



How statisticians think about data, and how they make graphs and summaries.



 About a single concept



 About the relationship between concepts

```

graph LR
    V1[Variable] --> V2[Variable]
    V2[Variable] --> V2
    V3[Variable] --> V2

```

The diagram illustrates two types of variables:

- Numerical** (also known as: Quantitative, Interval, Scale): Represented by a box labeled "variable name" containing a double-headed arrow. A note below states "(numbers: how far apart has meaning)".
- Categorical** (also known as: Qualitative): Represented by a box labeled "variable name" containing two small squares. A note below states "(words: how far apart has no meaning)".

The diagram shows a person interacting with a system. The person provides input to a process box, which then produces output. The output is further processed by another box, leading to a final output. The diagram is labeled with 'Input', 'Process', and 'Output'.

[illegible][illegible]

The diagram illustrates the process of aggregating data from multiple sources. On the left, two groups of people (represented by circles) are shown, each with arrows pointing towards a central box labeled "AGGREGATE DATA". This box then has an arrow pointing to a table on the right.

	1	2	3	4	5	6	7
1	15	10	12	18	20	22	25
2	18	15	14	16	19	21	23

- Subjects/rows are points
- Info represented by what the point lines up with on the axes...
... and by colour, shape etc
- Aim to see the variation, and see past the variation.

- Subjects are represented on the graph multiple times
- Info represented by what the point lines up with on the axes
- Points for the same subject are connected by lines

Shows how much of the data is in each zone.

Shows how spread out each quarter of the data is.

The figure is a composite image. On the left, there is a network graph with nodes and edges. Below it is a table with 5 columns: 'id', 'x', 'y', 'z', and 'w'. The table contains 5 rows of data. To the right of the table is a bar chart titled 'graph with errorbars'. The x-axis is labeled 'Node Type' and has two categories: 'A' and 'B'. The y-axis is labeled 'Frequency' and ranges from 0 to 10. There are two bars: a red bar for 'A' and a blue bar for 'B'. Both bars have error bars. The red bar is slightly taller than the blue bar.

id	x	y	z	w
1	0.1	0.2	0.3	0.4
2	0.2	0.3	0.4	0.5
3	0.3	0.4	0.5	0.6
4	0.4	0.5	0.6	0.7
5	0.5	0.6	0.7	0.8

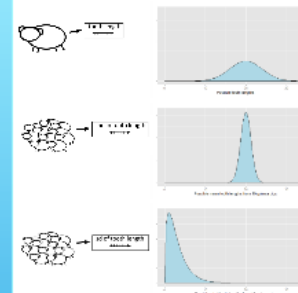
Figure 1: Schematic representation of the experimental design. The figure is divided into two main sections: 'Pre-Test' and 'Main Experiment'. The 'Pre-Test' section shows a flowchart for 'Pre-Test 1' (N=10) and 'Pre-Test 2' (N=10), leading to 'Pre-Test 3' (N=10). The 'Main Experiment' section shows a flowchart for 'Main Experiment 1' (N=10) and 'Main Experiment 2' (N=10), leading to 'Main Experiment 3' (N=10). The flowcharts include boxes for 'Pre-Test', 'Main Experiment', and 'Post-Test'.

- Your subjects are just **SOME** of the subjects you **COULD HAVE HAD**.
- Your group of subjects is just **ONE** of the groups you **COULD HAVE HAD**.
- Any one value of a variable is just **ONE** of the values you **COULD HAVE HAD** – even if it was calculated from other variables or recorded on the whole group.

EVERYTHING HAS A DISTRIBUTION

Tend to look like smooth "histograms"

PROBABILITY DENSITY FUNCTIONS



Maths Learning Centre
The University of Adelaide

Thinking About Data

How statisticians think about data, and how they make graphs and summaries.



