British Standards Institution Study Day

Types of data and how they can be analysed

Martin Bland

Prof. of Health Statistics
University of York

http://martinbland.co.uk

Types of data

We can look at how people present results in a leading journal: *The British Medical Journal*.

Link: http://www.bmj.com/archive/online/2011/05-30

Articles published between 30 May 2011 and 5 Jun 2011.

Four research papers.

Types of data

Effect of evidence based risk information on "informed choice" in colorectal cancer screening: randomised controlled trial

Results . . . 345/785 (44.0%) participants in the intervention group made an informed choice, compared with 101/792 (12.8%) in the control group (difference 31.2%, 99% confidence interval 25.7% to 36.7%; P<0.001). More intervention group participants had "good knowledge" (59.6% (n=468) v 16.2% (128); difference 43.5%, 37.8% to 49.1%; P<0.001). A "positive attitude" towards colorectal screening prevailed in both groups but was significantly lower in the intervention group (93.4% (733) v 96.5% (764); difference −3.1%, −5.9% to −0.3%; P<0.01). The intervention had no effect on the combination of actual and planned uptake (72.4% (568) v 72.9% (577); P=0.87) . . . .

Types of data

Effect of evidence based risk information on "informed choice" in colorectal cancer screening: randomised controlled trial

Outcome variables:

- making an informed choice,
- having “good knowledge”,
- having a “positive attitude” towards colorectal screening,
- uptake of screening.

All “yes or no” variables.

Qualitative or categorical.

Two categories ➔ dichotomous.

Types of data

Longer term effects of very low energy diet on obstructive sleep apnoea in cohort derived from randomised controlled trial: prospective observational follow-up study

Results . . . After one year the apnoea-hypopnoea index had improved by −17 events/hour (−13 to −21) and body weight by −12 kg (−10 to −14) compared with baseline (both P<0.001). . . . At one year, 30/63 (48%, 95% confidence interval 35% to 60%) no longer required continuous positive airway pressure and 6/63 (10%, 2% to 17%) had total remission of obstructive sleep apnoea (apnoea-hypopnoea index <5 events/hour).

Types of data

Longer term effects of very low energy diet on obstructive sleep apnoea in cohort derived from randomised controlled trial: prospective observational follow-up study

Outcome variables:

- body weight (Kg),
- apnoea-hypopnoea index = the total number of complete cessations of breathing (apnoea) and partial obstructions (hypopnoea) for at least 10 seconds during sleep (events/hour).

Quantitative, can take any value within a range ➔ continuous.

Types of data

Longer term effects of very low energy diet on obstructive sleep apnoea in cohort derived from randomised controlled trial: prospective observational follow-up study

Outcome variables:

- required continuous positive airway pressure,
- had total remission of obstructive sleep apnoea.

Qualitative, two categories ➔ dichotomous.

Types of data

Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from Ontario, Canada

Results . . . The risk of adverse events increased with the mean length of stay of similar patients in the same shift in the emergency department. For mean length of stay ≥6 *v* <1 hour the adjusted odds ratio (95% confidence interval) was 1.79 (1.24 to 2.59) for death and 1.95 (1.79 to 2.13) for admission in high acuity patients and 1.71 (1.25 to 2.35) for death and 1.66 (1.56 to 1.76) for admission in low acuity patients) . . .

Qualitative, two categories ➔ dichotomous.

Types of data

Effect of weekly vitamin D supplements on mortality, morbidity, and growth of low birthweight term infants in India up to age 6 months: randomised controlled trial

Results . . . Between group differences were not significant for death or hospital admissions (92 among 1039 infants in the vitamin D group *v* 99 among 1040 infants in the placebo group; adjusted rate ratio 0.93, 95% confidence interval 0.68 to 1.29; P=0.68), or referral to the outpatient clinic for moderate morbidity. Vitamin D supplementation resulted in better vitamin D status as assessed by plasma calcidiol levels at six months. . . .

Types of data

Effect of weekly vitamin D supplements on mortality, morbidity, and growth of low birthweight term infants in India up to age 6 months: randomised controlled trial

Outcome variables:

❖ death or hospital admissions (yes or no).

Qualitative, two categories ➔ dichotomous.

❖ plasma calcidiol levels at six months.

Quantitative, can take any value within a range ➔ continuous.

Types of data

We have two types of data in these abstracts:

❖ Qualitative, dichotomous,

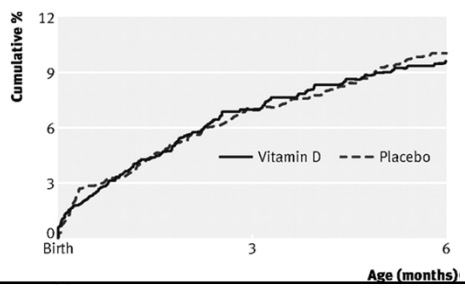
❖ Quantitative, continuous.

