Reporting statistical analyses

Martin Bland

Professor of Health Statistics

University of York

http://www-users.york.ac.uk/~mb55/msc/

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

Structure of a statistical report

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.

Look at 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
➤ categorical classifications

What analysis to do when

Analysis	Continuous variable	Categorical variable
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

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 control status, age in years, sex, serum cholesterol, ever smoked.

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 compared. 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:
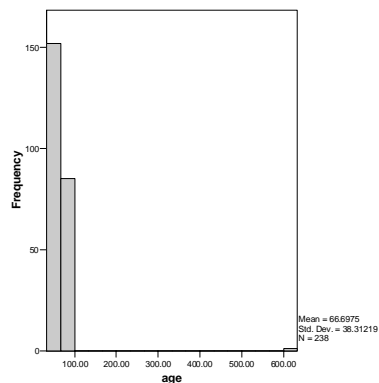
Results: *Checking for impossible values*

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)".
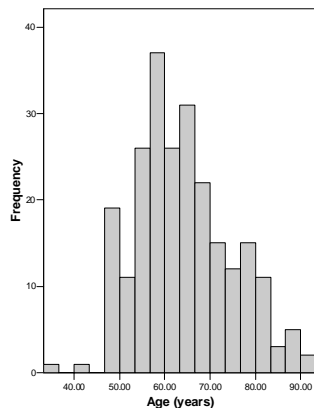

2
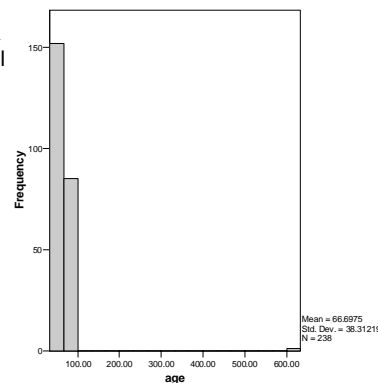
Results: *Checking for impossible values*

The graph now looks much better.

We can also remove the mean and standard deviation. Usually given to too many decimal places.
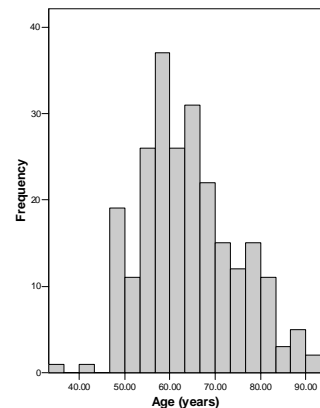
This information is better presented elsewhere.

Results: *Checking for impossible values*

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.

Results: *Checking for impossible values*

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 not need more than one decimal place for mean and standard deviation.
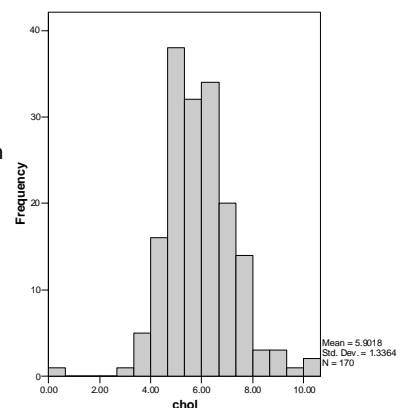
Just because the computer prints them out, you do not have to use them.

Results: *Checking for impossible values*

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.

Mean = 5.9018
Std. Dev. = 1.3364
N = 170

Results: *Checking for impossible values*

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.

Mean = 5.9367
Std. Dev. = 1.26018
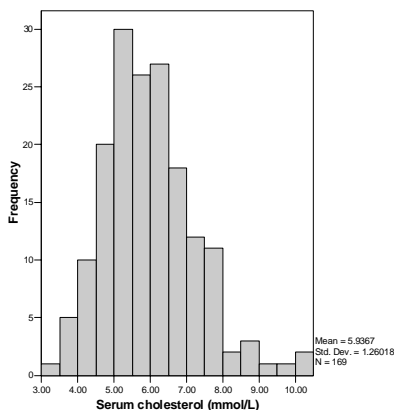N = 169

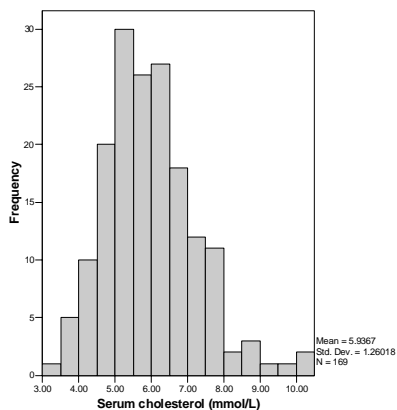**Results:** *Checking for impossible values*

Serum cholesterol has a positively skew distribution with mean 5.9 mmol/L and standard deviation 1.3 mmol/L.

