	3	4
Non obese	4	12	16
Total	5	15	20

$$\frac{{}_5C_1 \times {}_{15}C_3}{{}_{20}C_4} = 0.46965$$

⑤

	DM	Normal	Total
Obese	0	4	4
Non obese	5	11	16
Total	5	15	20

$$\frac{1 \times {}_{15}C_4}{{}_{20}C_4} = 0.28173$$

① + ② + ③ =

=0.24871 > *P*=0.05

McNemar's χ^2-test

Rationale

- The observed proportions with the characteristic in the two circumstances are $(w + y)/m$ and $(w + x)/m$. They will differ if x and y differ.

  Therefore, to compare the proportions with the characteristic concentrate on the discordant pairs, x and y.

Table 24.2 Observed frequencies of pairs in which the characteristic is present or absent.

	Circumstance 1		
	Present	Absent	Total no. of pairs
Circumstance 2			
Present	w	x	$w+x$
Absent	y	z	$y+z$
Total	$w+y$	$x+z$	$m=w+x+y+z$

	Radiographic diagnosis		
Diagnosis on section	Cavities absent	Cavities present	Total
Cavities absent	45	4	49
Cavities present	17	34	51
Total	62	38	100

$$\text{McNemar } \chi^2 = \frac{(|x - y| - 1)^2}{(x + y)}$$

McNemar's χ^2-test

1 Define the null and alternative hypotheses under study

H_0: the proportions with the characteristic are equal in the two groups in the population

H_1: these population proportions are not equal.

2 Collect relevant data from two samples

3 Calculate the value of the test statistic specific to H_0

$$\chi^2 = \frac{(|x-y|-1)^2}{x+y}$$

which follows the Chi-squared distribution with 1 degree of freedom. The 1 in the numerator is a continuity correction (Chapter 19).

4 Compare the value of the test statistic with values from a known probability distribution

Refer χ^2 to Appendix A3.

5 Interpret the *P*-value and results

Interpret the *P*-value and calculate the confidence interval for the difference in the true population proportions. The approximate 95% CI is:

$$\frac{x-y}{m} \pm \frac{1.96}{m}\sqrt{x+y-\frac{(x-y)^2}{m}}$$

McNemar's χ^2-test

McNemar $\chi^2 = (| x - y | - 1)^2 / (x + y)$

In order to compare two methods of establishing the cavity status (present or absent) of teeth, a dentist assessed the condition of 100 first permanent molar teeth that had either tiny or no cavities using radiographic techniques. These results were compared with those obtained using the more objective approach of visually assessing a section of each tooth. The percentages of teeth detected as having cavities by the two methods of assessment were compared using **McNemar's test**.

1 H_0: the two methods of assessment identify the same percentage of teeth with cavities in the population
H_1: these percentages are not equal.

2 The frequencies for the matched pairs are displayed in the table:

	Radiographic diagnosis		
Diagnosis on section	Cavities absent	Cavities present	Total
Cavities absent	45	4	49
Cavities present	17	34	51
Total	62	38	100

3 Test statistic, $\chi^2 = \dfrac{(|17-4|-1)^2}{17+4} = 6.86$

4 We refer χ^2 to Appendix A3 with 1 degree of freedom: $0.001 < P < 0.01$ (computer output gives $P = 0.009$).

5 There is substantial evidence to reject the null hypothesis that the same percentage of teeth are detected as having cavities using the two methods of assessment. The radiographic method has a tendency to fail to detect cavities. We estimate the difference in percentages of teeth detected as having cavities as 51% – 38% = 13%. An approximate confidence interval for the true difference in the percentages is given by 4.4% to 21.6%

$$\left(\text{i.e.} \left\{ \frac{|17-4|}{100} \pm \frac{1.96}{100} \times \sqrt{(17+4) - \frac{(17-4)^2}{100}} \right\} \times 100\% \right).$$

Adapted from Ketley, C.E. & Holt, R.D. (1993) Visual and radiographic diagnosis of occlusal caries in first permanent molars and in second primary molars. *British Dental Journal*, **174**, 364–370.

more than two categories, Chi-squared test

- Individuals can be classified by two factors.
  - For example, one factor may represent disease severity (mild, moderate or severe) and the other factor may represent blood group (A, B, O, AB).
  - We are interested in whether the two factors are associated.
  - Are individuals of a particular blood group likely to be more severely ill?

column

row

	Mild	Moderate	Severe
A			
B			
O			
AB			

Chi-squared test: Assumptions

- The data may be presented in an *r* ×*c* contingency table with *r* rows and *c* columns (Table 25.1).

Table 25.1 Observed frequencies in an $r \times c$ table.

Row categories	Col 1	Col 2	Col 3	...	Col *c*	Total
Row 1	f_{11}	f_{12}	f_{13}	...	f_{1c}	R_1
Row 2	f_{21}	f_{22}	f_{23}	...	f_{2c}	R_2
Row 3	f_{31}	f_{32}	f_{33}	...	f_{3c}	R_3
...	...	...	...	...	...	...
...	...	...	...	...	...	...
Row *r*	f_{r1}	f_{r2}	f_{r3}	...	f_{rc}	R_r
Total	C_1	C_2	C_3	...	C_c	n

- The entries in the table are **frequencies**; each cell contains the number of individuals in a particular row and a particular column.
- Every individual is represented once, and can only belong in one row and in one column, i.e. the categories of each factor are mutually exclusive.
- At least 80% of the expected frequencies are greater than or equal to 5.

Chi-squared test: Rationale

- The null hypothesis, there is no association between the two factors.
- Note that if there are only two rows and two columns, then this test of no association is the same as that of two proportions.
- We calculate the frequency that we expect in each cell of the contingency table if the null hypothesis is true.
- As explained, the expected frequency in a particular cell is the product of the relevant row total and relevant column total, divided by the overall total.
- We calculate a test statistic that focuses on the discrepancy between the observed and expected frequencies in every cell of the table. If the overall discrepancy is large, then it is unlikely the null hypothesis is true.