These are the kinds of data most often encountered in health research.

Types of data

There are other types of data.

❖ Time to event, combines quantitative, the time, with qualitative, whether the event happens. E. g. wound healing, admission to hospital or death (Vitamin D study).



Types of data

There are other types of data.

- Quantitative, discrete, only certain values possible. E.g. number of falls, number of attendances.
- Qualitative, ordered, more than two categories but ordered. E.g. physical condition or satisfaction with service rated as excellent, good, fair, poor, Vitamin D status as adequate (>50 nmol/L), mildly deficient (25-50 nmol/L), or severely deficient (<25 nmol/L).
- Qualitative, multinomial, more than two categories. E.g. single, married, divorced, widowed

These are the kinds of data most often encountered in health research.

How data can be analysed

Never collect data which you don't know how to analyse.

If in doubt, get advice before you collect anything.

Be sure you have suitable software which you know how to use.

How data can be analysed

Effect of evidence based risk information on "informed choice" in colorectal cancer screening: randomised controlled trial

Results . . . 345/785 (44.0%) participants in the intervention group made an informed choice, compared with 101/792 (12.8%) in the control group (difference 31.2%, 99% confidence interval 25.7% to 36.7%; P<0.001). More intervention group participants had "good knowledge" (59.6% (n=468) v 16.2% (128); difference 43.5%, 37.8% to 49.1%; P<0.001). . . .

Type of comparison: comparison of two groups.

How data can be analysed

Longer term effects of very low energy diet on obstructive sleep apnoea in cohort derived from randomised controlled trial: prospective observational follow-up study

Results . . . After one year the apnoea-hypopnoea index had improved by −17 events/hour (−13 to −21) and body weight by −12 kg (−10 to −14) compared with baseline (both P<0.001). . . . At one year, 30/63 (48%, 95% confidence interval 35% to 60%) no longer required continuous positive airway pressure and 6/63 (10%, 2% to 17%) had total remission of obstructive sleep apnoea (apnoea-hypopnoea index <5 events/hour).

Type of comparison: change within one group.

How data can be analysed

Types of comparison:

- comparison of two groups,
- change within one group.

Likely to be the main comparisons in health research.

Two main types of data:

- qualitative, dichotomous,
- quantitative, continuous.

Compare two groups, dichotomous data

Effect of evidence based risk information on "informed choice" in colorectal cancer screening: randomised controlled trial

Results . . . 345/785 (44.0%) participants in the intervention group made an informed choice, compared with 101/792 (12.8%) in the control group (difference 31.2%, 99% confidence interval 25.7% to 36.7%; P<0.001). . . .

Difference between two proportions.

Confidence interval for the difference.

Needs a “large” data set.

At least 5 “yes”s and 5 “no”s per group.

Compare two groups, dichotomous data

Effect of evidence based risk information on "informed choice" in colorectal cancer screening: randomised controlled trial

Results . . . 345/785 (44.0%) participants in the intervention group made an informed choice, compared with 101/792 (12.8%) in the control group (difference 31.2%, 99% confidence interval 25.7% to 36.7%; P<0.001). . . .

Difference between two proportions.

Test of significance: chi-squared test.

Needs a “large” data set.

At least 5 “yes”s and 5 “no”s per group, approximately. (Usual condition is a bit more complicated and a bit more liberal.)

Compare two groups, dichotomous data

Effect of evidence based risk information on "informed choice" in colorectal cancer screening: randomised controlled trial

Results . . . 345/785 (44.0%) participants in the intervention group made an informed choice, compared with 101/792 (12.8%) in the control group (difference 31.2%, 99% confidence interval 25.7% to 36.7%; P<0.001). . . .

Difference between two proportions.

Test of significance: Fisher’s exact test.

Any sample size.

Compare two groups, dichotomous data

Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from Ontario, Canada

Results . . . The risk of adverse events increased with the mean length of stay of similar patients in the same shift in the emergency department. For mean length of stay ≥6 *v* <1 hour the adjusted odds ratio (95% confidence interval) was 1.79 (1.24 to 2.59) for death and 1.95 (1.79 to 2.13) for admission in high acuity patients and 1.71 (1.25 to 2.35) for death and 1.66 (1.56 to 1.76) for admission in low acuity patients) . . .

Compare two groups, dichotomous data

Odds ratio

What is "odds"?

