

Beyond the Textbook: Resources for Statistics

Sidney Tyrrell
Coventry University

There is more here

- than we had time to cover
- but I have left it all in as you may find some of it useful
- especially the hyperlinks.

Starting with data

- [National Statistics.](#)
- [Neighbourhood Statistics](#)
- [Monthly digest of statistics](#)
- RPI and CPI
[Consumer Price Indices: A Brief Guide](#)

Social Trends

- [Social Trends 2010](#)
- Obtain the data by clicking on the table heading.

Data to info

- Involves communication
- Gapminder is an excellent site to make you rethink what world statistics represent
- <http://www.gapminder.org>

Data to info

- [Where does my money go?](#)
- A site which aims to promote transparency through visualisation of UK public spending.
- Do have a look!

In the news

- Michael Blastland at the BBC
- [Understanding Statistics in the News](#)
- [Straight Statistics](#)
- established by journalists and statisticians to improve the understanding and use of statistics.

In the news

- In the USA
- [Who's Counting?](#)
John Allen Paulos looks at the news.
- More topically with the General Election
- [Mori polls](#)

Data, data, what to do with it?

- First of all
- **Look** at it.

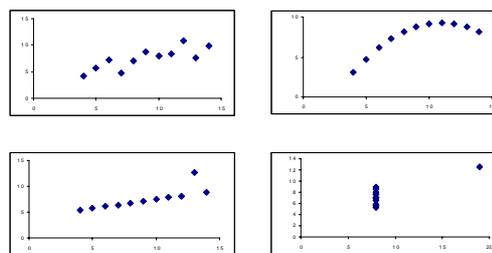
Anscombe's data sets

	x1	y1	x2	y2	x3	y3	x4	y4
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
mean	9	7.5	9	7.5	9	7.5	9	7.5
correlation coefficient		0.82		0.82		0.82		0.82

Looked? Now chart.

- Why chart?
- To help us understand the data.
- Sometimes a chart tells us far more than just computations.
- Frank Anscombe produced 4 data sets;
- data sets that were described by the **same linear model**

Was this what you expected?



The Visual Display of Quantitative Information: Edward R Tufte

Graphical excellence is the well designed presentation of interesting data

A matter of

- Substance
- Statistics
- Design

Graphical excellence consists of

- Complex ideas communicated with clarity, precision, and efficiency.
- It gives the viewer the greatest number of ideas in the **shortest** time with the **least ink** in the **smallest space**.
- It tells the **truth**.

Why chart?

- Having explored the data and discovered what it is telling us
- We need to communicate our findings:
 - Clearly
 - Succinctly.

Summaries

- A typical value
- The average
- Which to use:
 - Mean, median , mode

But is an average enough?

- Think of a simple data set with
 - **mean 10**
 - and
 - **median 10.**
- Ctrl P

Measures of spread

- Range
- Inter Quartile Range IQR
- Standard deviation
- What's the sd of **51, 53, 51** and **53** ?
- Write down **0 2 0 2**

Time for a break

- In pairs
- Take a packet of crisps
- Find the **e** symbol.

- What does this mean?

The E (e) mark

- Weights and Measures (Packaged Goods) Regulations 2006:

- These Directives set out three rules with which packers must comply:

First rule

- the actual contents of the packages should not be less, **on average**, than the nominal quantity;

Rules 2 and 3

- the proportion of packages which are short of the stated quantity by a defined amount (the “tolerable negative error” or TNE) should be less than a specified level; and
- no package should be short by more than twice the TNE.

The Tolerable Negative Error

Nominal quantity in grams and millilitres	Tolerable negative error	
	As a %age of nominal quantity	g or ml
5 to 50	9	-
from 50 to 100	-	4.5
from 100 to 200	4.5	-
from 200 to 300	-	9
from 300 to 500	3	-
from 500 to 1,000	-	15
from 1,000 to 10,000	1.5	-
from 10,000 to 15,000	-	150
above 15,000	1	-

How does the production manager cope?