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

Chi-squared test: Rationale

1 Define the null and alternative hypotheses under study

H_0: there is no association between the categories of one factor and the categories of the other factor in the population

H_1: the two factors are associated in the population.

2 Collect relevant data from a sample of individuals

3 Calculate the value of the test statistic specific to H_0

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

where O and E are the observed and expected frequencies in each cell of the table. The test statistic follows the Chi-squared distribution with degrees of freedom equal to $(r-1)\times(c-1)$.

Because the approximation to the Chi-squared distribution is reasonable if the degrees of freedom are greater than one, we do not need to include a continuity correction (as we did in Chapter 24).

4 Compare the value of the test statistic to values from a known probability distribution

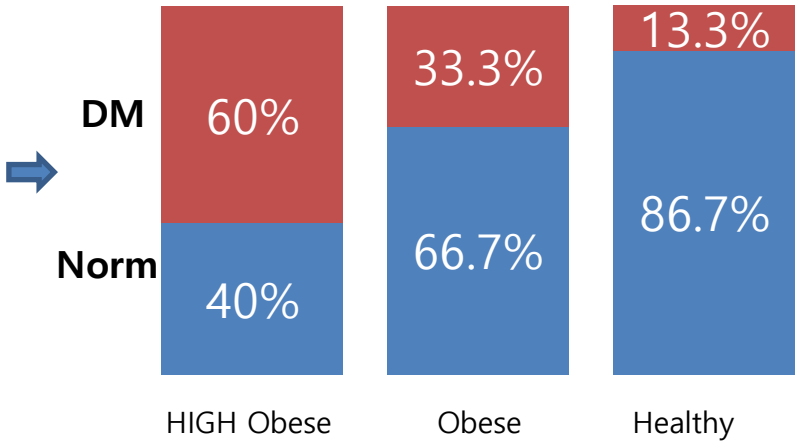
Refer χ^2 to Appendix A3.

5 Interpret the *P*-value and results

Chi-squared test for trend

- we investigate relationships in categorical data when one of the two factors has only two categories (e.g. the presence or absence of a characteristic) and the second factor can be categorized into k, say, mutually exclusive categories that are ordered in some sense
- We can then assess whether there is a trend in the proportions with the characteristic over the categories of the second factor

	HIGH Obese	Obese	Healthy	Total
DM	3 (60.0%)	1 (33.3%)	2 (13.3%)	6 (26.1)
Normal	2 (40.0%)	2 (66.7%)	13 (86.7%)	17 (73.9)
Total	5 (100%)	3 (100%)	15 (100%)	23 (100)



Chi-squared test for trend

1 Define the null and alternative hypotheses under study

H_0: there is no trend in the proportions with the characteristic in the population

H_1: there is a trend in the proportions in the population.

2 Collect relevant data from a sample of individuals

We estimate the proportion with the characteristic in each of the k categories. We assign a score to each of the column categories (Table 25.2). Typically these are the successive values, 1, 2, 3, . . . , k, but, depending on how we have classified the column factor, they could be numbers that in some way suggest the relative values of the ordered categories (e.g. the midpoint of the age range defining each category) or the trend we wish to investigate (e.g. linear or quadratic). The use of any equally spaced numbers (e.g. 1, 2, 3, . . . , k) allows us to investigate a *linear trend*.

3 Calculate the value of the test statistic specific to H_0

$$\chi^2 = \frac{\left(\sum w_i f_{1i} - R_1 \sum \frac{w_i C_i}{n}\right)^2}{\frac{R_1}{n}\left(1 - \frac{R_1}{n}\right)\left(\sum C_i w_i^2 - n\left(\sum \frac{w_i C_i}{n}\right)^2\right)}$$

using the notation of Table 25.2, and where the sums extend over all the k categories. The test statistic follows the Chi-squared distribution with 1 degree of freedom.

4 Compare the value of the test statistic to values from a known probability distribution

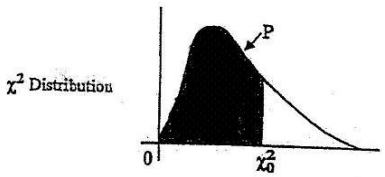
Refer χ^2 to Appendix A3.

5 Interpret the *P*-value and results

Interpret the *P*-value and calculate a confidence interval for each of the k proportions (Chapter 11).

Table 25.2 Observed frequencies and assigned scores in a $2 \times k$ table.

Characteristic	Col 1	Col 2	Col 3	...	Col k	Total
Present	f_{11}	f_{12}	f_{13}	...	f_{1k}	R_1
Absent	f_{21}	f_{22}	f_{23}	...	f_{2k}	R_2
Total	C_1	C_2	C_3	...	C_k	n
Score	w_1	w_2	w_3	...	w_k	



The table below gives the value x_0^2 for which $P[x^2 < x_0^2] =$ P for a given number of degrees of freedom and a given value of P.

Degrees of Freedom	Values of P									
	0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.01	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997

Table A3 Chi-squared distribution.