VARIABLES come in two TYPES

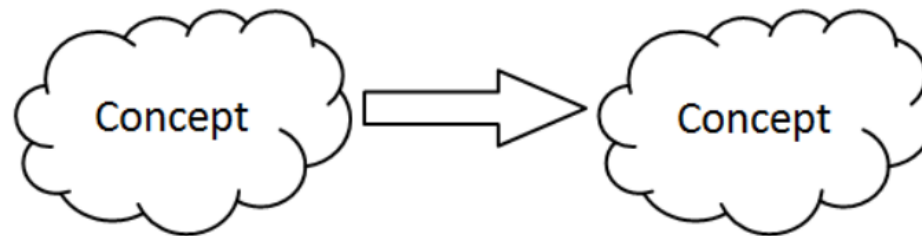


CU

STATS helps to answer QUESTIONS
Questions are about CONCEPTS



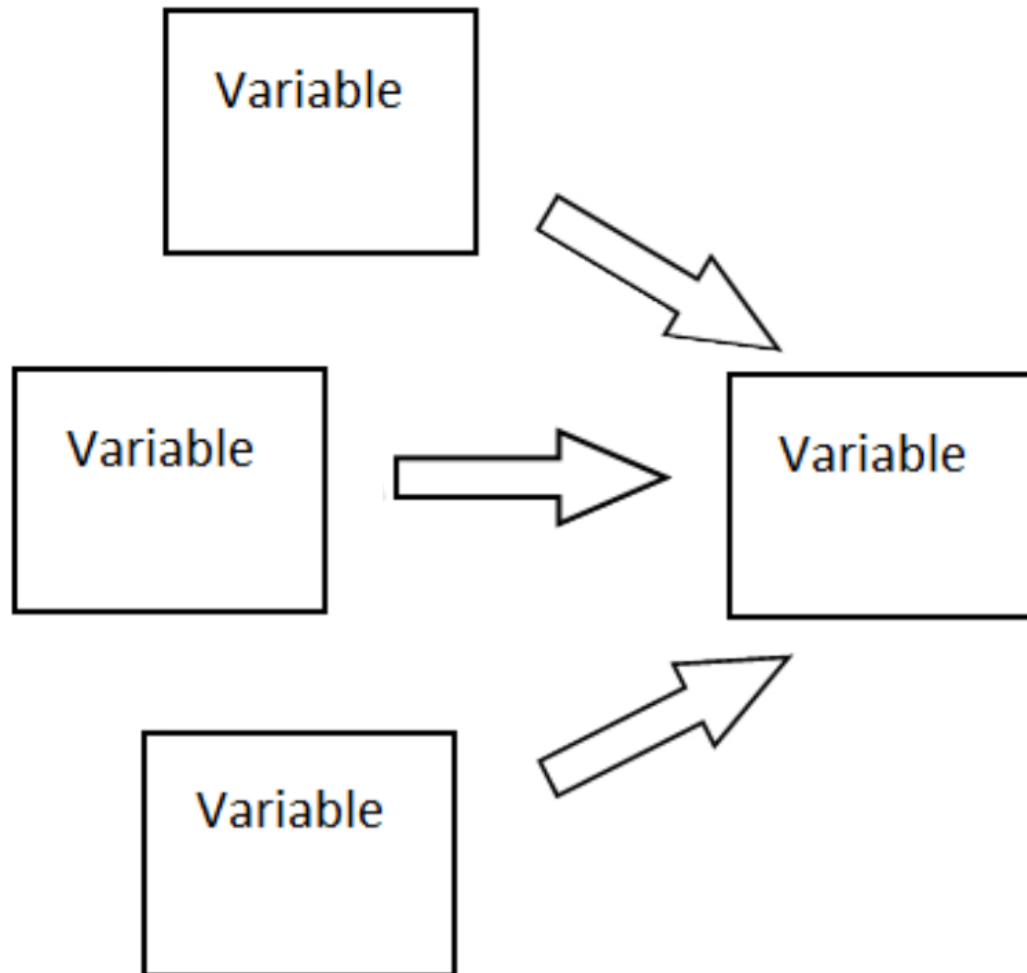
About a single concept



About the relationship
between concepts

CONCEPTS become VARIABLES

when you measure them



VARIABLES come in two TYPES

variable
name



Numerical

also known as:

Quantitative, Interval, Scale

(numbers: how far apart has meaning)

variable
name



Categorical

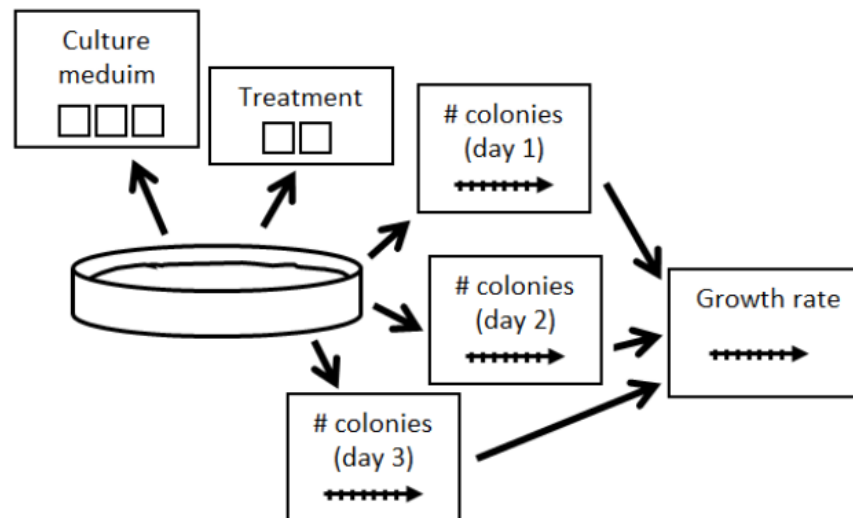
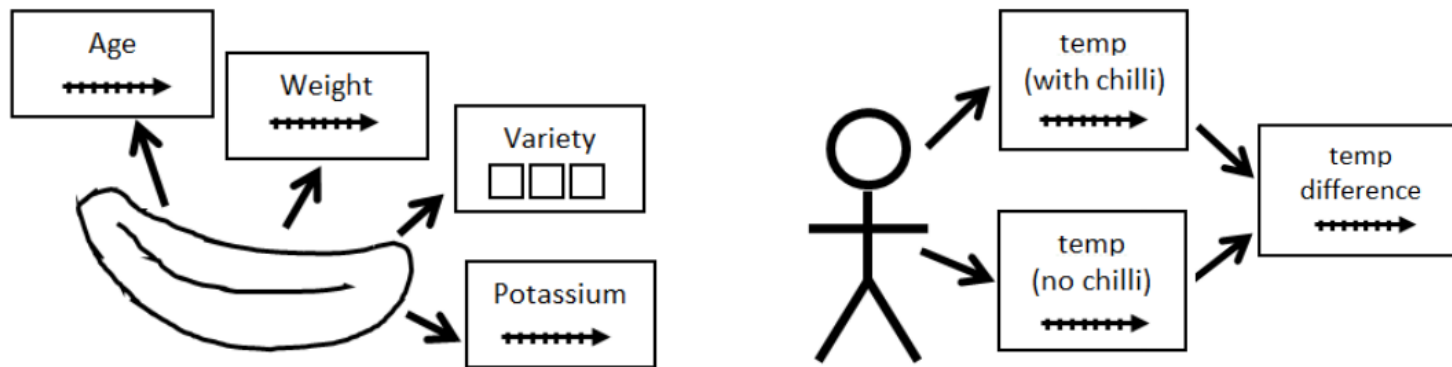
also known as:

Qualitative

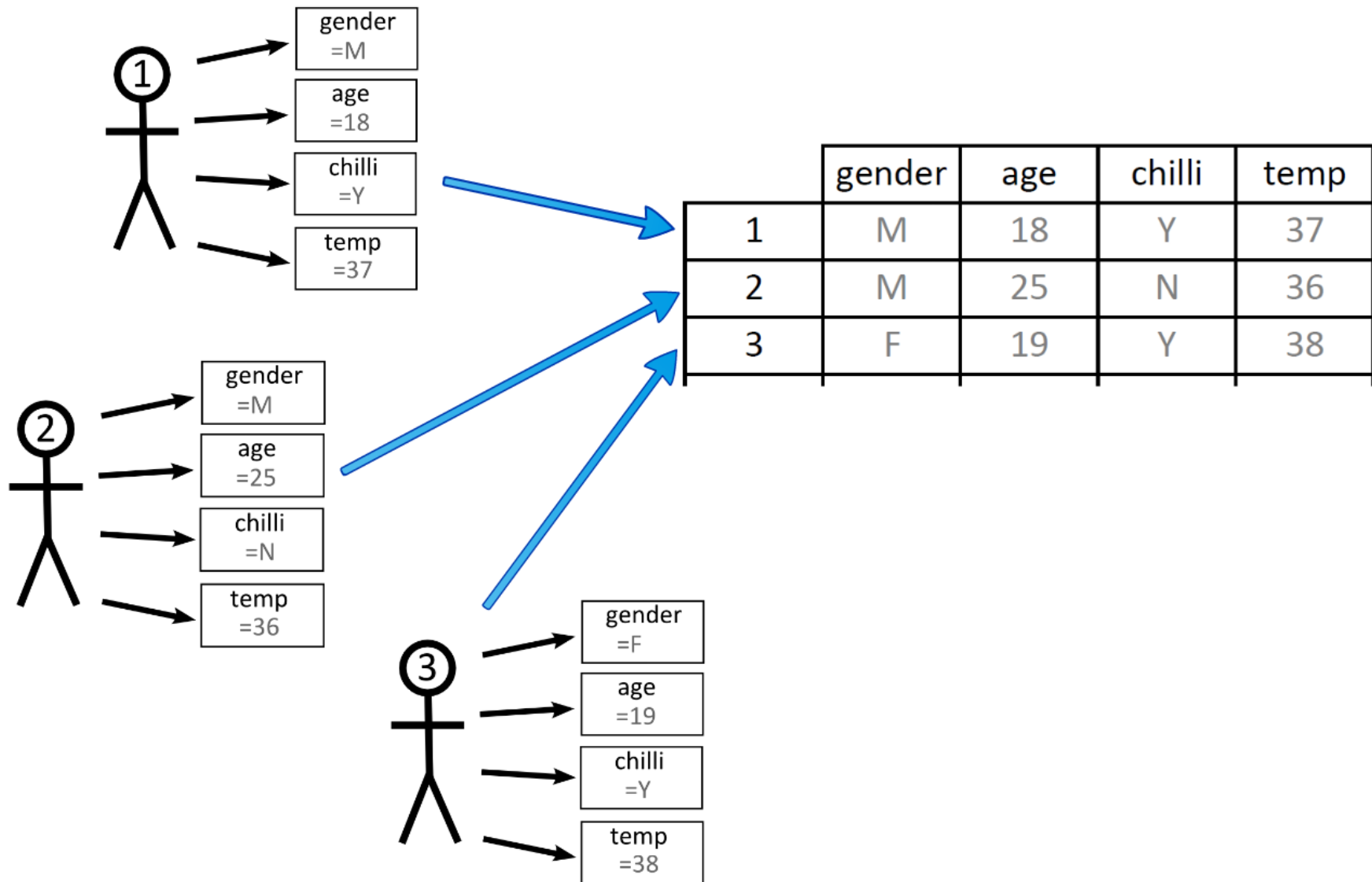
(words: how far apart has no meaning)

SUBJECTS are sources of VARIABLES

Variables can be used to calculate variables

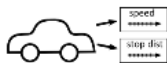


INFO about subjects goes in ROWS



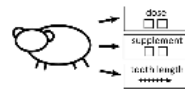
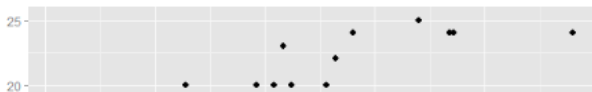
SCATTERPLOT

- Subjects/rows are points
- Info represented by what the point lines up with on the axes...
... and by colour, shape etc
- Aim to see the variation, and see past the variation.



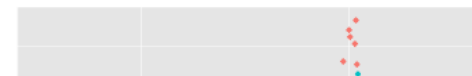
	speed	dist
1	4	2

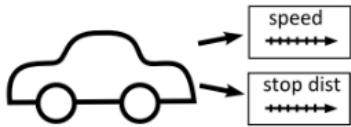
e



	len	supp	dose
1	4.2	VC	low

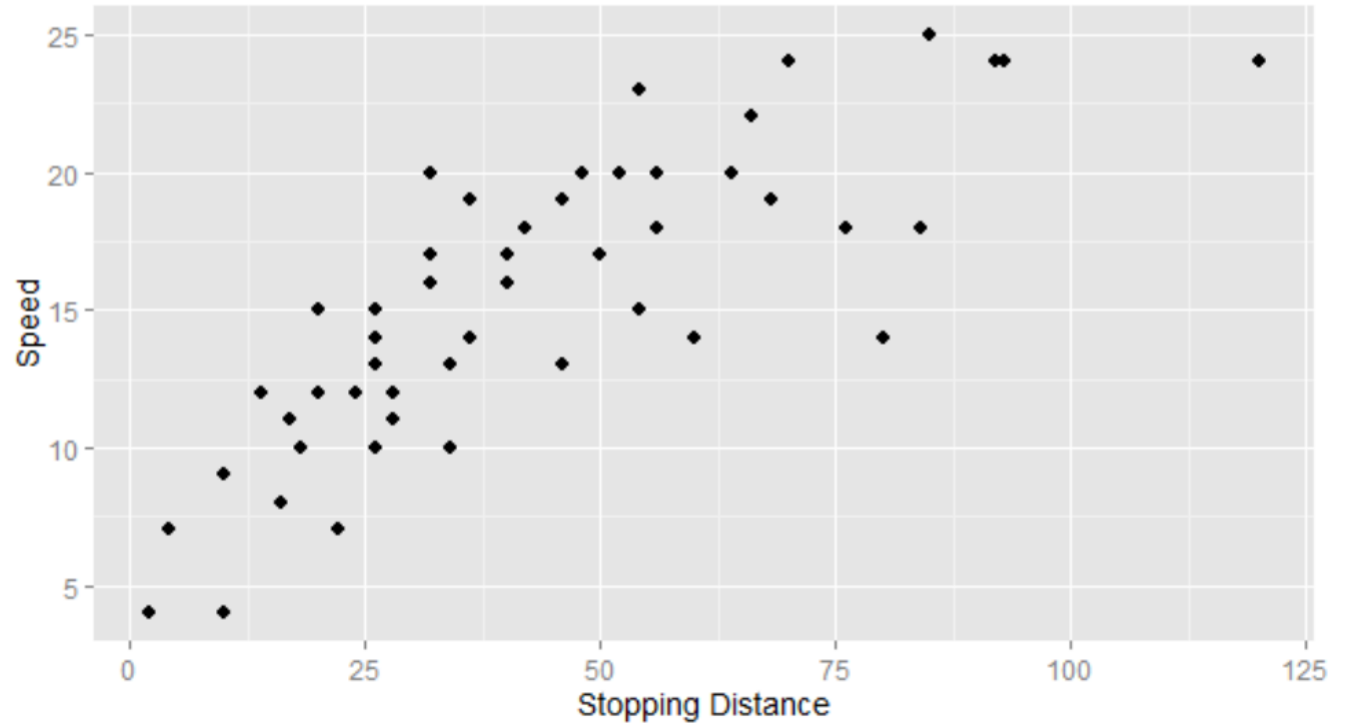
Jittered scatterplot





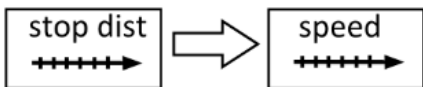
	speed	dist
1	4	2
2	4	10
3	7	4
4	7	22
5	8	16
6	9	10
7	10	18
8	10	26
9	10	34
10	11	17
11	11	28

outcome
also known as "dependent"
or "response"



explanatory

also known as "independent"