Odds of admission = $\frac{\text{number admitted}}{\text{number not admitted}}$

Odds ratio (OR) for admission, long wait vs. short wait =

$$\frac{\text{odds of admission, long wait}}{\text{odds of admission, short wait}}$$

Many useful mathematical properties.

Easy to find confidence interval if sample "large".

Significance test: chi-squared or Fisher's exact test.

Compare two groups, dichotomous data

Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from Ontario, Canada

Results . . . The risk of adverse events increased with the mean length of stay of similar patients in the same shift in the emergency department. For mean length of stay ≥6 *v* <1 hour the adjusted odds ratio (95% confidence interval) was 1.79 (1.24 to 2.59) for death and 1.95 (1.79 to 2.13) for admission in high acuity patients and 1.71 (1.25 to 2.35) for death and 1.66 (1.56 to 1.76) for admission in low acuity patients) . . .

What is "adjusted"?

Compare two groups, dichotomous data

What is "adjusted"?

Adverse events might be related to the particular emergency department.

They might be related to patient characteristics including age group, sex, calendar month, weekend/holiday versus weekday, time of day or night, average income level of the patient's neighbourhood and whether rural or urban, number of visits made to an emergency department in the past year, and main complaint.

We estimate the odds ratio of adverse events for patients who are the same on all of these but have different waits.

Method: logistic regression.

Easy when you know how!

Compare two groups, dichotomous data

Effect of weekly vitamin D supplements on mortality, morbidity, and growth of low birthweight term infants in India up to age 6 months: randomised controlled trial

Results . . . Between group differences were not significant for death or hospital admissions (92 among 1039 infants in the vitamin D group *v* 99 among 1040 infants in the placebo group; adjusted rate ratio 0.93, 95% confidence interval 0.68 to 1.29; P=0.68), . . .

Sometimes calculate the ratio of two proportions, called the risk ratio, or relative risk (RR).

Easy to find the confidence interval if sample 'large'.

Significance test, as for difference.

Adjustment very complicated.

Compare two groups, dichotomous data

Effect of weekly vitamin D supplements on mortality, morbidity, and growth of low birthweight term infants in India up to age 6 months: randomised controlled trial

Results . . . Between group differences were not significant for death or hospital admissions (92 among 1039 infants in the vitamin D group *v* 99 among 1040 infants in the placebo group; adjusted rate ratio 0.93, 95% confidence interval 0.68 to 1.29; P=0.68), . . .

In this study, the ratio presented is not actually a risk ratio.

A ratio of rates over time.

Calculated and adjusted by a very complicated statistical method, which we shall not consider further.

Compare two groups, time to event data

No examples in chosen week in the *BMJ.*

Following week (*BMJ* 2011; **342**: d3271) Sarah Cockayne *et al.* (Dept. of Health Sciences):

Cryotherapy versus salicylic acid for the treatment of plantar warts (verrucae): a randomised controlled trial

Abstract:

'There was no evidence of a difference between the salicylic acid and cryotherapy groups in self reported clearance of plantar warts at six months (29/95 (31%) v 33/98 (34%), difference –3.15% (–16.31 to 10.02), P=0.64) or in time to clearance **(hazard ratio 0.80 (95% CI 0.51 to 1.25), P=0.33)**.'

Compare two groups, time to event data

We give the word ‘hazard’ a special meaning:

the rate at which events happen.

In health research they are usually bad events, hence the choice of word, but not in this case.

The hazard can change over time.

Model: this process will be the same in each group and anything which increases the rate of clearance will do so in the same ratio throughout the follow-up.

The proportional hazards assumption.

We can check this in several ways.

Compare two groups, time to event data

The hazard ratio is the ratio of the rate of clearance in the cryotherapy group to the rate of clearance in the salicylic acid group.

Can adjust this as we did for an odds ratio.

Method: Cox proportional hazards regression.

Seldom have differences within one group for time-to-event data.

Change within one group, dichotomous data

Estimate and confidence interval: difference between proportions, conditional odds ratio.

Test of significance: McNemar's test.

Continuous data

Longer term effects of very low energy diet on obstructive sleep apnoea in cohort derived from randomised controlled trial: prospective observational follow-up study

Table 1. Baseline characteristics of obese men with moderate to severe obstructive sleep apnoea

	Mean (SD)	Range
Age (years)	48.7 (7.3)	33-61
Weight (kg)	113.1 (14.2)	86.9-139.9
Height (m)	1.80 (0.08)	1.65-2.03
Body mass index (BMI)	34.8 (2.9)	30.2-40.4

Continuous data

Table 1. Baseline characteristics of obese men with moderate to severe obstructive sleep apnoea

	Mean (SD)	Range
Age (years)	48.7 (7.3)	33-61
Weight (kg)	113.1 (14.2)	86.9-139.9
Height (m)	1.80 (0.08)	1.65-2.03
Body mass index (BMI)	34.8 (2.9)	30.2-40.4

(63 men)

Mean = average.