- Using sampling.
- What's the general idea?

- A random sample is one where every member of the population has an equal chance of being chosen, **and**
- **every subset.**

Sampling examples

- How much of the picture are you getting?
- Important for the pollsters!

Random numbers

- Types of samples
- Will you get the same answer each time?
- Even if you use the same technique but start in a different place?
- **Simulation**

The Salk Vaccine Trial

- What's the aim?
- How will that be achieved?
- What would you do?
- Resources almost boundless.
- Are there any ethical issues?

What happened

- Interpreting the results.
- Using tables!

tables are tricky
what % of men are very active?

	men	women
very active	75%	50%
not very active	25%	50%
total	100%	100%

	men	women	total
Very active	86%	14%	100%
Not very active	68%	32%	100%

what % of the very active are men?

	men	women
very active	75%	50%
not very active	25%	50%
total	100%	100%

	men	women	total
Very active	86%	14%	100%
Not very active	68%	32%	100%

This is not the same question

- Be clear in your own mind what you want to compare.
- Characteristics of men v women?
- Characteristics of the activity level groupings.

- Put each group in a separate 'room' in your mind

- you want %'s to add to 100% for that 'room'.



look at the tables again

	men	women
very active	75%	50%
not very active	25%	50%
total	100%	100%



	men	women	total
Very active	86%	14%	100%
Not very active	68%	32%	100%

Do men exercise more than women?

- This brings us to
- hypothesis testing.

I have a hypothesis

- I need to collect some data
- And then evaluate the evidence
- Here goes:

Choose and write down one of the following numbers:

- **1** **2** **3** **4**

Collect the data

The hypothesis was:

- More people choose
- 3
- than any other number.

What do you think?

- What does the evidence suggest?
- How do you make up your mind?

Here's one collected earlier.

	1	2	3	4
number	16	23	34	12

- What does the evidence suggest?
- How do you make up your mind?

What about this one?

	1	2	3	4
number	16	18	20	18

- What does the evidence suggest?
- Can you make up your mind?

This illustrates the basic steps

- State the hypothesis
- Collect the data
- Evaluate the evidence

The hypothesis test

- Uses the evidence to decide
- whether the null hypothesis
- can be accepted
- or not.

- A good analogy is

An English court of law:

- The defendant is assumed **not guilty** until there is sufficient evidence to find otherwise.

And there can be

- Mistakes!
- Which is why statisticians are picky

- **Do not accept** rather than reject.

- **Do not reject** rather than accept.

Significance level

- A practical example:
- ESP extra sensory perception using playing cards.

Probability and Risk

- Corks

- Expected values and boxes

- [Monty Hall problem](#)

- [Birthday problem](#)

Probability and Risk

- David Spiegelhalter
- Professor of the Public Understanding of Risk.
- On [YouTube](#)
- Understanding Uncertainty [website](#)

Beyond the Textbook: resources and Ideas for S1 and S2; Sidney Tyrrell, Coventry University s.tyrrell@coventry.ac.uk and Sidney@dayoff.co.uk

This particular session involves a variety of activities which are hard to put on paper. Some of the handouts are print outs from web pages, so what follows here is not the full experience of the session but attempts to provide some useful resources for those who are unable to attend.

A first look at data.

We start with meeting data in the press – newspapers are a good source of stats stories. They often involve the presentation of data, or misrepresentation.