The mean and standard deviation panel has been deleted for presentation.
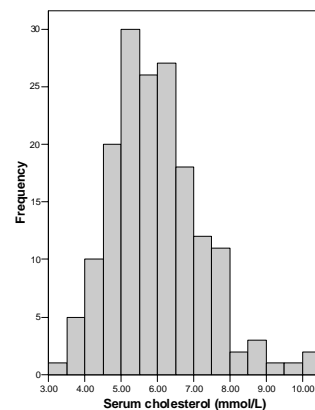
Results: *Checking for impossible values*

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%.

Mean = 5.9367
Std. Dev. = 1.26018
N = 169

(Histogram: Frequency vs Serum cholesterol (mmol/L))

Results: *Checking for impossible values*

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

Results: *Checking for impossible values*

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%)

Results: *Checking for impossible values*

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%)

Results: *Checking for impossible values*

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%)

Results: *Checking for impossible values*

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

The percentage we want is now the valid percentage.

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

Results

***Comparing age and sex distributions between cases
and controls***

For age, large sample comparison of two means.

Use two sample t method for unequal variances.

Results

***Comparing age and sex distributions between cases
and controls***

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.

Results

***Comparing age and sex distributions between cases
and controls***

Or give a reduced version of the table:

Group	Number	Mean	Standard deviation
Cases	100	64.9	9.1
Controls	137	63.9	11.4

Results

***Comparing age and sex distributions between cases
and controls***

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.

Results

***Comparing age and sex distributions between cases
and controls***

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.

Results

***Comparing age and sex distributions between cases
and controls***

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

Patient group	Female	Male	Total
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.

Results

Comparing age and sex distributions between cases and controls

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

Test of the null hypothesis that these variables are independent.

chi-squared = 14.91, df = 1, P < 0.001.

Do not quote P = 0.000 as shown in the SPSS output.

Results

Comparing age and sex distributions between cases and controls

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

Test of the null hypothesis that these variables are independent.

chi-squared = 14.91, df = 1, P < 0.001.

Ignore all the other tests. Never quote something when you don't know what it means and didn't want it.

Results

Comparing age and sex distributions between cases and controls

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.

Results

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.

Patient group	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.

Results

Comparison of cholesterol and smoking history between cases and controls

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

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.

Results

Comparison of cholesterol and smoking history between cases and controls

We cross-tabulate smoking by case control status:

Patient group	Never smoked	Has smoked	Total
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.

Results

Comparison of cholesterol and smoking history between cases and controls

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

The odds ratio is 0.152 (95% CI 0.09 to 0.27).

Not having ever been a smoker is associated with a much lower risk of stroke.

Results

Comparison of cholesterol and smoking history between cases and controls

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

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.

Results

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

Results

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

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).

Results

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

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.

Results

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

If we tabulate smoking history by sex, we get:

Sex	Never smoked	Has smoked	Total
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.

Results

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

We can investigate this a bit further by doing the case control versus smoking tabulation for men and women separately.

Females:

Patient group	Never smoked	Has smoked	Total
Cases	9 (28.1%)	23 (71.9%)	32 (100%)
Controls	59 (76.6%)	18 (23.4%)	77 (100%)

chi-squared = 26.66, df = 1, P < 0.001

Results

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

We can investigate this a bit further by doing the case control versus smoking tabulation for men and women separately.

Males:

Patient group	Never smoked	Has smoked	Total
Cases	21 (30.4%)	48 (69.6%)	69 (100%)
Controls	41 (69.5%)	18 (30.5%)	59 (100%)

chi-squared = 19.43 , df = 1, P < 0.001

Results

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

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.

Discussion

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.

Discussion

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.