University of York

Department of Health Sciences M.Sc. programme

Reporting statistical analyses

What you should know from the statistics course

- Frequencies and frequency distributions
- Percentages
- Histograms
- Means and standard deviations
- Standard errors
- Confidence intervals
- Significance tests
- t tests for means
- Chi-squared tests
- Correlation coefficients
- How to do these using SPSS
- What they mean

Structure of a statistical report

Reports are much easier to read if they have structure, with headings and subheadings. One way of structuring a report of a statistical analysis is to make it follow the plan of a scientific paper:

- Introduction
- Methods
- Results
- Discussion
- Conclusions

It gives the statistical parts of these:

- Introduction: the questions to be answered and the data available.
- Methods: the statistical methods to be used and why they have been chosen.
- Results: what has been found.
- Discussion: any limitation of these analyses.
- Conclusions: what we can conclude from these analyses.

This is the method which I would recommend and which I use myself.

What should go into the report? Look at published research papers. They do not contain lots of computer printout. They contain the results of the analysis, extracted from the computer printout.

What analysis to do when

Two types of variable: continuous measurements and categorical classifications

Analysis	Continuous	Categorical
Descriptive	Histogram, mean, standard deviation, median, range	Frequencies and percentages
Single group	Confidence interval for mean	Confidence interval for proportion
Changes in one group	Paired t method, large sample Normal method	McNemar's test*
Compare two groups	Two sample t method, large sample Normal method	Chi-squared test, odds ratio or relative risk
Relationship between two variables	Scatter diagram, correlation coefficient, linear regression*	Cross-tabulation, chi-squared test

* not included in this course

An example

The data file contains data from a case control study of stroke, a group of stroke patients and a group of unmatched controls. It contains the following variables:

case	Case control status	case=1, control=2
age	Age in years	
sex	Sex	female=1 male=2
chol	Serum cholesterol in mmol/L	
evsmok	Ever smoked	no=1 yes=2

Questions about these data

1. Is there anything to suggest that cases and controls were not comparable in terms of age and sex?
2. Do cases differ in cholesterol or smoking history?
3. Could age or sex differences have any affect on the relationships between stroke and cholesterol or smoking history?

Introduction

The data consist of five variables measured on a group of stroke patients and controls who have not had strokes. First the data will be described and any errors checked for. The age and sex distributions of cases and controls will be compares. The mean cholesterol levels and the smoking history will be compared between cases and controls. We then consider possible effects of age or sex differences between cases and controls on any cholesterol and smoking differences found.