	Two-tailed *P*-value			
df	0.10	0.05	0.01	0.001
1	2.706	3.841	6.635	10.827
2	4.605	5.991	9.210	13.815
3	6.251	7.815	11.345	16.266
4	7.779	9.488	13.277	18.466
5	9.236	11.070	15.086	20.515
6	10.645	12.592	16.812	22.457
7	12.017	14.067	18.475	24.321
8	13.362	15.507	20.090	26.124
9	14.684	16.919	21.666	27.877
10	15.987	18.307	23.209	29.588
11	17.275	19.675	24.725	31.264
12	18.549	21.026	26.217	32.909
13	19.812	22.362	27.688	34.527
14	21.064	23.685	29.141	36.124
15	22.307	24.996	30.578	37.698
16	23.542	26.296	32.000	39.252
17	24.769	27.587	33.409	40.791
18	25.989	28.869	34.805	42.312
19	27.204	30.144	36.191	43.819
20	28.412	31.410	37.566	45.314
21	29.615	32.671	38.932	46.796
22	30.813	33.924	40.289	48.268
23	32.007	35.172	41.638	49.728
24	33.196	36.415	42.980	51.179
25	34.382	37.652	44.314	52.619
26	35.563	38.885	45.642	54.051
27	36.741	40.113	46.963	55.475
28	37.916	41.337	48.278	56.892
29	39.087	42.557	49.588	58.301
30	40.256	43.773	50.892	59.702
40	51.805	55.758	63.691	73.403
50	63.167	67.505	76.154	86.660
60	74.397	79.082	88.379	99.608
70	85.527	90.531	100.43	112.32
80	96.578	101.88	112.33	124.84
90	107.57	113.15	124.12	137.21
100	118.50	124.34	135.81	149.45

Derived using Microsoft Excel Version 5.0.

Example

Example

A cross-sectional survey was carried out among the elderly population living in Southampton, with the objective of measuring the frequency of cardiovascular disease. A total of 259 individuals, ranging between 65 and 95 years of age, were interviewed. Individuals were grouped into four age groups (65–69, 70–74, 75–79 and 80+ years) at the time of interview. We used the **Chi-squared test** to determine whether the prevalence of chest pain differed in the four age groups.

1 H_0: there is no association between age and chest pain in the population

H_1: there is an association between age and chest pain in the population.

2 The observed frequencies (%) and expected frequencies are shown in the following table.

3 Test statistic, $\chi^2 = \left[\frac{(15-9.7)^2}{9.7} + \ldots + \frac{(41-39.1)^2}{39.1}\right] = 4.839$

4 We refer χ^2 to Appendix A3 with 3 degrees of freedom: $P > 0.10$ (computer output gives $P = 0.18$).

5 There is insufficient evidence to reject the null hypothesis of no association between chest pain and age in the population of elderly people. The estimated proportions (95% confidence intervals) with chest pain for the four successive age groups, starting with the youngest, are: 0.20 (0.11, 0.29), 0.12 (0.04, 0.19), 0.10 (0.02, 0.17) and 0.09 (0.02, 0.21).

	Age (years)				
Chest pain	65–69	70–74	75–79	80+	Total
Yes					
Observed	15 (20.3%)	9 (11.5%)	6 (9.7%)	4 (8.9%)	34
Expected	9.7	10.2	8.1	5.9	
No					
Observed	59 (79.7%)	69 (88.5%)	56 (90.3%)	41 (91.1%)	225
Expected	64.3	67.8	53.9	39.1	
Total	74	78	62	45	259

Example

As the four age groups in this study are ordered, it is also possible to analyse these data using a **Chi-squared test for trend**, which takes into account the ordering of the groups. We may obtain a significant result from this test, even though the general test of association gave a non-significant result. We assign the scores of 1, 2, 3 and 4 to each of the four age groups, respectively, and because of their even spacing, can test for a linear trend.

1 H_0: there is no linear association between age and chest pain in the population

H_1: there is a linear association between age and chest pain in the population.

2 The data are displayed in the previous table. We assign scores of 1, 2, 3 and 4 to the four age groups, respectively.

3 Test statistic is χ^2.

4 We refer χ^2 to Appendix A3 with 1 degree of freedom: $0.05 < P < 0.10$ (computer output gives $P = 0.052$).

5 There is insufficient evidence to reject the null hypothesis of no linear association between chest pain and age in the population of elderly people. However, the P-value is very close to 0.05 and there is a suggestion that the proportion of elderly people with chest pain decreases with increasing age.

$$\chi^2 = \frac{\left\{[(1\times 15)+\ldots+(4\times 4)]-34\times\left[\left(\frac{1\times 74}{259}\right)+\ldots+\left(\frac{4\times 45}{259}\right)\right]\right\}^2}{\frac{34}{259}\times\left(1-\frac{34}{259}\right)\times\left\{[(74\times 1^2)+\ldots+(45\times 4^2)]-259\times\left[\left(\frac{1\times 74}{259}\right)+\ldots+\left(\frac{4\times 45}{259}\right)\right]^2\right\}} = 3.79$$

Adapted from: Dewhurst, G., Wood, D.A., Walker, F., *et al.* (1991) A population survey of cardiovascular disease in elderly people: design, methods and prevalence results. *Age and Ageing* **20**, 353–360.