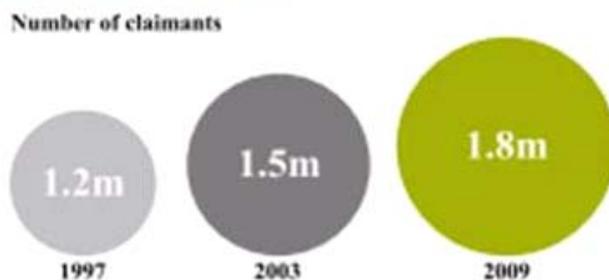
The Challenger: An Information Disaster

This is taken from <http://www.asktog.com/books/challengerExerpt.html>

Pie charts are a perennial problem:

From Radical Statistics, June 2010: “The Cabinet Office has issued an illuminating report on poverty, *The state of the nation: poverty, worklessness and welfare dependency in the UK*, which can be found at <http://www.cabinetoffice.gov.uk/media/410872/web-poverty-report.pdf> (it's a one-megabyte download). A picture is worth a thousand words, and the graphs on page 34 and 35 are models of their kind. Here's one by way of illustration:”

Figure 3.1: The numbers of working-age Disability Living Allowance claimants have increased by over 40% since 1997, from 1.2 million to 1.8 million



Source: DWP Tabulation Tool and Quarterly Statistical Releases

This is a reminder that there is a wealth of stats data available for free from a variety of Government offices and indeed from the [National Statistics](http://www.statistics.gov.uk) site.

This changes daily, and is a fascinating source of information particularly the [Neighbourhood Statistics](http://www.statistics.gov.uk/NeighbourhoodStatistics) section. Just type in your postcode and select either an overview or from a choice of data sets. It is a really powerful source of information, so do visit it if you haven't to find out about your own locality; please tell your local Councillors all about it.

My favourite reports on the NatsStats website are:

[Social Trends](http://www.statistics.gov.uk/socialtrends/) (<http://www.statistics.gov.uk/socialtrends/>)

and [Family Spending](http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=361), at <http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=361> which is a report on the Living Costs and Food Survey.

Click on the heading of a Table in any chapter of Social Trends and the underlying data will open in a spreadsheet for you. Magic!

For presentation of information in an interactive and imaginative way try [Gapminder](http://www.gapminder.org/) at <http://www.gapminder.org/>

Summarising data

Means, medians and modes – when to use which? Try thinking of an average as a typical value and that should help. There is a big difference between median household incomes and mean household incomes – politicians prefer the latter as it looks better.

Try calculating the mean and median incomes of 5 workers whose hourly rates are
 £7 £7 £7 £7 £22

And decide which gives the more representative value.

I can think of several data sets with a mean of 10 and a median of 10 but they are very different sets: 10; 10 10 10; 0 20 0 20; 9,10,11

What makes them different? It's the spread. I then explore the concept of the standard deviation using simple numbers.

Even knowing an average and a measure of spread does not tell us everything. In particular there are the interesting bits called outliers which are often mistakes or may be the really interesting and important bits of information, e.g. the outliers that revealed the ozone hole in the Antarctic and which were ignored for a time:

See <http://www.statsci.org/data/general/ozonehol.html>

Having got the data what can you or shouldn't you do with it. NB Anscombe's data sets

	x1	y1	x2	y2	x3	y3	x4	y4
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
mean	9	7.5	9	7.5	9	7.5	9	7.5
correlation coefficient		0.82		0.82		0.82		0.82

Having looked at this data you need to chart it before you do anything with it – resist the temptation to do a regression analysis until you have seen what they look like.

Data collection

How does all the data on Nat Stats get collected? For example, how is the RPI and CPI calculated using a typical basket of goods for a typical, and mythical family? See the article [on the 2010 basket of goods and services](http://www.statistics.gov.uk/CCI/article.asp?ID=2405&Pos=3&ColRank=1&Rank=160) at <http://www.statistics.gov.uk/CCI/article.asp?ID=2405&Pos=3&ColRank=1&Rank=160>

Just as interesting is how the sample is chosen, which leads to a discussion of sampling techniques and a handout of a map of Unitown showing houses in a town, and a list of their responses to a question, which can be adjusted according to the topic of the day, e.g. Do you support the government's plans for ...? The town can then be sampled using a variety of techniques and the results compared with each other and with the truth.