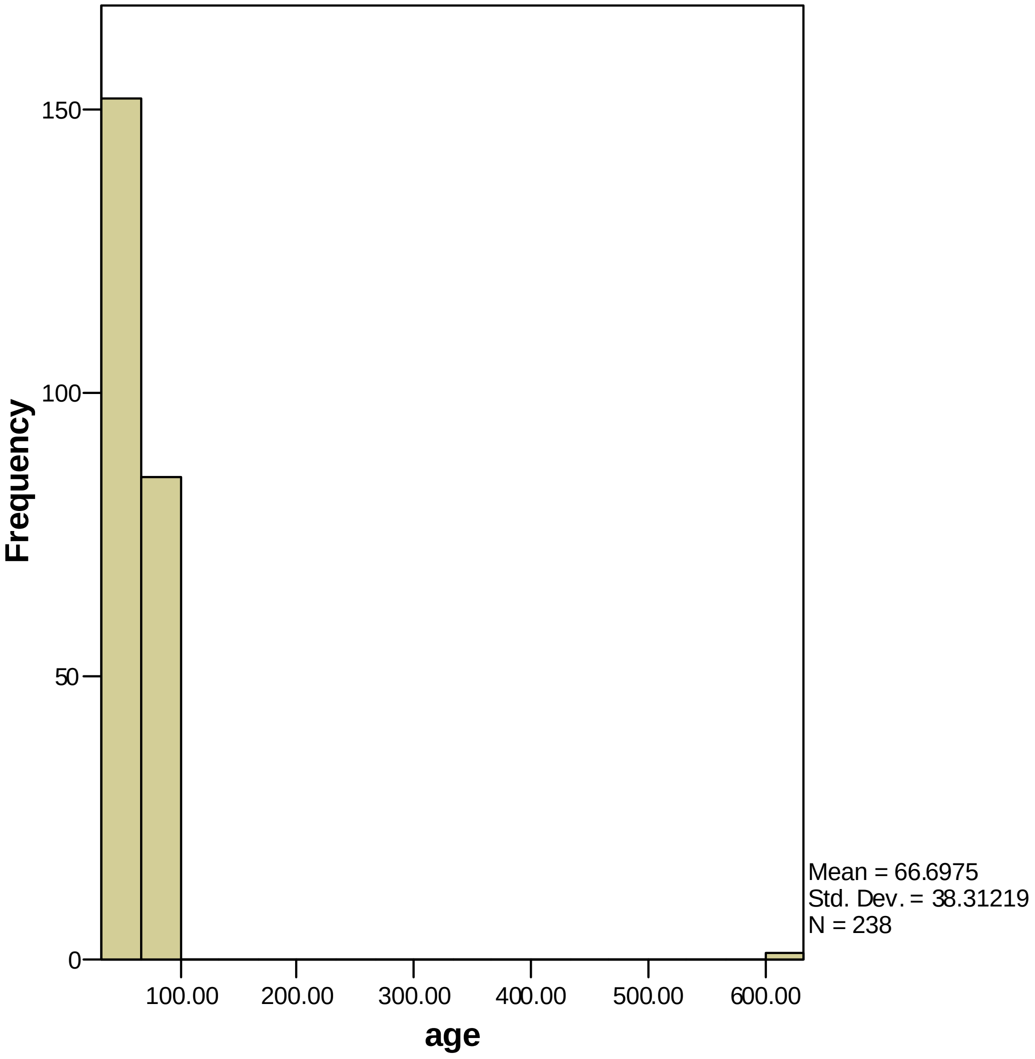
Methods

The distributions of the continuous variables, age and serum cholesterol, will be examined using histograms. The distributions of the categorical variable, case control status, sex, and smoking history, will be examined by tabulation. Any observations which appear to be mistakes will be identified and we will decide what to do about them. The mean age will be compared between the groups using a large sample Normal comparison of means, because this is a continuous variable and there are more than 100 observations in each group. The sex distribution will be compared using a chi-squared test, because this is a categorical variable. Mean cholesterol levels will be compared between the groups using a large sample Normal comparison of means and the smoking history will be compared using a chi-squared test, because cholesterol is a continuous variable and smoking history is categorical.