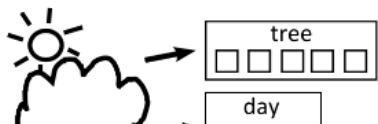


e

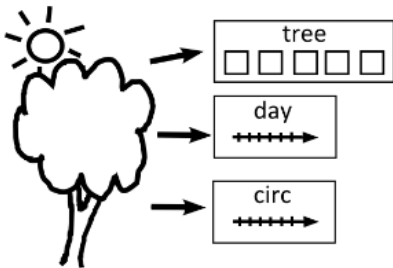
littered scatterplot

LINE GRAPH

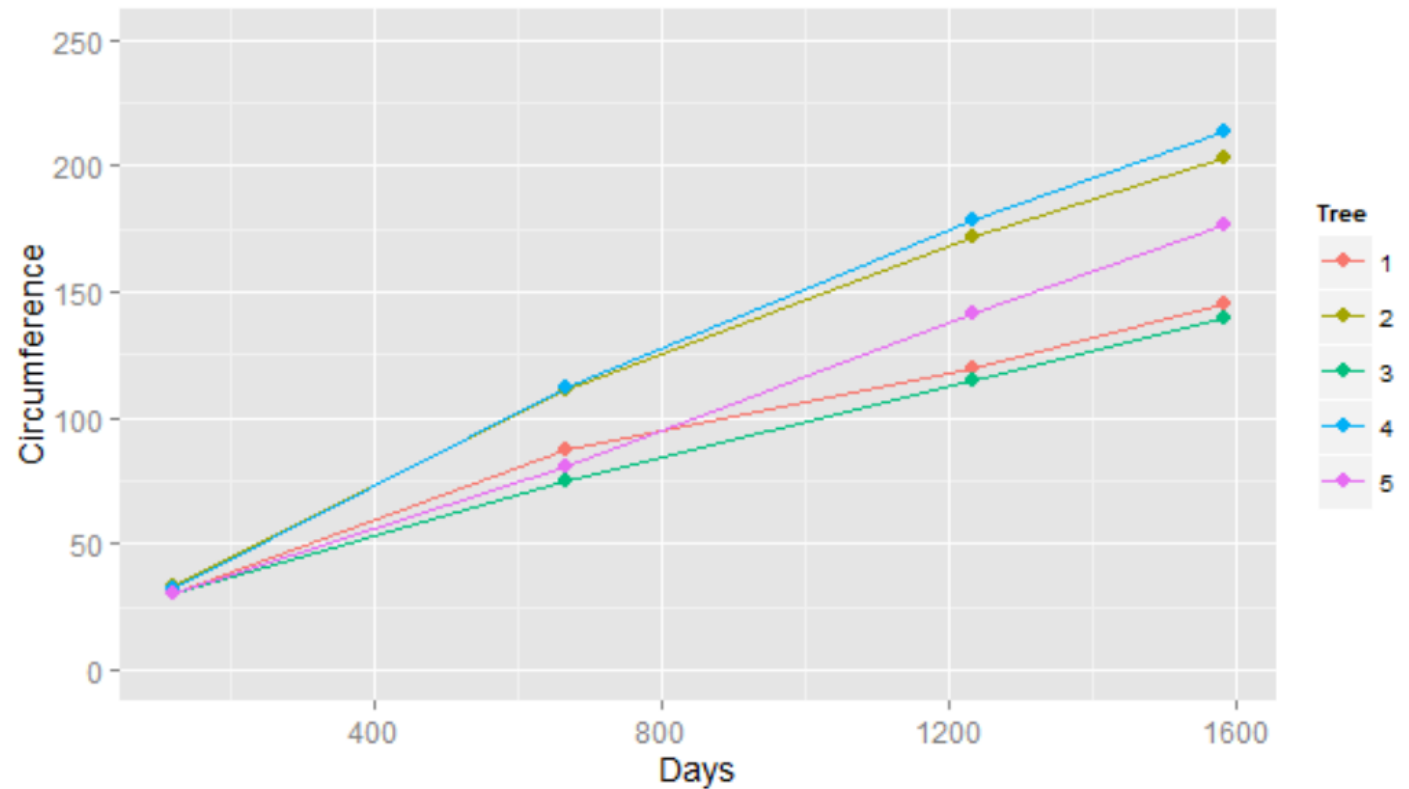
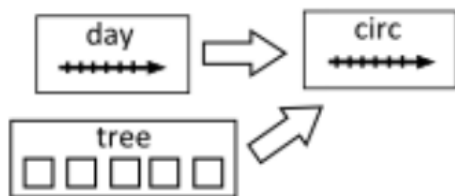
- Subjects are represented on the graph multiple times
- Info represented by what the point lines up with on the axes
- Points for the same subject are connected by lines



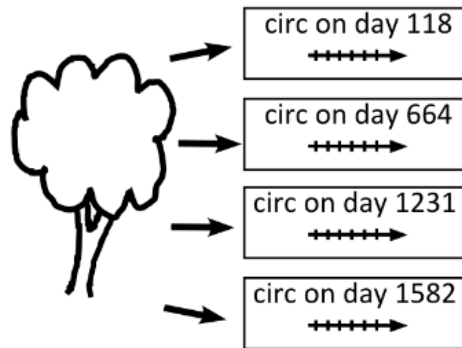
connected by lines



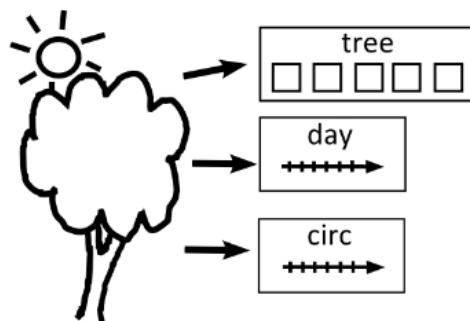
	Tree	days	circ
1	1	118	30
2	1	664	87
3	1	1231	120
4	1	1582	145
5	2	118	33
6	2	664	111
7	2	1231	172
8	2	1582	203
9	3	118	30
10	3	664	75
11	3	1231	115



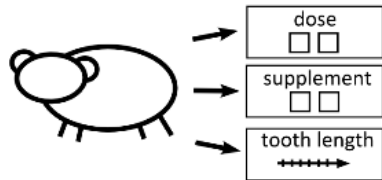
Repeated measurements: rethink your "subjects"



