Lecture 6
Quantitative data:
Maternal Smoking Data

Dr. Franco Ubaudi

The Nature of Data
Western Sydney University

Spring 2021

Outline

- Difference between probability and statistics
- Quantitative data: birth weight + sales office
- Data and experiment design

Probability vs statistics

Credit here to Victor Panaretos for the explanation and Franco Ubaudi for the examples

Probability:

1. Process of interest conceptualised as a probability model
2. Use model to learn about probability of potential outcomes
3. e.g. you have a biased coin, say 60/40, work out probability of getting 10 heads from 20 coin flips

Statistics:

1. Process of interest conceptualised as a probability model
2. Data viewed as observed outcomes from model
3. Use outcomes to learn about the model
4. e.g. pretend you don't know if above coin was biased, determine if a fair coin could be responsible for the result

Probabilist vs statistician

Job of the probabilist

Given a probability model P on a space Ω find the probability $P(A)$ that the outcome of the experiment is A.

Job of the statistician

Given an outcome of $A \subset \Omega$ – the data – of a probability experiment on Ω, tell me something "interesting" about the (unknown) probability model P that generated it.

Interesting questions

1. Are the data more more consistent with one or another model?
2. Given a family of models, can we determine which model generated the data?
3. What range of models are consistent with a given set of data?
4. How to best answer these questions? (is there even a best way?)

10 coin flips $X_1, X_2, \ldots X_{10}$

$X_i \overset{iid}{\sim} Bernoulli(\theta)$

$$\text{Outcome } (0, 0, 0, 1, 0, 1, 1, 1, 1, 1).$$

Probabilist asks:

- Probability of outcome as function of θ
- Probability of k-long run?
- If keep tossing, how many k-long runs?
- How does the sum of observations behave?

Statistician asks:

- Is the coin fair?
- What is a good guess for θ given the observations?
- What range of θ is plausible given the observations?
- How much error do we make?
- Is there a "best" solution to the above problems?
- How sensitive are answers to departures from the model?

Statistician: presented with the data and estimates a model, makes a hypothesis.

Probabilist: concerned with the consequences of a model.

Quantitative dataset

Recall maternal smoking dataset:

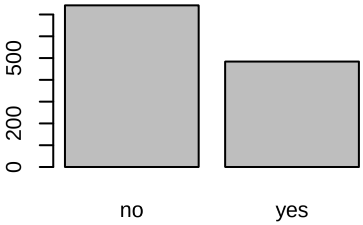


Figure: Count of smokers and non-smokers.

Does data support the hypothesis?

`bwt`, the infant birth weight, is quantitative – a continuous measurement of weight.

`smoke`, the smoking status of the mother – "yes" or "no" – is qualitative.

Summarising data

We first examine the qualitative variable `smoke`. It is a qualitative (not quantitative) so the sensible thing is to tabulate the counts.

no	yes
742	484

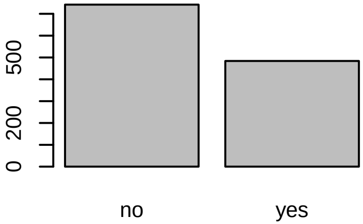


Figure: Count of smokers and non-smokers..

Numerical summaries of data

birth weight (bwt) is quantitative data So what is the “centrality” of the data, the “average”?

mean and **median**.

Mean

Mean

The *sample* mean of a set of numbers $\{x_1, x_2, \ldots, x_n\}$ is given by

$$\frac{x_1 + x_2 + \cdots + x_n}{n}$$

$\{4.8, 5.2, 3.9, 5.3, 3.8\}$

$23/5 = 4.6$

The mean birth weight from our data is 3,466.83 gm

Sample mean vs population mean

Median

Median

The median of a set of numbers $\{x_1, x_2, \ldots, x_n\}$ for n odd is a number such that half the data is greater and half less.

$\{4.8, 5.2, 3.9, 5.3, 3.8\}$?

Median is 4.8, 2 are above and 2 below.

n even has 2 contenders for middle point.

Pick half way between the two

$\{14.2, 9.1, 10.1, 8.1, 8.7, 12.2\}$ Median is 9.6

Median birth weight = 3,471.5 gm

Similar to mean.