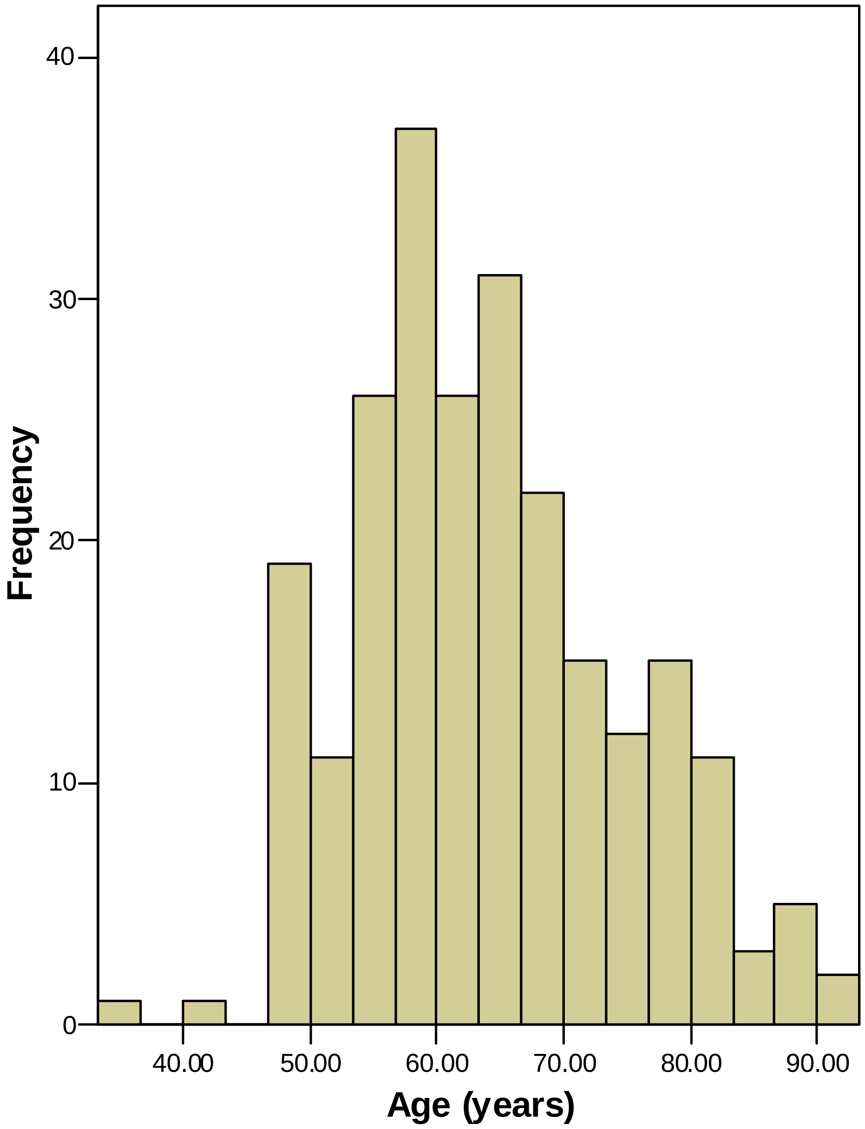
Results

Checking for impossible values (N.B. Use headings and subheadings)

The histogram for age is as follows:

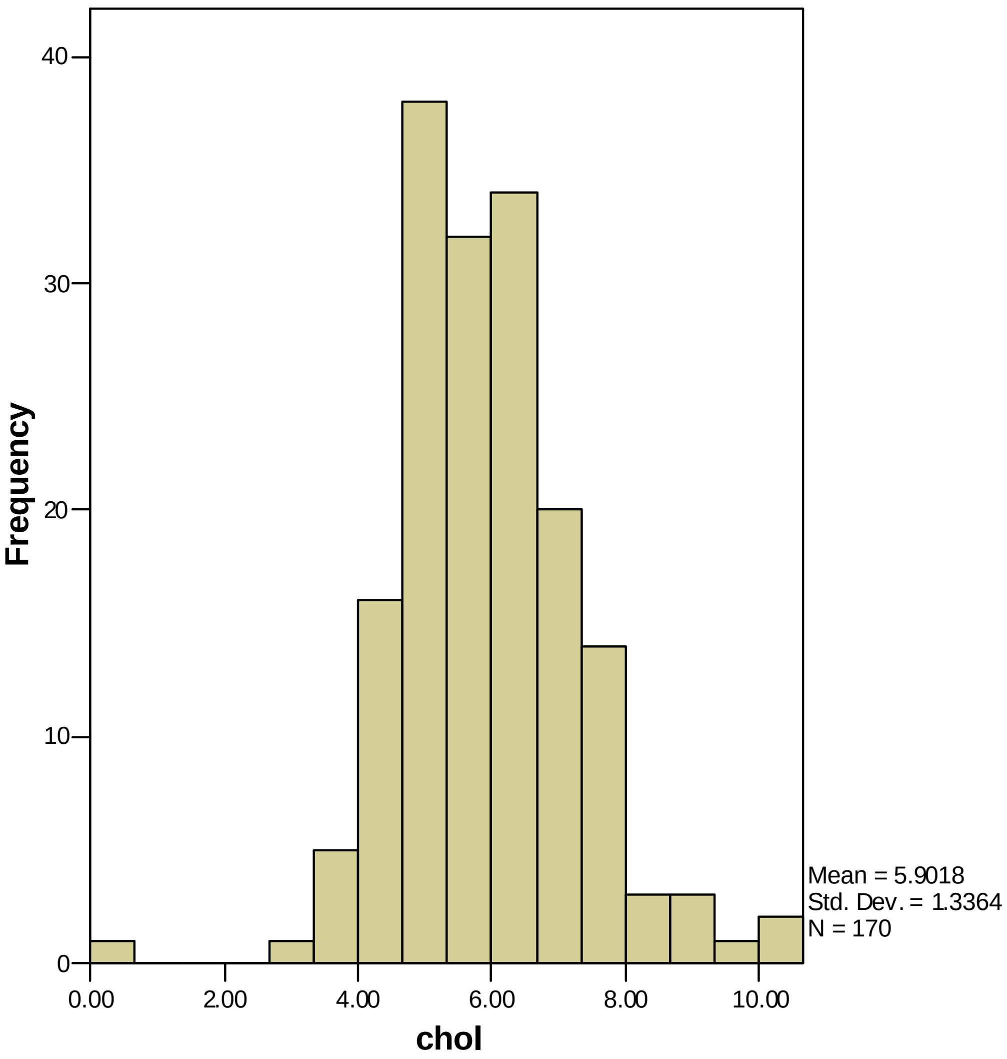


There is an impossible age: 633 years. If we could, we would check this against the original records, but as we are unable to do this, we will set it to missing. There are no units on the label for the horizontal axis. We can correct this by adding a label for the variable in the variable view: “Age (years)”. Another improvement to the presentation is to remove the side legend giving mean and standard deviation. These are usually given to too many decimal places and this information is better presented elsewhere. The graph now looks much better.

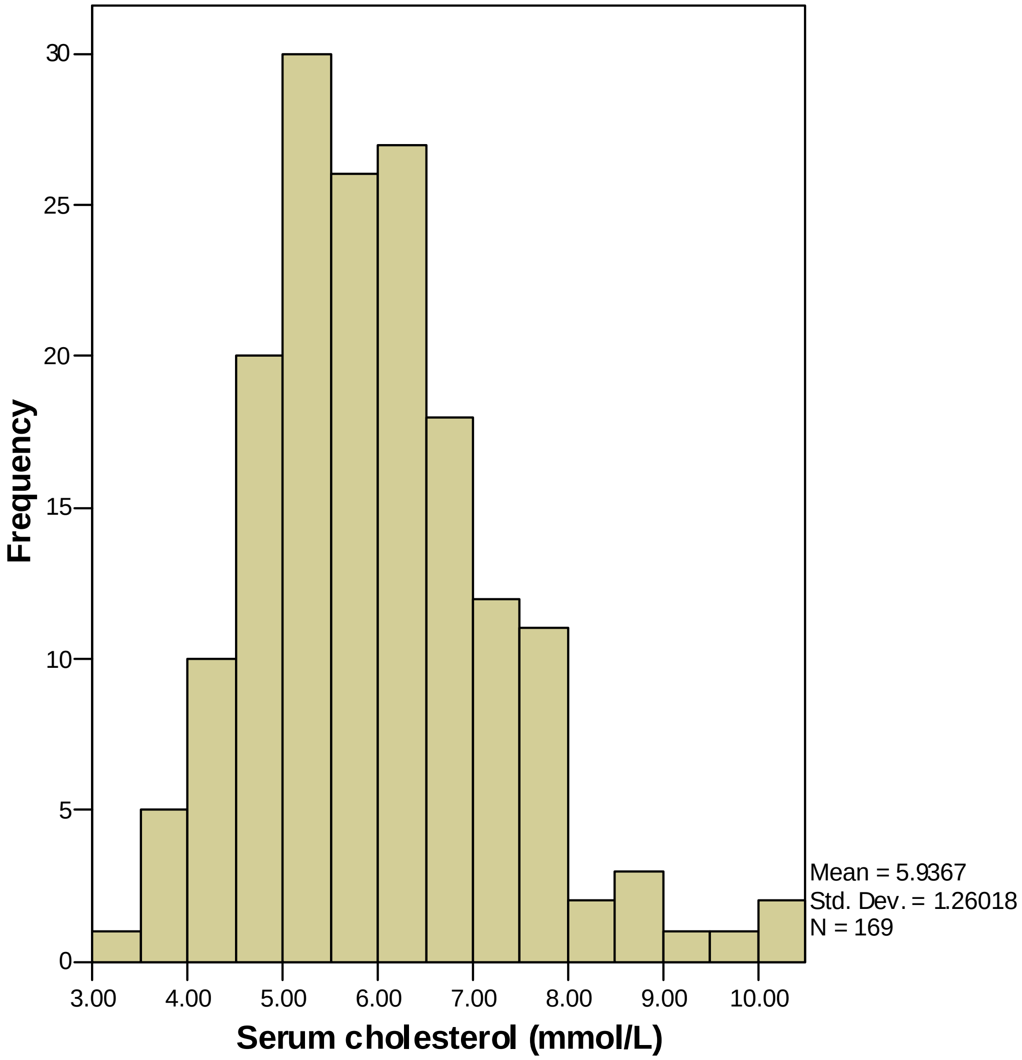


The distribution has a positively skew shape with mean age 64.3 years and standard deviation 10.5 years. (Note that age is recorded in whole years, so we do need more than one decimal place for mean and standard deviation. Just because the computer prints them out, you do not have to use them. It is a good idea to give units when you quote means and standard deviations.)