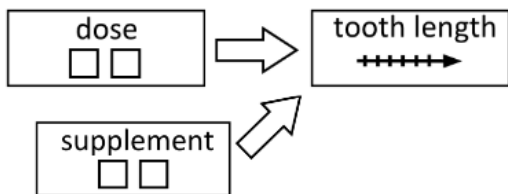
	Tree	circ.118	circ.664	circ.1231	circ.1582
1	1	30	87	120	145
2	2	33	111	172	203
3	3	30	75	115	140
4	4	32	112	179	214
5	5	30	81	142	177



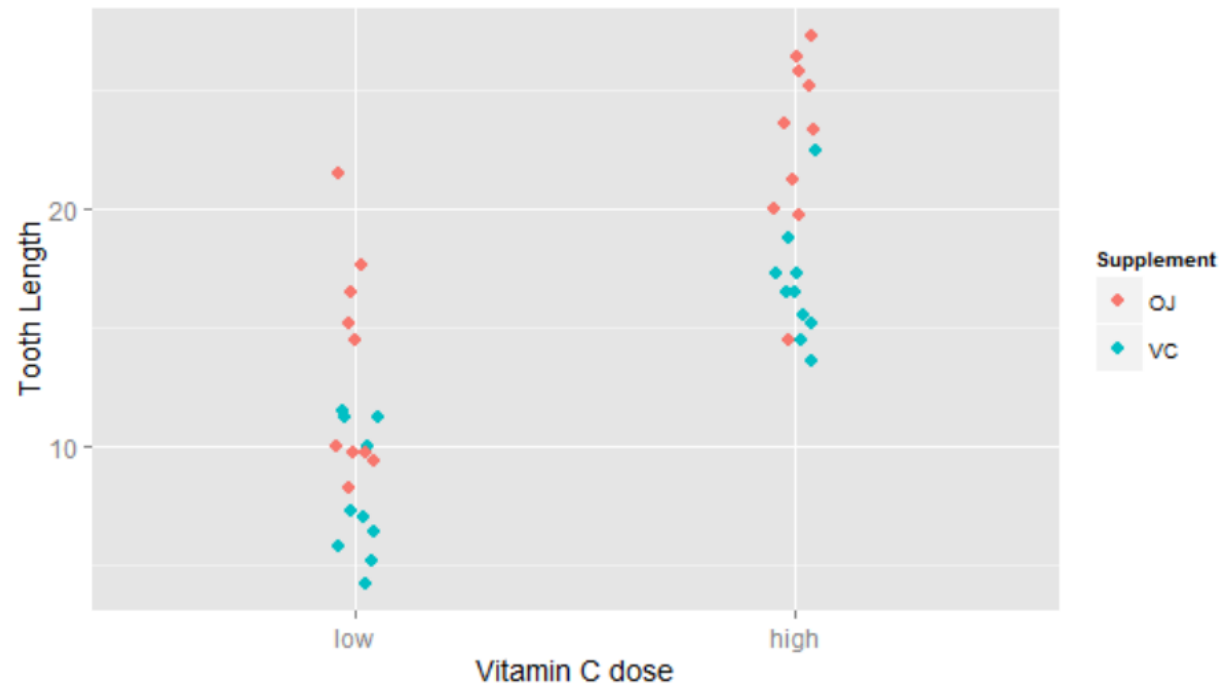
	Tree	days	circ
1	1	118	30
2	1	664	87
3	1	1231	120
4	1	1582	145
5	2	118	33
6	2	664	111
7	2	1231	172
8	2	1582	203
9	3	118	30
10	3	664	75
11	3	1231	115



	len	supp	dose
1	4.2	VC	low
2	11.5	VC	low
3	7.3	VC	low
4	5.8	VC	low
5	6.4	VC	low
6	10.0	VC	low
7	11.2	VC	low
8	11.2	VC	low
9	5.2	VC	low
10	7.0	VC	low
11	16.5	VC	high



Jittered scatterplot

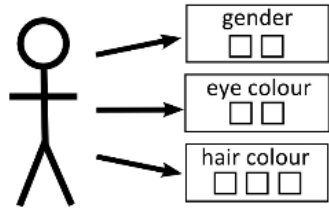


CATEGORICAL explanatory



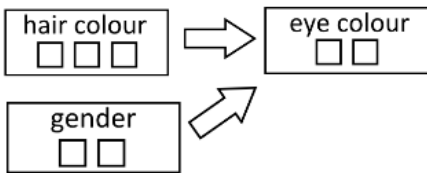
explanatory

also known as "independent"



	Hair	Eye	Gender
1	Dark	Brown	Male
2	Dark	Blue/Green/Hazel	Female
3	Dark	Blue/Green/Hazel	Male
4	Dark	Brown	Female
5	Dark	Blue/Green/Hazel	Male
6	Dark	Blue/Green/Hazel	Male
7	Dark	Brown	Male
8	Dark	Blue/Green/Hazel	Male
9	Red	Blue/Green/Hazel	Female
10	Blond	Blue/Green/Hazel	Male
11	Dark	Brown	Female