Outliers

$\{4.8, 5.2, 3.9, 5.3, 3.8\} \rightarrow \{4.8, 50.2, 3.9, 5.3, 3.8\}$

Mean $= 13.6$, median $= 4.8$

Connection with expected value

$E[X]$ of a random variable X

Outcomes labelled

$$x_1, x_2, x_3, \ldots, x_n$$

have sample mean

$$m_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

As $n \to \infty$, $m_n \to E[X]$

m_n is an estimate for $E[X]$

How spread out is the data?

Range

For $\{x_1, x_2, \ldots, x_n\}$,

$$\text{range} = \max_i x_i - \min_i x_i$$

$\{4.8, 5.2, 3.9, 5.3, 3.8\}$?

Outliers: Range of $\{4.8, 50.2, 3.9, 5.3, 3.8\}$?

The range of the birth weights is 800 gm

How far is data typically away from the mean?

Variance

The variance of a set of numbers $\{x_1, x_2, \ldots, x_n\}$ is

$$s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$

The $1/(n-1)$ term is Bessel's correction.

n would give a biased estimate for $E[(X-\mu)^2]$

Standard deviation

Standard deviation

$s =$ the square root of its variance.

If data were $3, 3, 3, 3$ variance $= 0$. If often deviating from mean, the variance gets larger.

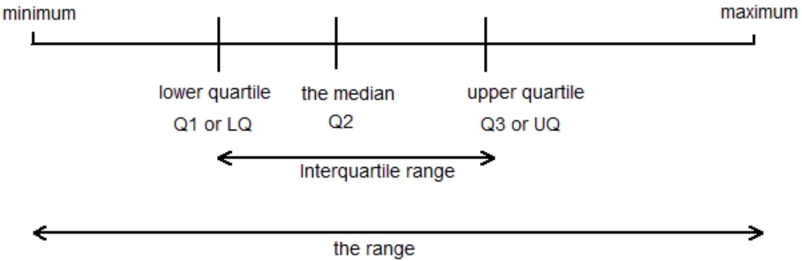
Yes we're MAD

Mean Absolute Deviation (MAD)

$$\mathrm{MAD} = \frac{1}{n} \sum_{i} |x_i - \bar{x}|$$

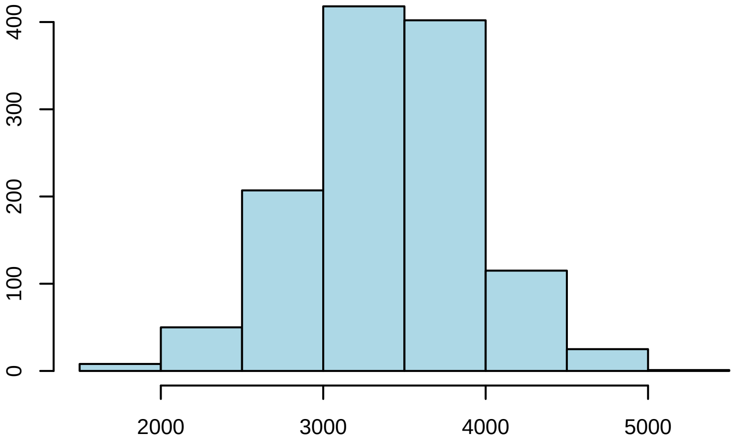
This is an estimate for $E[|X - \mu|]$.

Quartiles



For birth weight data, the first quartile Q1 = 3279, the third quartile Q3 = 3621.25, while the IQR is 342.25.