We do the same for cholesterol:



We have one obvious mistake, an observation equal to zero, which we should set to missing. We should also label the variable properly, including units. Also, there are only 170 of the 238 observations present, the others are missing. A lot of cholesterol measurements are missing.



For the report, it would be better to remove the mean and standard deviation from the graph and give them in the text instead: “Serum cholesterol has a positively skew distribution with mean 5.9 mmol/L and standard deviation 1.3 mmol/L.” There are serum cholesterol measurements for only 169 of the 238 subjects, so there are quite a high proportion of observations missing, 69/238 = 29%. We should mention this as it will be an important limitation which we should mention in the discussion.

The tables for the three categorical variables produces the following SPSS output:

Frequencies

Statistics

		case	evsmok	sex
N	Valid	238	238	238
	Missing	0	0	0

Frequency Table

case

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	101	42.4	42.4	42.4
	2.00	137	57.6	57.6	100.0
	Total	238	100.0	100.0	

evsmok

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	1	.4	.4	.4
	1.00	130	54.6	54.6	55.0
	2.00	107	45.0	45.0	100.0
	Total	238	100.0	100.0	

sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	110	46.2	46.2	46.2
	2.00	128	53.8	53.8	100.0
	Total	238	100.0	100.0	

You do not see this sort of thing in published papers. We need only the frequencies and the percentages and we should label the categories. We could do this either in SPSS using values in the variables view, or in the word processed report. We could give something like this:

There were no missing values for case control status, sex or smoking history. The proportions in each category were:

Case control status	Frequency
Case	101 (42.4%)
Control	137 (57.6%)

Sex	Frequency
Female	110 (46.2%)
Male	128 (53.8%)

Smoking history	Frequency
0	1 (0.4%)
Never smoked	130 (54.6%)
Has smoked	107 (45.0%)

You could improve the report by having table numbers and legends for these tables and refer to them in the text as you would in a paper, e.g. "Table 2 shows the numbers of male and female subjects."

Table 2. Numbers of male and female subjects

Sex	Frequency
Female	110 (46.2%)
Male	128 (53.8%)

I haven’t done this in these notes for reasons of time and space.

Smoking history has one impossible code, which we shall set to missing:

evsmok

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	130	54.6	54.9	54.9
	2.00	107	45.0	45.1	100.0
	Total	237	99.6	100.0	
Missing	System	1	.4		
Total		238	100.0		

The percentage we want is now the valid percentage:

Smoking history	Frequency
Never smoked	130 (54.9%)
Has smoked	107 (45.1%)

Comparing age and sex distributions between cases and controls.

The SPSS output is

T-Test

Group Statistics

	case	N	Mean	Std. Deviation	Std. Error Mean
Age (years)	1.00	100	64.9100	9.08111	.90811
	2.00	137	63.8686	11.37452	.97179

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Age (years)	Equal variances assumed	7.247	.008	.756	235	.450	1.04139	1.37706	-1.67156	3.75434
	Equal variances not assumed			.783	233.074	.434	1.04139	1.33005	-1.57908	3.66185

Do not put this chunk of printout into the report. You do not see this sort of thing in papers, it has information we don't need, and it has far too many decimal places. We might quote the mean and standard deviation in each group: "The mean (standard deviation) of age was 64.9 (9.1) years for cases and 63.9 (11.4) years for controls, or give a reduced version of the table:

Table: Age (years) of cases and controls

Group	Number	Mean	Std. Deviation
Cases	100	64.9	9.1
Controls	137	63.9	11.4

Because we are using the large sample Normal test, we need the comparison of means when equal variances are not assumed. We can say the difference in mean age (cases minus controls) is 1.0 years (SE=1.3, 95% confidence interval –1.6 to 3.7 years, P=0.4). Hence there is no evidence that the groups differ in mean age. The standard deviation is larger in the control group and the Levene test shows that this is significant (P=0.008), so the controls vary more in age than do the cases.