CATEGORICAL outcome



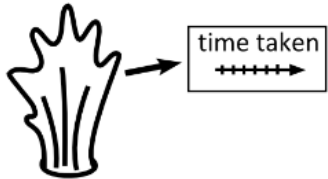
Jittered scatterplot



CATEGORICAL explanatory

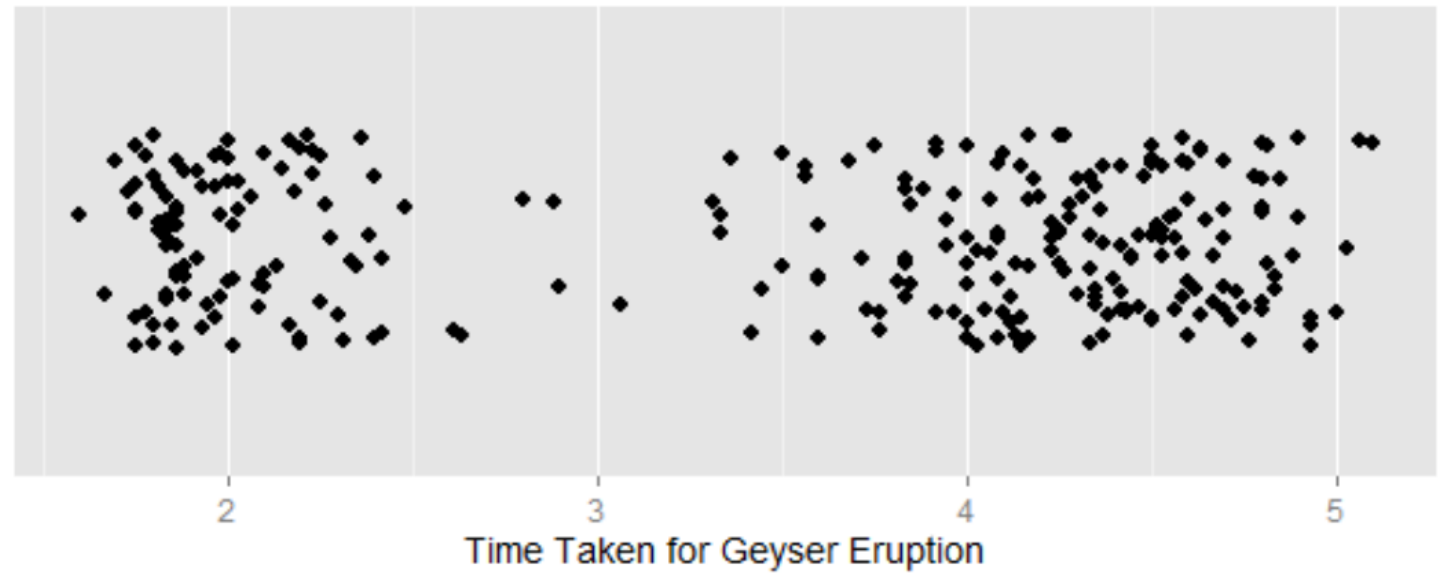
CATEGORICAL explanatory

supplement
□ □



	time
1	3.600
2	1.800
3	3.333
4	2.283
5	4.533
6	2.883
7	4.700
8	3.600
9	1.950
10	4.350
11	1.822

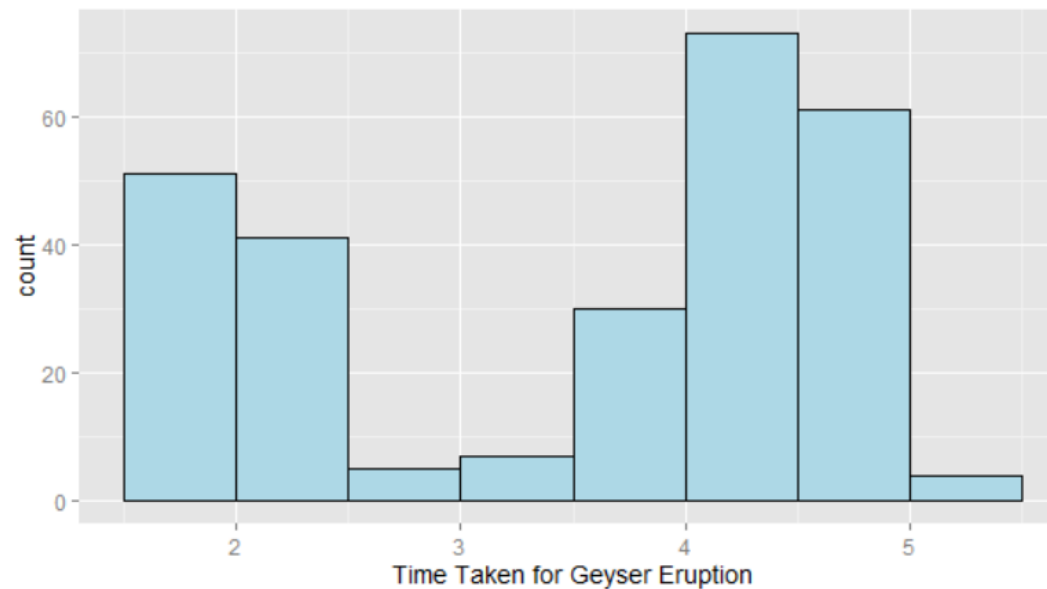
time taken
+++++→ ?



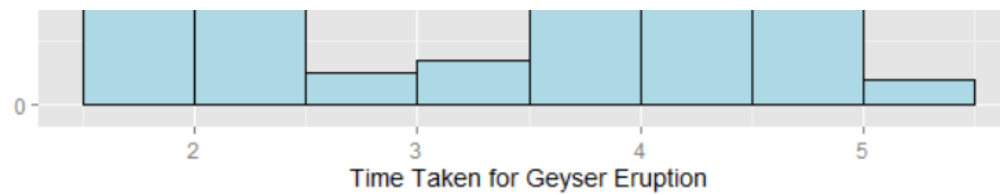
Only one variable

HISTOGRAM

Shows how much of the data is in each zone.