Simpson's paradox

UC Berkeley 1973 grad school admissions

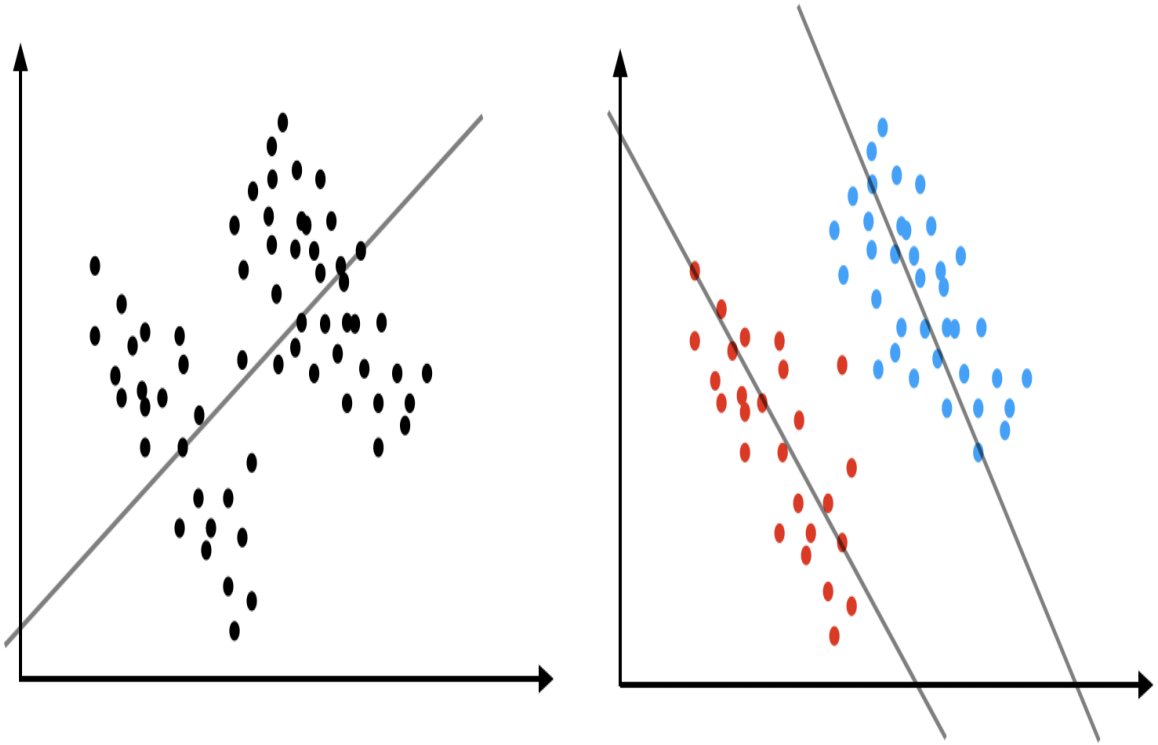
	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	272	6%	341	**7%**

Simpson's paradox

- Simpson's paradox (Yule-Simpson effect)
  - an effect that occurs when the marginal association between two categorical variables is qualitatively different from the partial association between the same two variables after controlling for one or more other variables.
  - A paradox in which a correlation (trend) present in different groups is reversed when the groups are combined.

Simpson's paradox



Relative Risk

Table 15.1. Observed frequencies (see Fig. 15.1)

	Exposed to factor		
	Yes	No	Total
Disease of interest			
Yes	a	b	$a+b$
No	c	d	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

$$\mathbf{RR} = \frac{\text{Estimated risk of disease in exposed group}}{\text{Estimated risk of disease in unexposed group}} = \frac{\mathbf{risk_{exp}}}{\mathbf{risk_{unexp}}} = \frac{\mathbf{a / (a+c)}}{\mathbf{b / (b+d)}}$$

$\mathbf{H_0}$: the risk is the same in the exposed and unexposed groups

$\mathbf{H_1}$: not $\mathbf{H_0}$

Odds ratio

Table 16.1 Observed frequencies (see Fig. 16.1).

	Exposed to factor		
	Yes	No	Total
Disease status			
Case	a	b	$a+b$
Control	c	d	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

$$\text{OR} = \frac{\text{Odds of being a case in exposed group}}{\text{Odds of being a case in unexposed group}} = \frac{a / c}{b / d} = \frac{a \times d}{b \times c}$$

Kappa

Cohen's Kappa

$$K = \frac{observed\ \%\ agreement - chance\ expected\ \%\ agreement}{100 - chance\ expected\ \%\ agreement}$$

- **Interpretation of Kappa**

Value of K	Strength of agreement
< 0.20	Poor
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Good
0.80 – 1.00	Very good

Ref: Altman DG (1991) Practical statistics for medical research. London: Chapman and Hall

Kappa

Cohen's Kappa

Example

Rater A	Rater B		Total
	1	2	
1	35	20	55
2	5	40	45
Total	40	60	100

$$\widehat{\gamma}_\kappa = \frac{p_a - p_e}{1 - p_e}.$$

- Overall agreement probability

$$p_a = \frac{35 + 40}{100} = 0.75$$

- Chance agreement probability

$$p_e = \frac{55}{100} \times \frac{40}{100} + \frac{45}{100} \times \frac{60}{100} = \frac{49}{100} = 0.49.$$

$$\widehat{\gamma}_\kappa = \frac{0.75 - 0.49}{1 - 0.49} \cong 0.51.$$

Ref: Gwet, K. (2010). "Handbook of Inter-Rater Reliability (Second Edition)

Kappa

inter-observer agreement

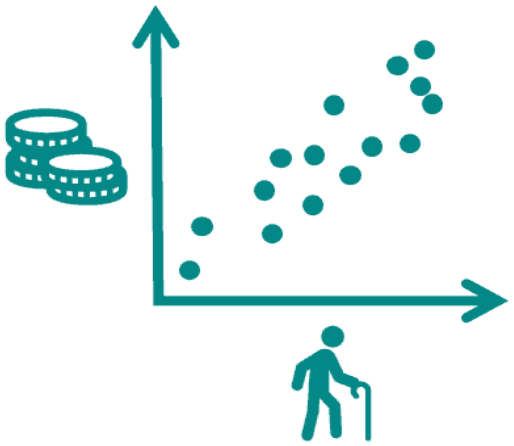
CONSISTENCY OF Readings				
		A expert		
		Tuberculosis	normal	normal
B expert	Tuberculosis	136	92	228
	normal	69	240	309
	TOTAL	205	332	537

Observed agreement = (136 + 240)/537 = 70.0%

Kappa (κ)= 0.378 (0.298~0.459)

Correlation

- Correlation analysis is concerned with measuring the degree of association between two variables, *x* and *y*.
- Initially, we assume that both *x* and *y* are **numerical**, e.g. height and weight.
- Suppose we have a pair of values, (*x*, *y*), measured on each of the *n* individuals in our sample. We can mark the point corresponding to each individual's pair of values on a two-dimensional **scatter diagram.**
- Plotting the points for all *n* individuals, we obtain a scatter of points that may suggest a relationship between the two variables.