For sex, we cross-tabulate case control status by sex. As is conventional, I have put case control status as the row variable. I have also asked for row percentages in the cells, because this gives me the percentage of cases who are female. The percentage of females who are cases would be meaningless.

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
case * sex	238	100.0%	0	.0%	238	100.0%

case * sex Crosstabulation

			sex		Total
			1.00	2.00	
case	1.00	Count	32	69	101
		% within case	31.7%	68.3%	100.0%
	2.00	Count	78	59	137
		% within case	56.9%	43.1%	100.0%
Total		Count	110	128	238
		% within case	46.2%	53.8%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	14.913(b)	1	.000		
Continuity Correction(a)	13.915	1	.000		
Likelihood Ratio	15.156	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	14.851	1	.000		
N of Valid Cases	238				

a Computed only for a 2x2 table
b 0 cells (.0%) have expected count less than 5. The minimum expected count is 46.68.

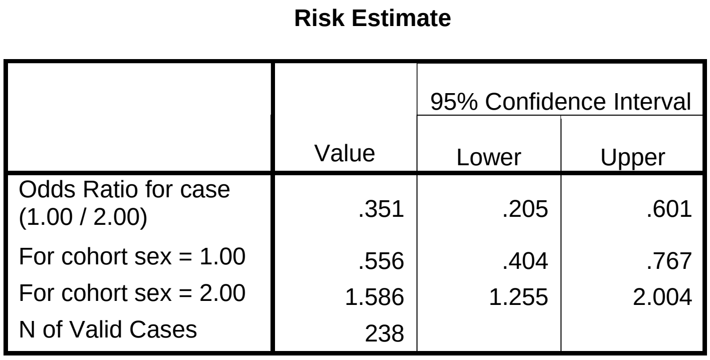
As before, we do not want this chunk of printout in the report. We could use a version of the cross-tabulation

Patient group	Sex		Total
	Female	Male	
Cases	32 (31.7%)	69 (68.3%)	101 (100%)
Controls	78 (56.9%)	59 (43.1%)	137 (100%)

Note that I have omitted the column totals, because in this case control study they do not mean anything. Cases and controls were sampled as two separate groups and there is a far greater proportion of cases in this sample than there would be in a general sample from the adult population. The same would be true in a clinical trial.

We can quote a test of the null hypothesis that these variables are independent. This is a large sample and all the expected frequencies are greater than 5 (check if you like, but SPSS tells you that the minimum is 46.68) and so we can use the chi-squared test, labelled “Pearson Chi-square” by SPSS: chi-squared = 14.91, df = 1, $P < 0.001$. Do not quote $P = 0.000$ as shown in the SPSS output. This is a technically correct representation of the probability to 3 decimal places, but because P cannot be actually equal to zero the convention is to emphasise this by putting $P<0.001$. Ignore all the other tests. Never quote something when you don’t know what it means and didn’t want it. SPSS often gives you more than you ask for.

We could also give an odds ratio for the table, this being a case control study. SPSS has this labelled “Risk” in Crosstabs:



Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for case (1.00 / 2.00)	.351	.205	.601
For cohort sex = 1.00	.556	.404	.767
For cohort sex = 2.00	1.586	1.255	2.004
N of Valid Cases	238		

The first row gives the odds ratio. As we don’t know what the second and third rows mean, we should ignore them. The odds ratio for being female given stroke is 0.35 (95% CI 0.21 to 0.60). As this is a case control study and stroke is a rare condition in the population we can use this as an estimate of the relative risk of stroke for women compared to men: 0.35 (95% CI 0.21 to 0.60).

Hence we have a big difference in sex between cases and controls but not in mean age.

Comparison of cholesterol and smoking history between cases and controls

We do analyses very similar to those for age and sex, so I won’t go through them in detail. We get:

Serum cholesterol (mmol/L)

	Number	Mean	Standard deviation
Cases	85	6.33	1.40
Controls	84	5.53	0.95

(Cholesterol was recorded to one decimal place, so I chose two decimal places for the mean and SD.) The difference in mean serum cholesterol (cases minus controls) was 0.79 mmol/L (SE = 0.18, 95% CI 0.43 to 1.16 mmol/L, P < 0.001). We have good evidence that cases had higher mean cholesterol, and we estimate this difference to be between 0.43 and 1.16 mmol/L.