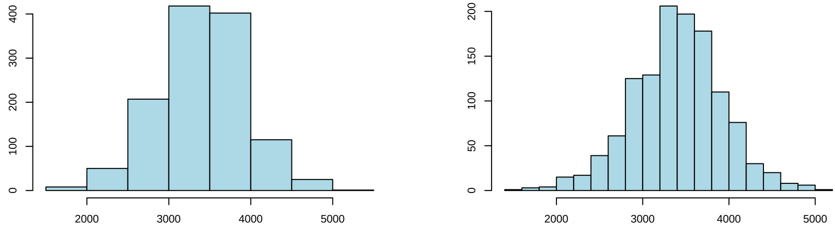
Histograms



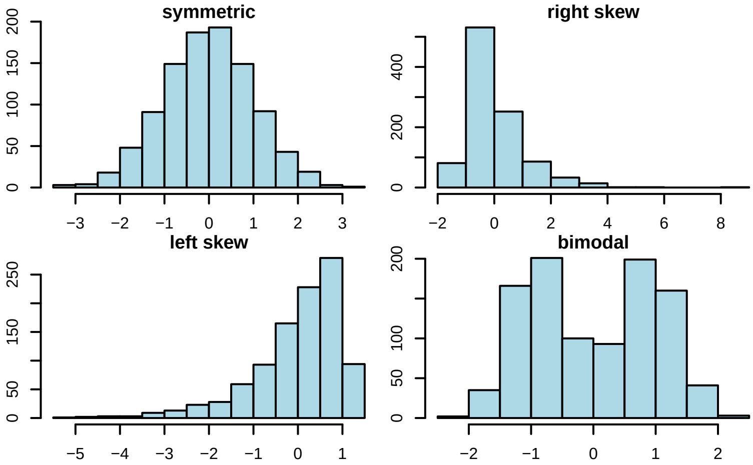
To create a histogram

1. Split the range into a **good** number of bins.
2. Assign each observation to one bin, increasing the bin count by 1.
3. Plot of the count of observations assigned to each bin.

How many bins?

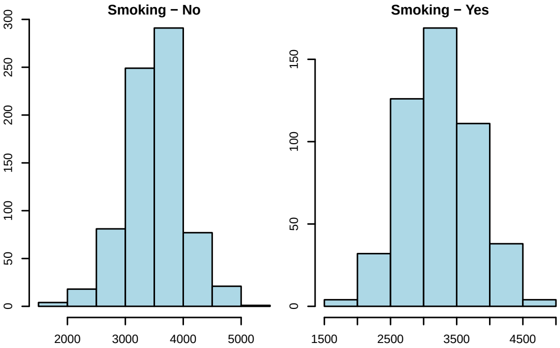


Note that a histogram is an estimate of the distribution of the data.

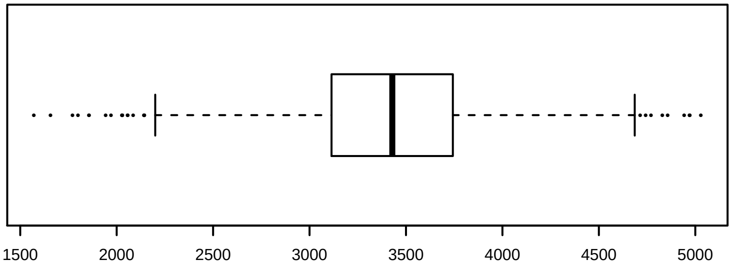

symmetric
0
50
100
150
200
−3
−2
−1
0
1
2
3
right skew
0
200
400
−2
0
2
4
6
8
left skew
0
50
150
250
−5
−4
−3
−2
−1
0
1
bimodal
0
50
100
200
−2
−1
0
1
2

Back to birth weight and smoking

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
no	1571	3229	3514	3515.639	3829	5029
yes	1657	2914	3286	3260.285	3600	4657



Box plots



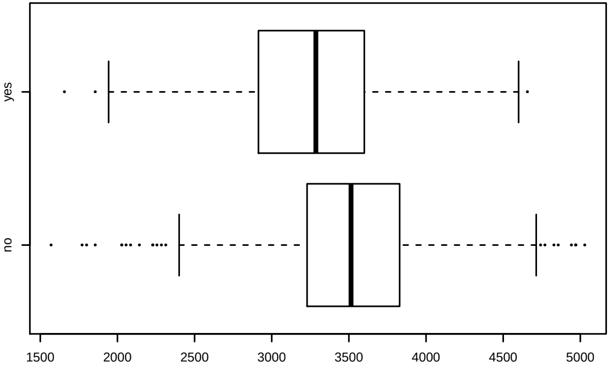
Central box contains the lower and upper quartiles (50% data)

Thick line is the median.

Dotted lines extend to furthest data point that is no more than 1.5 times the IQR from the box.

Data further away plotted as points.

Box plots by smoking status



Spread of groups about the same.

“no group” has higher median birth weight, more points outside 1.5 times the IQR.

Arriving at a hypothesis

“Does smoking affect maternal birth weight?”

Infants of smoking mothers have lower **average** birth weight: *mean* or *median*.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
no	1571	3229	3514	3515.639	3829	5029
yes	1657	2914	3286	3260.285	3600	4657

median & *mean* seem lower for smoking mothers.

The difference in *means*, for example, is 255.35

Is the difference statistically significant?

Hypothesis

Could the 255 gm difference be a chance occurrence and $\therefore$ the population *means*, actually be the same?

mean birth weight from smoking sample $m_1 = 3260.285$, non-smoking sample $m_2 = 3515.639$.