Pearson correlation coefficient

- We say that we have a **linear relationship** between *x* and *y* if a straight line drawn through the midst of the points provides the most appropriate approximation to the observed relationship.
- We measure how close the observations are to the straight line that best describes their linear relationship by calculating the **Pearson product moment correlation coefficient**, usually simply called the **correlation coefficient**. Its true value in the *population*, ρ (the Greek letter, rho), is estimated in the *sample* by *r*, where

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

- which is usually obtained from computer output.

Properties

- *r* ranges from -1 to +1.
- Its **sign** indicates whether one variable increases as the other variable increases (positive *r*) or whether one variable decreases as the other increases (negative *r*).
- Its **magnitude** indicates how close the points are to the straight line.
  - In particular if *r* = +1 or -1, then there is perfect correlation with all the points lying on the line (this is most unusual, in practice); if *r* = 0, then there is no **linear** correlation (although there may be a non-linear relationship).

    The closer *r* is to the extremes, the greater the degree of linear association.
- It is dimensionless, i.e. it has no units of measurement.
- Its value is valid only within the range of values of *x* and *y* in the sample.
  - Its absolute value (ignoring sign) tends to increase as the range of values of *x* and/or *y* increases and therefore you cannot infer that it will have the same value when considering values of *x* or *y* that are more extreme than the sample values.
- *x* and *y* can be interchanged without affecting the value of *r*.
- A correlation between *x* and *y* does not necessarily imply a 'cause and effect' relationship.
- r^2 represents the proportion of the variability of *y* that can be attributed to its linear relationship with *x*.

Properties

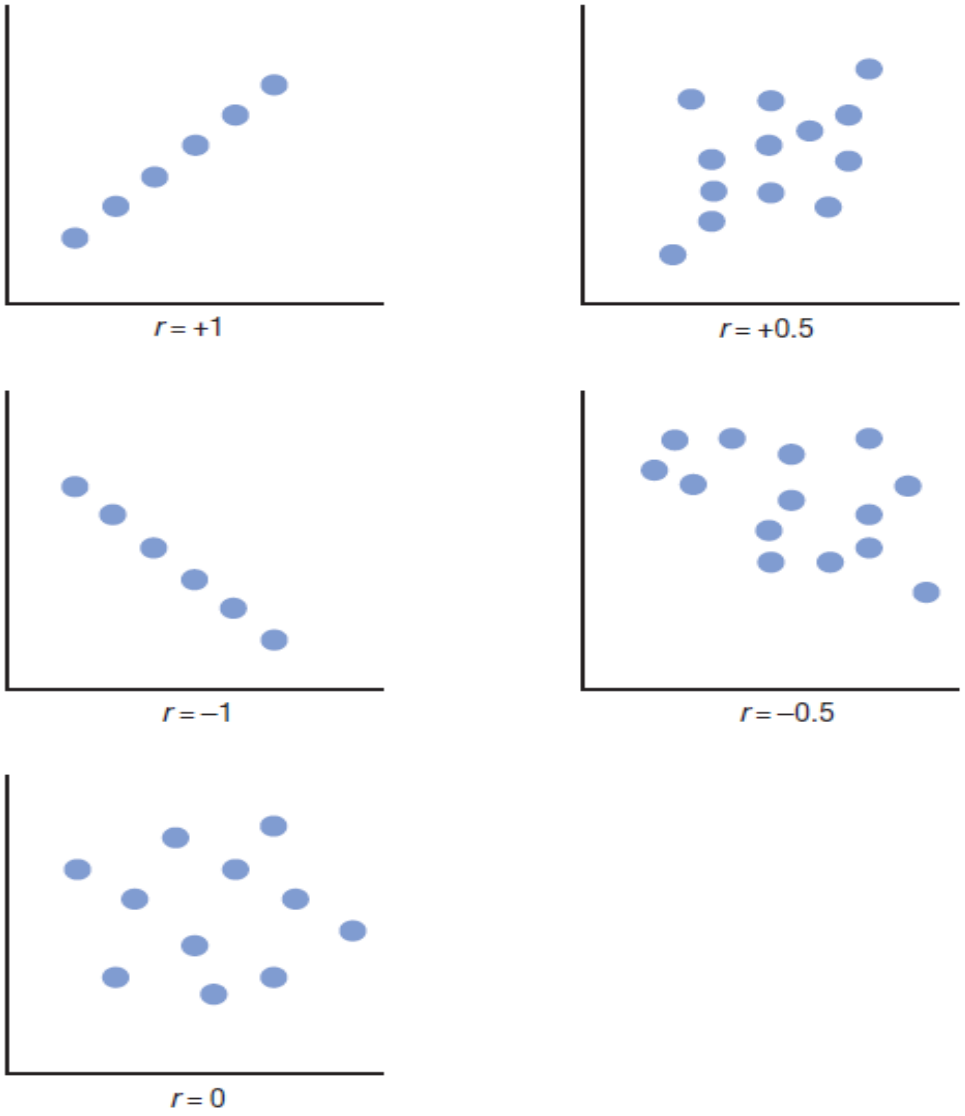


Figure 26.1 Five diagrams indicating values of r in different situations.