What can be inferred from the sample? This leads on to confidence intervals using [DISCUS](#) a free download of Excel spreadsheets from http://www.coventry.ac.uk/ec/research/discus/discus_home.html of which the sampling one and the confidence interval one can be quite useful here.

What's in an e number?

This refers to the E(e) mark on packaged goods thanks to the Weights and Measures (Packaged Goods) Regulations 2006. These Directives set out three rules with which packers must comply:

- the actual contents of the packages should not be less, on average, than the nominal quantity;
- the proportion of packages which are short of the stated quantity by a defined amount (the “tolerable negative error” or TNE) should be less than a specified level; and
- no package should be short by more than twice the TNE.

Nominal quantity in grams and millilitres	Tolerable negative error	
	As a %age of nominal quantity	g or ml
5 to 50	9	-
from 50 to 100	-	4.5
from 100 to 200	4.5	-
from 200 to 300	-	9
from 300 to 500	3	-
from 500 to 1,000	-	15
from 1,000 to 10,000	1.5	-
from 10,000 to 15,000	-	150
above 15,000	1	-

From there it is a short hop to hypothesis testing and asking people to choose one of the numbers 1, 2, 3 or 4. You might be quite surprised to see what happens. Does the evidence collected support my hypothesis that the number most people choose is 3? This leads to significance levels and another example using a pack of cards to illustrate how a significance level might be chosen.

An excellent example of experimental design and sample size, ethics and common sense is the world's largest public health trial – the Salk Vaccine trial. This was the American trial to see if the Salk vaccine really could prevent paralytic polio. Again see the end for further details.

Interpreting the results requires an understanding of tables, which can be tricky.

What % of men are very active?

	men	women
very active	75%	50%
not very active	25%	50%
total	100%	100%

	men	women	total
very active	86%	14%	100%
not very active	68%	32%	100%

Distributions

There's hardly time to fit this in but a use of random sampling numbers to simulate distributions is revealing.

Probability and Risk

- Expected values and boxes
- [Monty Hall problem](#) and [Birthday problem](#)
- David Spiegelhalter, Professor of the Public Understanding of Risk, on [YouTube](#)
- Understanding Uncertainty [website](#)

From [Tog on Software Design](#)

A case study; the Salk Vaccine Trial.

The Problem:

To conduct an experiment to test the efficacy of a new polio vaccine in the United States of America.

This was a real problem that resulted in the biggest public health experiment ever: the 1954 field trial of the Salk Poliomyelitis Vaccine.

Your task is to decide how YOU would have conducted this experiment, bearing in mind that your subjects are children, what you do to them will affect their health, maybe their lives and that certainly the outcome of your experiment could dramatically affect the lives of millions.

There are three important things to consider.

1. How will you set up a control group, especially bearing in mind that it was difficult to predict where the next outbreak might be?
2. How large will your sample be?
3. What are the ethical issues involved?

The Background.

In the early 1950's polio was a frightening disease, and one with the most inexplicable behaviour. It appeared in epidemic waves in one community one year and in another the next, hitting hardest at young children and surprisingly usually affecting those who were best off in terms of nutrition, housing, hygiene etc! Although it was responsible for relatively few deaths (about 6% of all deaths in the 5 to 9 age group) it left many children paralysed or crippled and some who could only survive in respirators. In 1952 there were over 60,000 cases in the United States.

These factors together with the earlier involvement of President Franklin D Roosevelt led to a very great concern about the nature of polio, with the result that money and effort were spent on research into the disease. The cause of polio was found to be a virus, and the search was on for an effective vaccine.

Smallpox and 'flu illustrate two different approaches to the preparation of vaccines. For smallpox a closely related live virus is used, cowpox, which is ordinarily incapable of causing serious disease in humans but which nevertheless gives rise to antibodies which protect against smallpox. In the case of 'flu, instead of using a closely related live virus, the vaccine is a solution of the dead 'flu virus, which still has enough antigenic activity to produce the required antibodies.