Population mean for all smoking pregnant women be μ_1, and all those who don't smoke be μ_2.

$$H_0 : \mu_1 \equiv \mu_2$$

$$H_1 : \mu_1 < \mu_2$$

An alternative alternative hypothesis

$$H_0 : \mu_1 \equiv \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

This choice changes meaning of p-value

The method of calculating the p-value is determined by H_1

Can we reject the null hypothesis?

Suppose H_0 was true

Then smoking labels do not affect the birth weight

So randomly allocating them should not change things substantially

Card piles

Two piles of cards. Have they been randomly dealt out?

Statistic = difference in means of numbers in each pile.

Randomise the data to simulate random piles:

1. Shuffle and deal two piles of same size.
2. Compute new difference in means.
3. Repeat to obtain 1000 mean differences.

If difference in means of original pile is not like any from randomised piles, then it is not likely the original two piles were randomly created.

First 10 observations from birth weight data:

	1	2	3	4	5	6	7	8	9	10
bwt	3429	3229	3657	3514	3086	3886	3943	3771	3429	4086
smoke	no	no	yes	no	yes	no	no	no	no	yes
smoke.random	no	yes	no	no	no	no	yes	yes	no	no

Difference in mean birth weights for 10,000 label shuffles:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-117.8866	-19.99175	0.2129742	0.3055107	20.87941	122.2384

Using actual smoking labels gives a difference of 255.35.

H_0 looks shaky, and we can indeed reject it.

Sales vs. office location

East and west offices sales.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
west	102.3928	140.7893	158.6182	154.0425	166.5258	193.9367
east	110.0611	144.6482	166.9310	162.6992	175.9951	206.5422

Is there evidence that the average sales per sales person between the two offices are different?

Does location affect sales?

Difference in mean sales between the east and west is 8.66

Null hypothesis = assume that the location does not effect the sales.

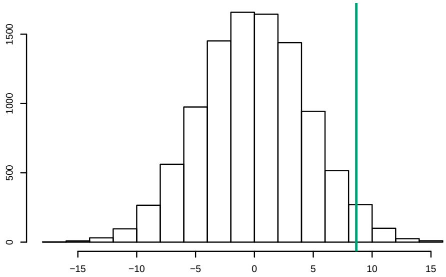
If true, we could shuffle the labels and no substantial change would occur.

10,000 differences in means:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-16.32221	-3.133544	-0.0633669	-0.0485031	3.077553	15.55492

289 of the 10,000 simulated differences in means are greater than the data difference in means 8.66

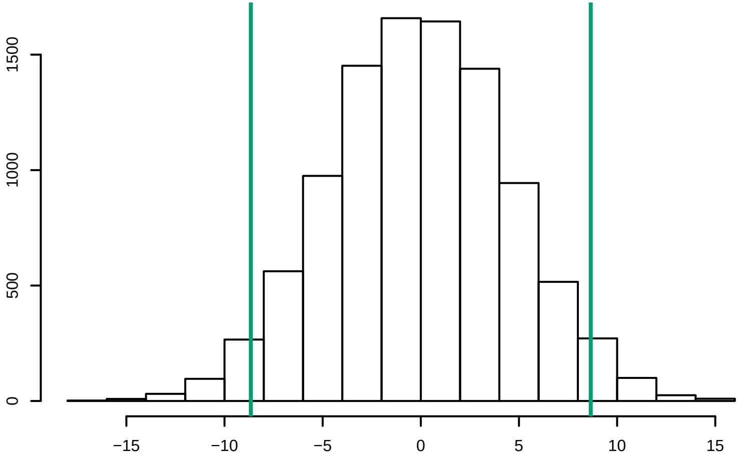
Difference in mean sales from randomly allocated sales offices.



The east office out performs the west office, so
$H_1 : \mu_{east} - \mu_{west} > 0$
So a one-sided test

At first sight, it seems that H_0 might explain the observation shown

Counting them all



For a two-sided test, or
$H_1 : \mu_{east} - \mu_{west} \neq 0$

Two-sided tests

two-sided test is when we want both sides.

For sales data there are 585: estimated **p-value** is 0.0585.

p-value

Proportion of samples from the randomisation that provide a statistic that is more extreme than the original sample, *assuming the null hypothesis*.

0.0585 was an estimate of the chance that randomisation can produce a difference in means larger than the sales sample produced.