When not to calculate r

- It may be misleading to calculate *r* when:
  - there is a non-linear relationship between the two variables (Fig. 26.2a), e.g. a quadratic relationship;
  - the data include more than one observation on each individual;
  - one or more outliers are present (Fig. 26.2b);
  - the data comprise subgroups of individuals for which the mean levels of the observations on at least one of the variables are different (Fig. 26.2c);

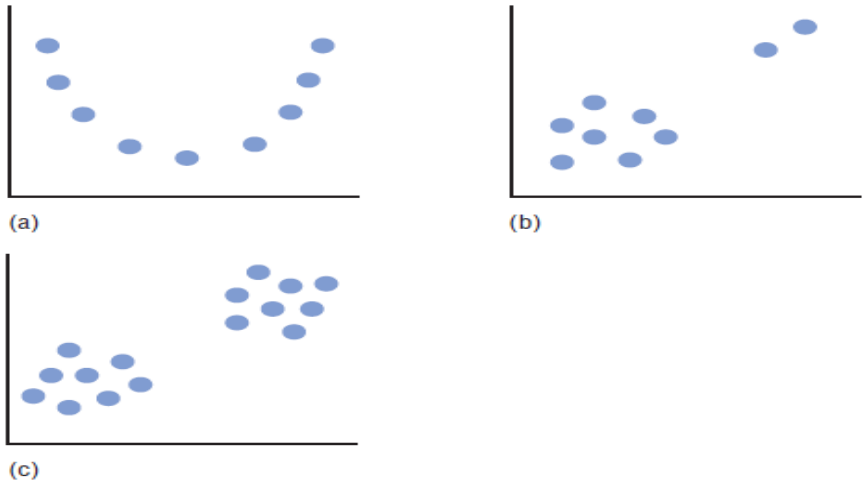


Figure 26.2 Diagrams showing when it is inappropriate to calculate the correlation coefficient. (a) Relationship not linear, $r = 0$. (b) In the presence of outlier(s). (c) Data comprise subgroups.

Hypothesis test for the Pearson correlation coefficient

- We want to know if there is any linear correlation between two numerical variables.
- Our sample consists of *n* independent pairs of values of *x* and *y*.
- We assume that at least one of the two variables is Normally distributed.

1 Define the null and alternative hypotheses under study

$H_0: \rho = 0$

$H_1: \rho \neq 0$

2 Collect relevant data from a sample of individuals

3 Calculate the value of the test statistic specific to H_0

Calculate r.

- If $n \leq 150$, r is the test statistic
- If $n > 150$, calculate $T = \sqrt{\frac{(n-2)}{(1-r^2)}}$

which follows a t-distribution with $n-2$ degrees of freedom.

4 Compare the value of the test statistic to values from a known probability distribution

- If $n \leq 150$, refer r to Appendix A10
- If $n > 150$, refer T to Appendix A2.

5 Interpret the *P*-value and results

Calculate a confidence interval for ρ. Provided *both variables are approximately Normally distributed*, the approximate 95% confidence interval for ρ is:

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

where $z_1 = z - \frac{1.96}{\sqrt{n-3}}$, $z_2 = z + \frac{1.96}{\sqrt{n-3}}$,

and $z = 0.5\log_e\left[\frac{(1+r)}{(1-r)}\right]$.

Note that, if the sample size is large, H_0 may be rejected even if r is quite close to zero. Alternatively, even if r is large, H_0 may not be rejected if the sample size is small. For this reason, it is particularly helpful to calculate r^2, the proportion of the total variance of one variable explained by its linear relationship with the other. For example, if $r = 0.40$ then $P < 0.05$ for a sample size of 25, but the relationship is only explaining 16% ($= 0.40^2 \times 100$) of the variability of one variable.

Table A2 *t*-distribution.

df	Two-tailed *P*-value			
	0.10	0.05	0.01	0.001
1	6.314	12.706	63.656	636.58
2	2.920	4.303	9.925	31.600
3	2.353	3.182	5.841	12.924
4	2.132	2.776	4.604	8.610
5	2.015	2.571	4.032	6.869
6	1.943	2.447	3.707	5.959
7	1.895	2.365	3.499	5.408
8	1.860	2.306	3.355	5.041
9	1.833	2.262	3.250	4.781
10	1.812	2.228	3.169	4.587
11	1.796	2.201	3.106	4.437
12	1.782	2.179	3.055	4.318
13	1.771	2.160	3.012	4.221
14	1.761	2.145	2.977	4.140
15	1.753	2.131	2.947	4.073
16	1.746	2.120	2.921	4.015
17	1.740	2.110	2.898	3.965
18	1.734	2.101	2.878	3.922
19	1.729	2.093	2.861	3.883
20	1.725	2.086	2.845	3.850
21	1.721	2.080	2.831	3.819
22	1.717	2.074	2.819	3.792
23	1.714	2.069	2.807	3.768
24	1.711	2.064	2.797	3.745
25	1.708	2.060	2.787	3.725
26	1.706	2.056	2.779	3.707
27	1.703	2.052	2.771	3.689
28	1.701	2.048	2.763	3.674
29	1.699	2.045	2.756	3.660
30	1.697	2.042	2.750	3.646
40	1.684	2.021	2.704	3.551
50	1.676	2.009	2.678	3.496
100	1.660	1.984	2.626	3.390
200	1.653	1.972	2.601	3.340
5000	1.645	1.960	2.577	3.293

Derived using Microsoft Excel Version 5.0.

Table A10 Pearson's correlation coefficient.