Both these methods were explored for polio. Vaccines were produced but at least one actually caused cases of paralytic polio ... so all were withdrawn! In the early 1950's Jonas Salk at the University of Pittsburgh developed a killed virus vaccine, which was shown to be safe and to induce high levels of antibodies in children on whom it had been tested. However, it was thought unsafe to release the vaccine for general use without convincing proof of its effectiveness; hence the decision to undertake a large scale field trial.

Polio is not easy to diagnose in its early stages, but its effects range from mild fever and weakness to paralysis and death. The virus turns out to be quite common, and it is thought that most adults had actually experienced a polio infection sometime in their lives without being aware of it. The rate of occurrence of polio at this time was about 50 per 100 000.

The Salk Vaccine Trial what happened

The table below summarises the results of the experiment actually carried out, which was conducted in two different ways. The first thing to note, however, are the numbers involved - 1,829,916 children!! In both areas there were control groups, but the difference lay in how these groups were chosen.

In the placebo control area every child was vaccinated, but some of them only received salt solution instead of the vaccine. Each vial of the fluid was identified by a code number so that NO-ONE involved in the vaccination, or in subsequently diagnosing a suspected case of polio, knew which child had received the vaccine and which child had received salt solution. This is called a DOUBLE-BLIND experiment. Only AFTER a final diagnosis had been made in each case, was the code broken to identify whether or not that child had received the vaccine.

Such an experiment does pose some ethical problems, and the practical ones of persuading parents to allow their children to be vaccinated with something which might be totally ineffective ... or might be the real thing (which of course might also be totally ineffective).

In the observed control areas vaccination was offered to all children in the second grade of participating schools - the TREATED GROUP -while the children in the first and third grades were the CONTROL GROUP who were also followed up for the incidence of polio.

Summary of study cases by diagnostic class and vaccination status (rates per 100,000)

<i>Poliomyelitis Cases</i>													
<i>Study Group</i>	<i>Study Population</i>	<i>All Reported Cases</i>				<i>Non-paralytic</i>				<i>Fatal polio</i>		<i>Not Polio</i>	
		<i>No.</i>	<i>Rate</i>	<i>No.</i>	<i>Rate</i>	<i>No.</i>	<i>Rate</i>	<i>No.</i>	<i>Rate</i>	<i>No.</i>	<i>Rate</i>	<i>No.</i>	<i>Rate</i>
<i>All areas: Total</i>	1,829,916	1,013	55	863	47	685	37	178	10	15	1	150	8
<i>Placebo control areas: Total</i>	749,236	428	57	358	48		36	88	12	4	1	70	9
Vaccinated	200,745	82	41	57	28	33	16	24	12	—		25	12
Placebo	201,229	162	81	142	71	115	57	27	13	4	2	20	10
Not inoculated*	338,778	182	54	157	46	121	36	36	11	-	-	25	7
Incomplete vaccinations	8,484	2	24	2	24	1	12	1	12				
<i>Observed control areas: Total</i>	1,080,680	585	54	505	47	415	38	90	8	11	1	80	7
Vaccinated	221,998	76	34	56	25	38	17	18	8		-	20	9
Controls**	725,173	439	61	391	54	330	46	61	8	11	2	48	6
Grade 2 not inoculated	123,605	66	53	54	44	43	35	11	9	—		12	10
Incomplete vaccinations	9,904	4	40	4	40	4	40	—	—	—	-	—	—

*Includes 8,577 children who received one or two Infections of placebo,

**First and third-grade total population

Source.- Adapted from T. Francis, Jr. (1955), Tables 2 and 3.

And Afterwards

After the report of the field trial was published the Salk vaccine was released for general use. Tragically a few batches actually caused serious polio, and the vaccine was withdrawn for a while. Meanwhile research on live-virus vaccines continued and these are largely used today.