We cross-tabulate smoking by case control status:

Patient group	Smoking history		Total
	Never smoked	Has smoked	
Cases	30 (29.7%)	71 (70.3%)	101 (100%)
Controls	100 (73.5%)	36 (26.5%)	136 (100%)

The difference is highly significant by a chi-squared test, chi-squared = 44.95, df = 1, P < 0.001. The odds ratio is 0.152 (95% CI 0.09 to 0.27). So not having ever been a smoker is associated with a much lower risk of stroke. We could turn this the other way round and recode the variable so that smoking has the lower code, to get an odds ratio for smoking and stroke and hence an estimate of the relative risk of a stroke for smokers. This gives OR = 6.6

(95% CI 3.7 to 11.6). We estimate that smoking increases the risk of a stroke by between 3.7 and 11.6 times.

Could age or sex differences have any affect on the relationships between stroke and cholesterol or smoking history?

The cases and controls did not differ greatly in mean age, but they do in sex. Far more of the cases were males than were the controls. If we found that males had higher mean cholesterol than females, this could explain the difference in mean cholesterol between cases and controls.

Serum cholesterol (mmol/L)

Sex	Number	Mean	Standard deviation
Females	80	6.03	1.48
Males	89	5.85	1.02

The difference in mean serum cholesterol (females minus males) was 0.18 mmol/L (SE = 0.20, 95% CI –0.21 to 0.56 mmol/L, P = 0.4). We have no evidence that men had higher mean cholesterol than women, and it is not plausible that this was the mechanism by which cases had higher mean cholesterol than controls.

If we tabulate smoking history by sex, we get:

Sex	Smoking history		Total
	Never smoked	Has smoked	
Females	68 (62.4%)	41 (37.6%)	109 (100%)
Males	62 (48.4%)	66 (51.6%)	128 (100%)

The difference is significant by a chi-squared test, chi-squared = 4.63, df = 1, P = 0.03. Men were more likely to smoke than women. Hence the excess of men among the cases could explain the association between stroke and smoking. We can investigate this a bit further by doing the case control versus smoking tabulation for men and women separately:

Sex	Patient group	Smoking history		Total	Chi-squared test with 1 d.f.
		Never smoked	Has smoked		
Females	Cases	9 (28.1%)	23 (71.9%)	32 (100%)	26.66, P < 0.001
	Controls	59 (76.6%)	18 (23.4%)	77 (100%)	
Males	Cases	21 (30.4%)	48 (69.6%)	69 (100%)	19.43, P < 0.001
	Controls	41 (69.5%)	18 (30.5%)	59 (100%)	

Hence for men and for women separately there is a highly significant association between stroke and smoking. (In practice, we would not analyse the females and males separately, but would use a method called logistic regression, outside the scope of this course. This gives the relative risk of stroke for smokers, adjusted for sex, RR = 6.32, P < 0.001.)

Discussion

We have found clear associations between stroke and higher serum cholesterol, and between stroke and having smoked. Neither of these can be explained by differences in age or sex distribution between cases and controls.

Quite a lot of the serum cholesterol measurements are missing, and we would like to know why this is. Is there any systematic difference in what led them to be missing between the two groups? Such differences might produce spurious relationships. Also, we have a small number of missing observations for other variables in the analysis, which weakens it a little.

We cannot conclude from these data that high cholesterol causes stroke. It may be that having a stroke increases cholesterol, or that some factor that increase the risk of stroke also increases serum cholesterol. It is implausible that history of smoking is a result of stroke, but it is possible that some other factor both increases the risk of stroke and the risk of smoking. We would have to use other knowledge to shed light on these possibilities, which is beyond the scope of this report. We can say from the analysis neither sex nor age is a third variable which produces a non-causal relationship between stroke and either cholesterol or sex.

Conclusions

Stroke is associated with raised serum cholesterol and with a history of cigarette smoking. The mean serum cholesterol is estimated to be between 0.43 and 1.16 mmol/L higher in stroke patients, the risk of stroke for those with a history of smoking is between 3.7 and 11.6 times the risk for those who have never smoked. These associations do not appear to be explained by age or sex differences between stroke cases and people who have not had strokes.

Key points in writing statistical reports

- Do not put in chunks of unmodified SPSS output.
- Do not put in analyses you do not want or do not understand.
- Watch out for impossible values. Do not leave them in the analysis.
- Always be aware that there may be more than one explanation for a relationship.

Martin Bland
13 December 2005