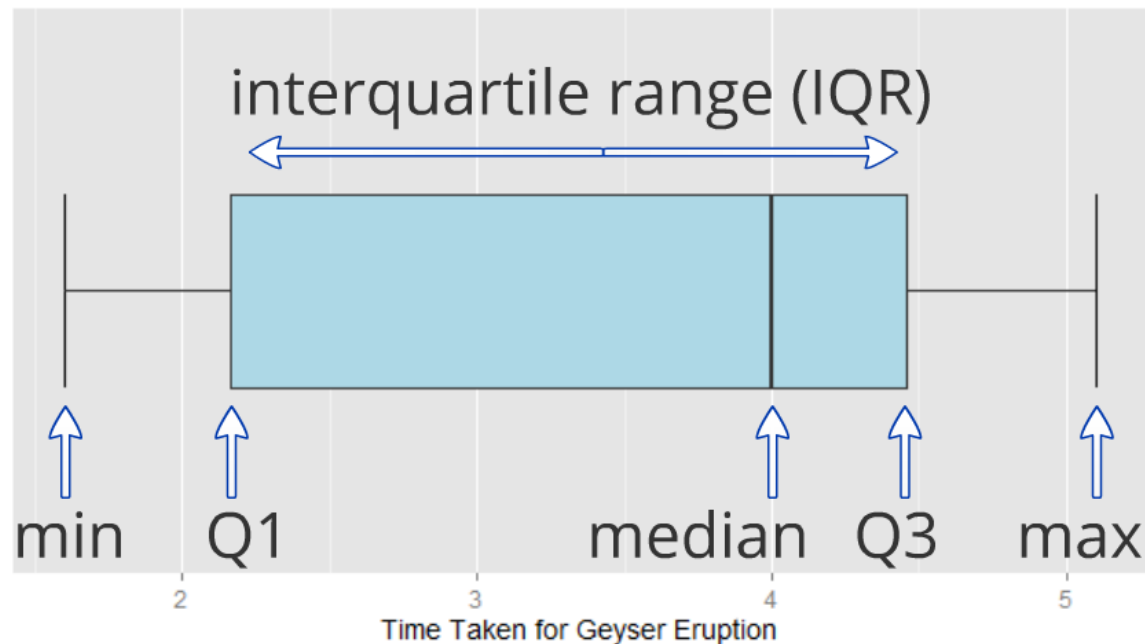


BOXPLOT



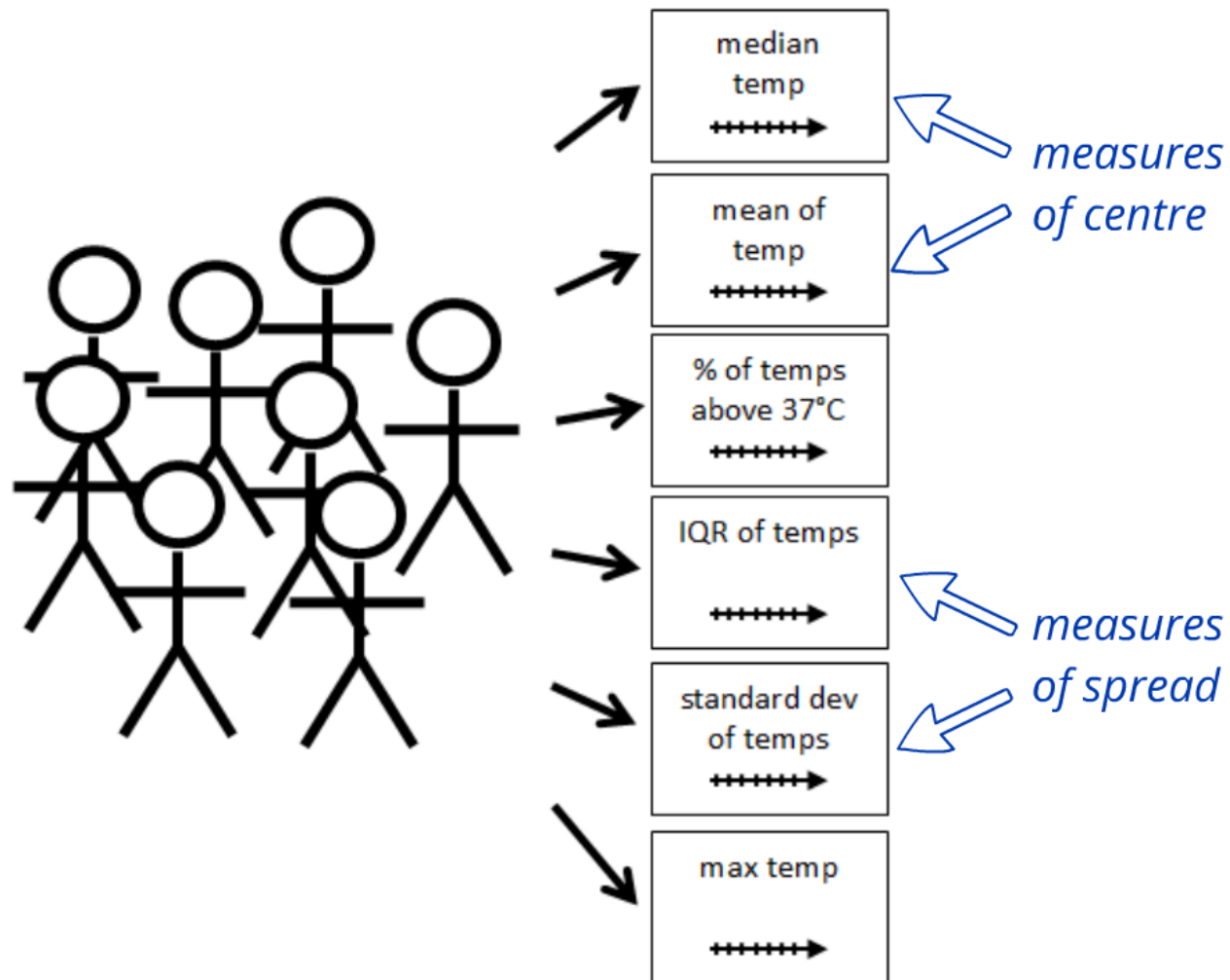
BOXPLOT

Shows how spread out each quarter of the data is.



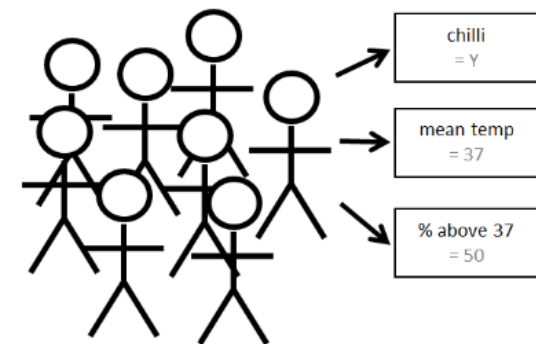
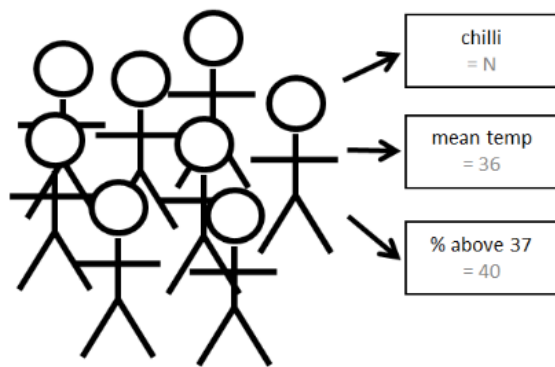
The individual subjects are not here!

VARIABLES can be recorded about a GROUP



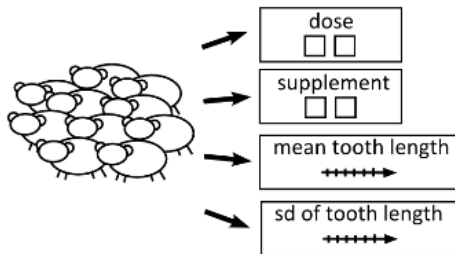
INFO about GROUPS goes in ROWS in a NEW DATASET