Using medians

For the sales data 10,000 differences using shuffled location provides:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-21.21642	-3.1677	-0.2059212	0.0226339	3.222071	21.00838

The difference in medians west-east is 8.3128681

670 of the randomised differences in means > 8.3128681 and 624 were < -8.3128681

In the next worksheet, you will calculate the p-value of the above test.

Recap of hypothesis testing

Null hypothesis H_0 of no difference (in mean or median), alternative hypothesis H_1 is whether

- the means are different (**two-sided**),
- or one is greater than or one is less than other (**one-sided**).

To simulate the null hypothesis, we generated randomisations of the groups (shuffle) and compute the difference (in mean or median).

We then compared the actual difference (from the data) to randomised differences to get a p-value.

Using medians can give a different p-value (in the sales case).

The number of randomisations used determines *how accurately we estimate the p-value*.

Summary

- Can summarise numerical data using its mean, median, variance, quartiles and interquartile range.
- Numerical data can be visualised using a histogram.
- Boxplots allow us to compare numeric data from one or more groups.
- Shuffling labels allows us to test if the mean depends on the label.
- We can also test the difference in medians.
- These tests are hypothesis tests, that require a null hypothesis H_0 (that we can simulate), and an alternative Hypothesis H_1

What is Data?

Data can take many forms.

- Quantitative - numerical measurements or counts eg. bank balance, item cost, . . .
- Qualitative - non-numerical, categorical data eg. hair colour, group membership (treatment versus control)
- Unstructured Data - text, audio, images, video etc. eg. tweets, accident descriptions, audio transcripts, Instagram,. . .

Populations and samples

Data usually takes the form of a number of **observations** on one or more **variables**.

id	gender	race	ses	schtyp	prog	read	write	math	science	socst
90	female	white	high	public	academic	42	54	50	50	52
27	male	asian	middle	public	academic	53	61	61	57	56
96	female	white	high	public	academic	65	54	61	58	56
22	male	hispanic	middle	public	vocation	42	39	39	56	46
82	female	white	high	public	academic	68	62	65	69	61
56	male	white	middle	public	vocation	55	45	46	58	51
173	female	white	low	public	general	50	62	61	63	51
121	female	white	middle	public	vocation	68	59	53	63	61
146	male	white	high	public	academic	55	62	64	63	66
179	female	white	middle	private	academic	47	65	60	50	56

The Population

Data is rarely collected on all individuals of interest.

Usually only small subset e.g., survey of 1000 residents

population = set of all individuals of interest

sample = a subset of the population.

census: data is collected on all individuals of interest

It may represent only one time period e.g., current policy holders

Estimation versus Inference

What can the sample tell us about the population?

Estimate the mean income of 30 to 40 year old Australian males.

Inference
birth weight/smoking sample: is smoking **associated** with lower birth weight in the general population?

eels: do two species populations **generally** live in different habitats?

The sample is used to either *infer* something about the population, or *estimate* something in the population.

$Data + Method \implies Inference$

sample data + hypothesis test $\implies$ generalisation about the population

Inference: answer a specific question about the population using a sample.

Estimation: estimate something in the population using a sample.

Example estimation: est. average difference in birth weight for smoking and nonsmoking pregnant women

est. average number of deaths by horse kick per regiment per year in the Prussian army

Experiments versus observational studies

Observational studies: no intervention is made to the individuals in the population

e.g. surveys where a number of individuals are asked a specific set of questions.

Experiments involve an intervention in individuals sampled from a population

e.g. trialling a new drug on cancer patients.

Random sampling

A sample should be in some sense *representative* of the population of interest.

Is sampling those students who come to lectures representative of the whole population?

If a newspaper only asked its readers is that representative?

A representative sample allows us to **generalise** to the population.

The simplest way is **random sampling**.

Experiments - Random Allocation

Suppose we compare a new drug to a standard treatment. Best to draw patients at random and allocate drugs at random.

To compare the effect of a single drug on males and females, we can draw males and females at random, but we cannot allocate gender randomly.

Random allocation allows us to draw conclusions about *causality*: the treatment *causes* the effect.

Studies that cannot feasibly allocate **treatments** at random are observational.

We cannot allocate gender.

Environmental study: when contaminated sites to pristine ones, we cannot allocate which sites are to be contaminated.

Most environmental studies are observational.