Range = smallest to largest.

SD = Standard Deviation = ?

Continuous data

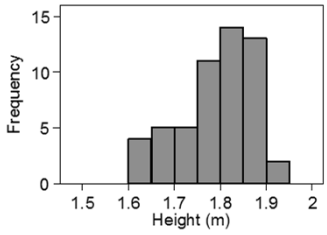
SD = Standard Deviation = ?

Standard deviation is a measure of variability.

	Mean (SD)	Range
BMJ: Height (m)	1.80 (0.08)	1.65-2.03
Research Methods, 2011	1.79 (0.08)	1.60-1.93

A histogram:

54 men.



Continuous data

SD = Standard Deviation = ?

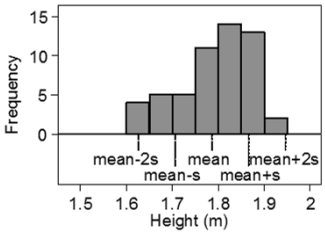
Standard deviation is a measure of variability.

	Mean (SD)	Range
BMJ: Height (m)	1.80 (0.08)	1.65-2.03
Research Methods, 2011	1.79 (0.08)	1.60-1.93

SD = 0.08 m.

About 2/3 within 1 SD of Mean.

About 95% within 2 SD of mean.



Compare two groups, continuous data

Effect of weekly vitamin D supplements on mortality, morbidity, and growth of low birthweight term infants in India up to age 6 months: randomised controlled trial

Results . . . Vitamin D supplementation resulted in better vitamin D status as assessed by plasma calcidiol levels at six months. . . .

	Vitamin D group (n=216)	Placebo group (n=237)	P value
Mean (SD) calcidiol level (nmol/L)	55.0 (22.5)	36.0 (25.5)	<0.001

Difference between means: confidence interval or test of significance by large sample Normal method or two sample t method (small samples).

What is a small sample? <50 in either group.

Compare two groups, continuous data

Two sample t method

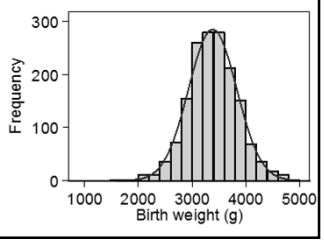
Small samples, <50 in either group.

Conditions the data must meet:

- Normal distribution,
- same variability in both populations.

Several ways to check these.

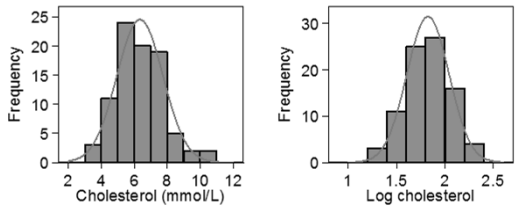
Birthweights, >37 weeks gestation



Compare two groups, continuous data

Two sample t method, data do not meet the conditions:

- find a mathematical transformation of the data which does,
- use other methods which don't need them (not so good).



Compare two groups, continuous data

Effect of weekly vitamin D supplements on mortality, morbidity, and growth of low birthweight term infants in India up to age 6 months: randomised controlled trial

Infants in the vitamin D treatment group had significantly higher plasma calcidiol levels at six months; crude mean difference 19.0 nmol/L (95% confidence interval 14.7 to 23.5; P<0.001). After adjustment for sunlight exposure and for factors associated with not having a result for calcidiol, the adjusted mean difference was 18.7 nmol/L (14.2 to 23.5; P<0.001).

Adjustment: multiple regression.

Change within one group, continuous data

Longer term effects of very low energy diet on obstructive sleep apnoea in cohort derived from randomised controlled trial: prospective observational follow-up study

Results . . . After one year the apnoea-hypopnoea index had improved by −17 events/hour (−13 to −21) and body weight by −12 kg (−10 to −14) compared with baseline (both P<0.001). . . .

Mean difference and confidence interval: large sample Normal method or one sample t method (small samples).

Test of significance: large sample paired Normal test or paired t test (small samples).

What is a small sample? <100.

Change within one group, continuous data

One sample t method

Small samples, <100.

Conditions the data must meet:

- ❖ Normal distribution for differences,
- ❖ mean and variability for differences same throughout scale.

Several ways to check these.

How data can be analysed

Still to come:

- how to analyse data in practice using SPSS,
- data entry,
- histograms, means, standard deviations and other statistics,
- comparing means,
- comparing proportions, odds ratios and risk,
- regression and adjustment.