AGGREGATE DATA

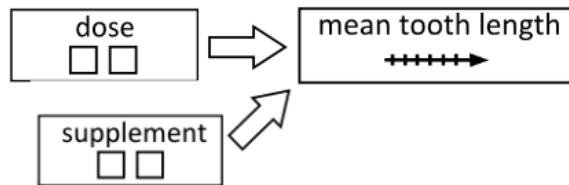


	chilli	mean temp	% above 37
1	Y	37	50
2	N	36	40

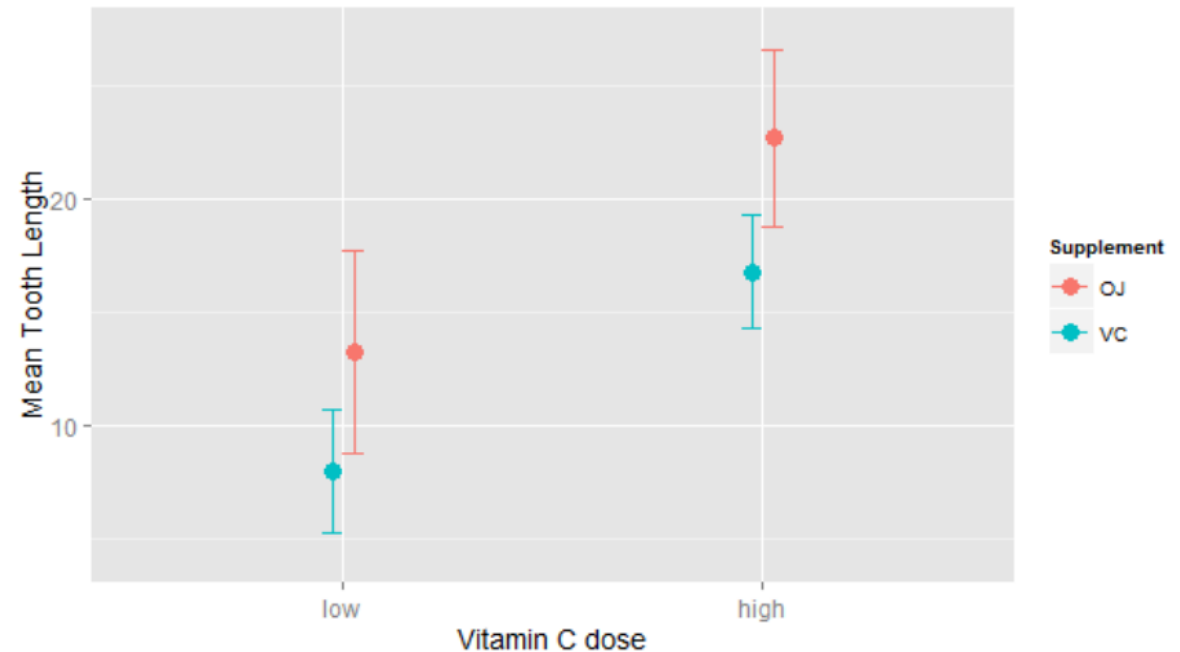
GRAPHS OF GROUP STATS



	supp	dose	mean.len	sd.len
1	OJ	low	13.23	4.459709
2	OJ	high	22.70	3.910953
3	VC	low	7.98	2.746634
4	VC	high	16.77	2.515309

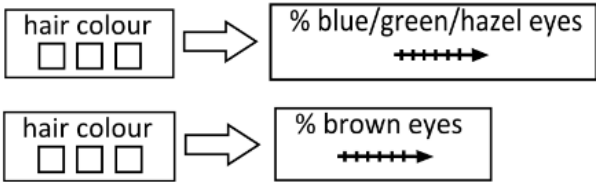


graph with errorbars

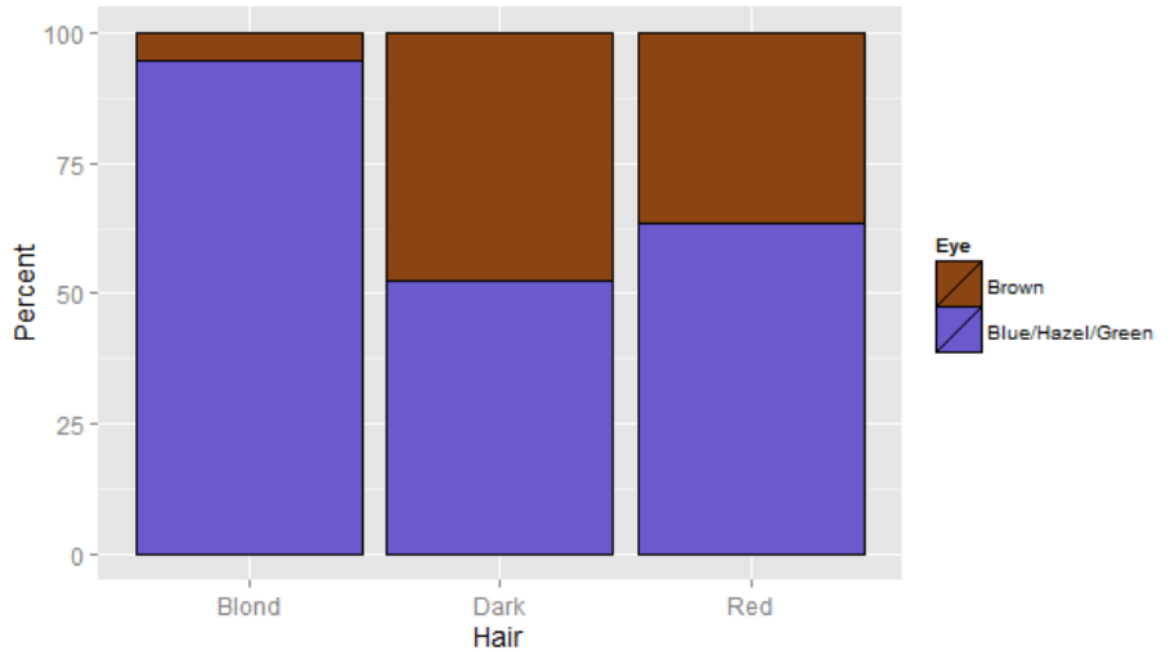




	Hair	percent.blue	percent.brown
1	Blond	94.48819	5.511811
2	Dark	52.53807	47.461929
3	Red	63.38028	36.619718



stacked bar chart



THE BIGGEST IDEA IN STATS

- Your subjects are just SOME of the subjects you COULD HAVE HAD.
- Your group of subjects is just ONE of the groups you COULD HAVE HAD.
- Any one value of a variable is just ONE of the values you COULD HAVE HAD -- even if it was calculated from other variables or recorded on the whole group.

The description of all the values that COULD HAVE BEEN and how likely they are is called the DISTRIBUTION.

EVERYTHING HAS A DISTRIBUTION

If you choose your variables right, you might even know what the distribution is called.

TS

jects is
roups
HAD.

0.

HAVE BEEN
UTION.

UTION

ight

.

GRAPHS OF DISTRIBUTIONS

Tend to look like smooth "histograms"

PROBABILITY DENSITY FUNCTIONS