Prospective vs. retrospective studies

Prospective studies: set up the sample and follow the outcomes (for example, clinical trials)

Retrospective studies: look at already collected data and try and associate risk factors with outcomes

Study design

Bias is caused when the sample is not representative.

Estimates may not be representative of the population but vary depending on the sample.

For example, estimate the size of a crab population by measuring those caught in a particular trap. Larger crabs might having a higher chance of being caught.

Or responder/non-responder bias in surveys: only those with strong opinions may reply.

Regression to the mean

One type of biased sampling.

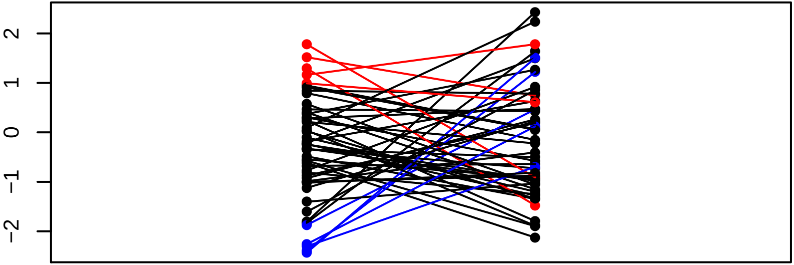
When measuring a group of individuals it may be that some are extreme, but when measured again they are not.

Suppose a new blood pressure treatment is tested.

Those with high blood pressure are selected for treatment. They are trialled on the drug for a period, and then retested.

Generally they would have lower second measurements.

Regression to the mean



50 pairs of $N(0,1)$ points. Highest 5 in red, lowest in blue.

Controlling variability

Which weight loss program study of diving people into two groups would you think is best?

1. Put one group on the program and the other not. Weigh (only) after they have been on the program.
2. Match so that for each person in the program group there is a person of similar weight in the non-program group. Weighing (only) after they have been on the program.
3. Put one group on the program and the others not. Then weighing before the program and after, and comparing the changes in weight.

Given big enough samples we could most likely distinguish any difference from any of them.

The third design eliminates uninteresting variability by considering changes.

Multiple sources of variability

- Population average: eucalyptus trees would have an average height for example.
- Natural variation: Each tree just naturally has a different height due to genetics and/or environment.
- Temporal variation: A single tree may have different heights in its life cycle.
- Measurement error: Measuring the same tree at the same time, may give different heights at different measurement attempts

Confounding

When sub-populations of interest happen to coincide with features not of primary interest. It can occur by chance, or by bad design.

Do more lichen grow on north or south facing sides of trees?

If we look at the north side of trees on Hawkesbury campus, and the south side of trees on Parramatta campus, then “side” and “campus” are confounded.

We don’t know if differences are due to “side” or “campus”

Confounding can usually be avoided by careful choice of a sample.

Placebo effect

A kind of confounding.

Compare treated individuals to untreated individuals: the act of offering *any* treatment may have an effect.

New treatments can be compared to a control (placebo):

- an inactive treatment that mimics the active treatment: chalk pill versus active pill;
- saline injection versus active drug injection;
- “cup of tea and a chat” versus formalised counselling.

Sampling methods

The simplest form of sampling is **random sampling**: every member of the population is given the same random chance of being chosen for a sample. But can be expensive.

Cluster sampling: if the population is all residents of a region, the clusters might be the towns in the region.

Done correctly, cluster sampling does not introduce bias.

Stratified sampling: accounts for a potentially confounding variable.

To survey people on internet attitudes, you might first divide the population into age groups, and split the sample across groups.

Unless age groups are sampled in proportion to their size, adjustments will need to be made.

Blocking

4 fields, 4 treatments. Compare plant growth.

Option 1: randomly allocate treatments to fields.

Option 2: Divide each field into four plots, and use each treatment once in each field.

Option 2 (fields are blocks) accounts for any random variation due to the fields alone.

Principles of study design

In general there are a few key principles of study design:

1. randomisation: eliminates bias and allows generalisation.
2. control variation: allows more subtle effects to be detected.
3. replication: helps to overcome variation due to unknown sources.
4. blocking: helps to overcome variation due to known sources

Summary

- Data comes in many forms. In its simplest form it is quantitative or qualitative.
- The population contains all items of interest. A sample is a randomly selected subset of these.
- We can infer answers to questions about a population, or estimate parameters.
- We can control the variables of an experiment, but only observe the variables of an observational study.
- Studies must be designed to account for different sources of variability.