Sample size	Two-tailed *P*-value		
	0.05	0.01	0.001
5	0.878	0.959	0.991
6	0.811	0.917	0.974
7	0.755	0.875	0.951
8	0.707	0.834	0.925
9	0.666	0.798	0.898
10	0.632	0.765	0.872
11	0.602	0.735	0.847
12	0.576	0.708	0.823
13	0.553	0.684	0.801
14	0.532	0.661	0.780
15	0.514	0.641	0.760
16	0.497	0.623	0.742
17	0.482	0.606	0.725
18	0.468	0.590	0.708
19	0.456	0.575	0.693
20	0.444	0.561	0.679
21	0.433	0.549	0.665
22	0.423	0.537	0.652
23	0.413	0.526	0.640
24	0.404	0.515	0.629
25	0.396	0.505	0.618
26	0.388	0.496	0.607
27	0.381	0.487	0.597
28	0.374	0.479	0.588
29	0.367	0.471	0.579
30	0.361	0.463	0.570
35	0.334	0.430	0.532
40	0.312	0.403	0.501
45	0.294	0.380	0.474
50	0.279	0.361	0.451
55	0.266	0.345	0.432
60	0.254	0.330	0.414
70	0.235	0.306	0.385
80	0.220	0.286	0.361
90	0.207	0.270	0.341
100	0.217	0.283	0.357
150	0.160	0.210	0.266

Extracted with permission from Diem, K. (1970) *Documenta Geigy Scientific Tables*, 7th edn, Blackwell Publishing, Oxford.

Pearson correlation coefficient

Spearman (rank) correlation coefficient

When two variables X and Y are normally distributed

When two variables X and Y are **not normally distributed**

Spearman's rank correlation coefficient

- We calculate **Spearman's rank correlation coefficient**, the nonparametric equivalent to Pearson's correlation coefficient, if one or more of the following points is true:
  - at least one of the variables, *x* or *y*, is measured on an ordinal scale;
  - neither *x* nor *y* is Normally distributed;
  - the sample size is small;
  - we require a measure of the association between two variables when their relationship is non-linear.

Example

As part of a study to investigate the factors associated with changes in blood pressure in children, information was collected on demographic and lifestyle factors, and clinical and anthropometric measures in 4245 children aged from 5 to 7 years. The relationship between height (cm) and systolic blood pressure (mmHg) in a sample of 100 of these children is shown in the scatter diagram (Fig. 28.1); there is a tendency for taller children in the sample to have higher blood pressures. **Pearson's correlation coefficient** between these two variables was investigated. Appendix C contains a computer output from the analysis.

1 H_0: the population value of the Pearson correlation coefficient, ρ, is zero (there is no linear relationship)

H_1: the population value of the Pearson correlation coefficient is not zero. (there is a linear relationship)

2 We can show (Fig. 37.1) that the sample values of both height and systolic blood pressure are approximately Normally distributed.

3 We calculate r as 0.33. This is the test statistic since $n \leq 150$.

4 We refer r to Appendix A10 with a sample size of 100: $P < 0.001$.

5 There is strong evidence to reject the null hypothesis; we conclude that there is a linear relationship between systolic blood pressure and height in the population of such children. However, $r^2 = 0.33 \times 0.33 = 0.11$. Therefore, despite the highly significant result, the relationship between height and systolic blood pressure explains only a small percentage, 11%, of the variation in systolic blood pressure.

In order to determine the 95% confidence interval for the true correlation coefficient, we calculate:

$$z = 0.5 \ln\left(\frac{1.33}{0.67}\right) = 0.3428$$

$$z_1 = 0.3428 - \frac{1.96}{9.849} = 0.1438$$

$$z_2 = 0.3428 + \frac{1.96}{9.849} = 0.5418$$

Thus the confidence interval ranges from

$$\frac{(e^{2\times 0.1438} - 1)}{(e^{2\times 0.1438} + 1)} \text{ to } \frac{(e^{2\times 0.5418} - 1)}{(e^{2\times 0.5418} + 1)}, \text{ i.e. from } \frac{0.33}{2.33} \text{ to } \frac{1.96}{3.96}.$$

We are thus 95% certain that ρ lies between 0.14 and 0.49.

26, 28 및 29장에서 사용된 인체 측정 자료에 대한 분석 (SAS 분석결과)

OBS	SBP	Height	Weight	Sex
1	91.00	119.7	20.0	0
2	122.50	124.6	42.5	0
3	109.50	111.3	19.8	0
4	100.50	110.3	18.9	0
5	99.00	112.5	19.0	0
6	103.50	115.1	19.3	0
7	101.00	116.3	19.6	0
8	103.00	111.1	17.1	1
9	106.50	117.2	20.7	1
10	102.50	113.2	22.1	1

첫 10명의 아이들에 대한 출력 자료

Correlation Analysis

4 'VAR' Variables: SBP Height Weight Age

Simple Statistics

Variable	N	Mean	Std Dev	Sum
SBP	100	104.414700	9.430933	10441
Height	100	120.054000	6.439986	12005
Weight	100	22.826000	4.223303	2282.600000
Age	100	6.696900	0.731717	669.690000

Simple Statistics

Variable	Minimum	Maximum
SBP	81.500000	128.850000
Height	107.1000000	136.800000
Weight	15.900000	42.500000
Age	5.130000	8.840000

각 변수에 대한 요약 통계량

Pearson Correlation Coefficients/Prob> |R| under Ho:Rho=0
/N=100

	SBP	Height	Weight	Age
SBP	1.00000	0.33066	0.51774	0.16373
	0.0	0.0008	0.0001	0.1036
Height	0.33066	1.00000	0.69151	0.64486
	0.0008	0.0	0.0001	0.0001
Weight	0.51774	0.69151	1.00000	0.38935
	0.0001	0.0001	0.0	0.0001
Age	0.16373	0.64486	0.38935	1.00000
	0.1036	0.0001	0.0001	0.0

SBP와 Age 간 Pearson의 상관계수

이에 해당하는 P-value

Spearman Correlation Coefficients/Prob> |R| under Ho:Rho=0
/N=100

	SBP	Height	Weight	Age
SBP	1.00000	0.31519	0.45453	0.14778
	0.0	0.0014	0.0001	0.1423
Height	0.31519	1.00000	0.82298	0.61491
	0.0014	0.0	0.0001	0.0001
Weight	0.45453	0.82298	1.00000	0.51260
	0.0001	0.0001	0.0	0.0001
Age	0.14778	0.61491	0.51260	1.00000
	0.1423	0.0001	0.0001	0.0

Hight와 Age 간 Spearman의 상관계수

P-value

26장

R

R is an open source programming language and soft ware environment for statistical computing and gra phics.

The R language is widely used among statisticians a nd data miners for developing statistical software an d data analytics tools

History of R

- **Modelled after S & S-plus, developed at AT&T labs in late 1 980s.**
- **R project was started by Robert Gentleman and Ross Ihaka Department of Statistics, University of Auckland (1995).**
- **Currently maintained by R core development team – an inte rnational team of volunteer developers (since 1997).**

Download R and RStudio

Download R :

http:/www.r-project.org/

Download RStudio :

http:// posit.co/download/rstudio-desktop /

aov

```
aov(formula, data = , ...)
```

* **oneway.test or lm ( )**

```
TukeyHSD(x, conf.level = 0.95, ...)
```

```
kruskal.test(formula, data)
```

```
dunnTest(formula, data, method = "bonferroni")
```

Data : chickwts (N=71, variable: weight, feed)

H_0 : $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_6$

H_1 : not H_0

Exercise

Data: stroke_CI.xls

1. Is the average body weight of all study participants greater than 60 kg?
(one-sample t-test: t.test)

2. Is there a difference in body weight between CI patients (200 patients) and the normal group (200 patients)?
(two-sample t-test :t.test)

3. Among CI patients, subjects were divided into three groups based on drinking status.
Is there a difference in Hct levels among the three groups?
"drinking" 1: Non-drinker 2: Moderate 3: Heavy
(anova, multiple comparison: aov, TukeyHSD)

#HW2

KIOM
KOREA INSTITUTE OF ORIENTAL MEDICINE

Data: stroke_CI.xls

1. Is there a difference in WBC values between the CI patients (200 patients) and the normal group (200 patients)??

2. Among CI patients, Hct (HDL) levels were investigated by dividing them into three groups according to whether they smoked or not (1: Non-smoker 2: Former smoker 3: Current smoker). Are there differences between Hct (HDL) values for the three groups?

chisq.test

chisq.test(x, y = NULL, correct = TRUE,...)

*** prop.test (x, n, p=NULL)**

Arguments

x	a numeric vector or matrix. x and y can also both be factors.
y	a numeric vector; ignored if x is a matrix. If x is a factor, y should be a factor of the same length.
correct	a logical indicating whether to apply continuity correction when computing the test statistic for 2 by 2 tables: one half is subtracted from all |O - E| differences; however, the correction will not be bigger than the differences themselves. No correction is done if simulate.p.value = TRUE.
p	a vector of probabilities of the same length of x. An error is given if any entry of p is negative.
rescale.p	a logical scalar; if TRUE then p is rescaled (if necessary) to sum to 1. If rescale.p is FALSE, and p does not sum to 1, an error is given.
simulate.p.value	a logical indicating whether to compute p-values by Monte Carlo simulation.
B	an integer specifying the number of replicates used in the Monte Carlo test.

* correct=TRUE Yate's continuity correction

fisher.test

fisher.test(x, y = NULL, conf.int = TRUE,...)

Arguments

x	either a two-dimensional contingency table in matrix form, or a factor object.
y	a factor object; ignored if x is a matrix.
workspace	an integer specifying the size of the workspace used in the network algorithm. In units of 4 bytes. Only used for non-simulated p-values larger than *2 by 2* tables. Since R version 3.5.0, this also increases the internal stack size which allows larger problems to be solved, however sometimes needing hours. In such cases, `simulate.p.values=TRUE` may be more reasonable.
hybrid	a logical. Only used for larger than *2 by 2* tables, in which cases it indicates whether the exact probabilities (default) or a hybrid approximation thereof should be computed.
hybridPars	a numeric vector of length 3, by default describing "Cochran's conditions" for the validity of the chisquare approximation, see 'Details'.
control	a list with named components for low level algorithm control. At present the only one used is `"mult"`, a positive integer ≥ *2* with default 30 used only for larger than *2 by 2* tables. This says how many times as much space should be allocated to paths as to keys: see file '`fexact.c`' in the sources of this package.
or	the hypothesized odds ratio. Only used in the *2 by 2* case.
alternative	indicates the alternative hypothesis and must be one of `"two.sided"`, `"greater"` or `"less"`. You can specify just the initial letter. Only used in the *2 by 2* case.
conf.int	logical indicating if a confidence interval for the odds ratio in a *2 by 2* table should be computed (and returned).
conf.level	confidence level for the returned confidence interval. Only used in the *2 by 2* case and if `conf.int = TRUE`.
simulate.p.value	a logical indicating whether to compute p-values by Monte Carlo simulation, in larger than *2 by 2* tables.
B	an integer specifying the number of replicates used in the Monte Carlo test.

mcnemar.test

KIOM
KOREA INSTITUTE OF ORIENTAL MEDICINE

```
mcnemar.test(x, y = NULL, correct = TRUE)
```

Arguments

`x`	either a two-dimensional contingency table in matrix form, or a factor object.
`y`	a factor object; ignored if `x` is a matrix.
`correct`	a logical indicating whether to apply continuity correction when computing the test statistic.

Arthritis: Arthritis Treatment Data

Data from Koch \& Edwards (1988) from a double-blind clinical trial investigating a new treatment for rheumatoid arthritis.

- A data frame with 84 observations and 5 variables.

-ID: patient ID
-Treatment : factor indicating treatment (Placebo, Treated)
-Sex : factor indicating sex (Female, Male)
-Age : age of patient.
-Improved : ordered factor indicating treatment outcome (None, Some, Marked)

KIOM
KOREA INSTITUTE OF ORIENTAL MEDICINE

Thank you!

kmm1836@kiom.re.kr