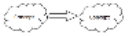
Maths Learning Centre
The University of Adelaide

Thinking About Data

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

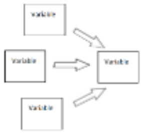
STATS helps to answer QUESTIONS
Questions are about CONCEPTS

About a single concept



About the relationship
between concepts

CONCEPTS become VARIABLES
when you measure them



VARIABLES come in two TYPES

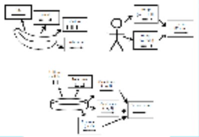
variable
name

Numerical
also known as:
Quantitative, Interval, Scale
(numbers: how far apart has meaning)

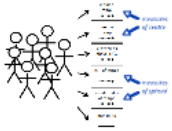
variable
name

Categorical
also known as:
Qualitative
(words: how far apart has no meaning)

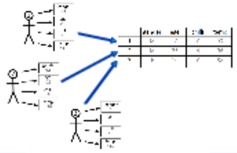
SUBJECTS are sources of VARIABLES
Variables can be used to calculate variables



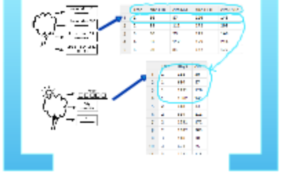
VARIABLES can be recorded about a GROUP



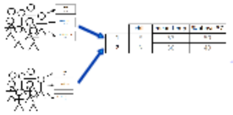
INFO about subjects goes in ROWS



Repeated measurements: rethink your "subjects"

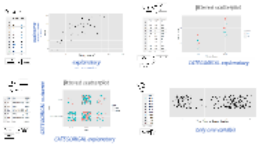


INFO about GROUPS goes in ROWS in a NEW DATASET
AGGREGATE DATA



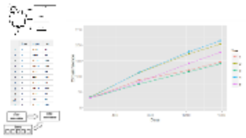
SCATTERPLOT

- 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.



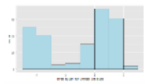
LINE GRAPH

- 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



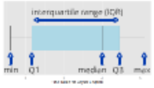
HISTOGRAM

Shows how much of the data is in each zone.



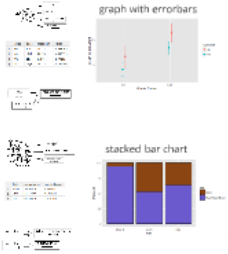
BOXPLOT

Shows how spread out each quarter of the data is.



The individual subjects are not here!

GRAPHS OF GROUP STATS



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.

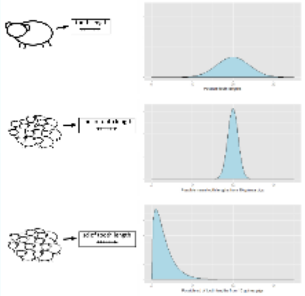
The description of all the values that COULD HAVE BEEN and how likely they are is called the DISTRIBUTION.

EVERYTHING HAS A DISTRIBUTION

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

GRAPHS OF DISTRIBUTIONS
Tend to look like smooth "histograms"

PROBABILITY DENSITY FUNCTIONS



Maths Learning Centre

The University of Adelaide

Thinking About Data

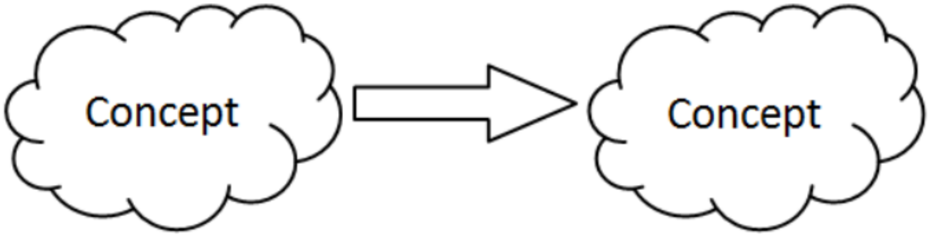
How statisticians think about data, and how they make graphs and summaries.

VARIABLES come in two TYPES

STATS helps to answer QUESTIONS

Questions are about CONCEPTS

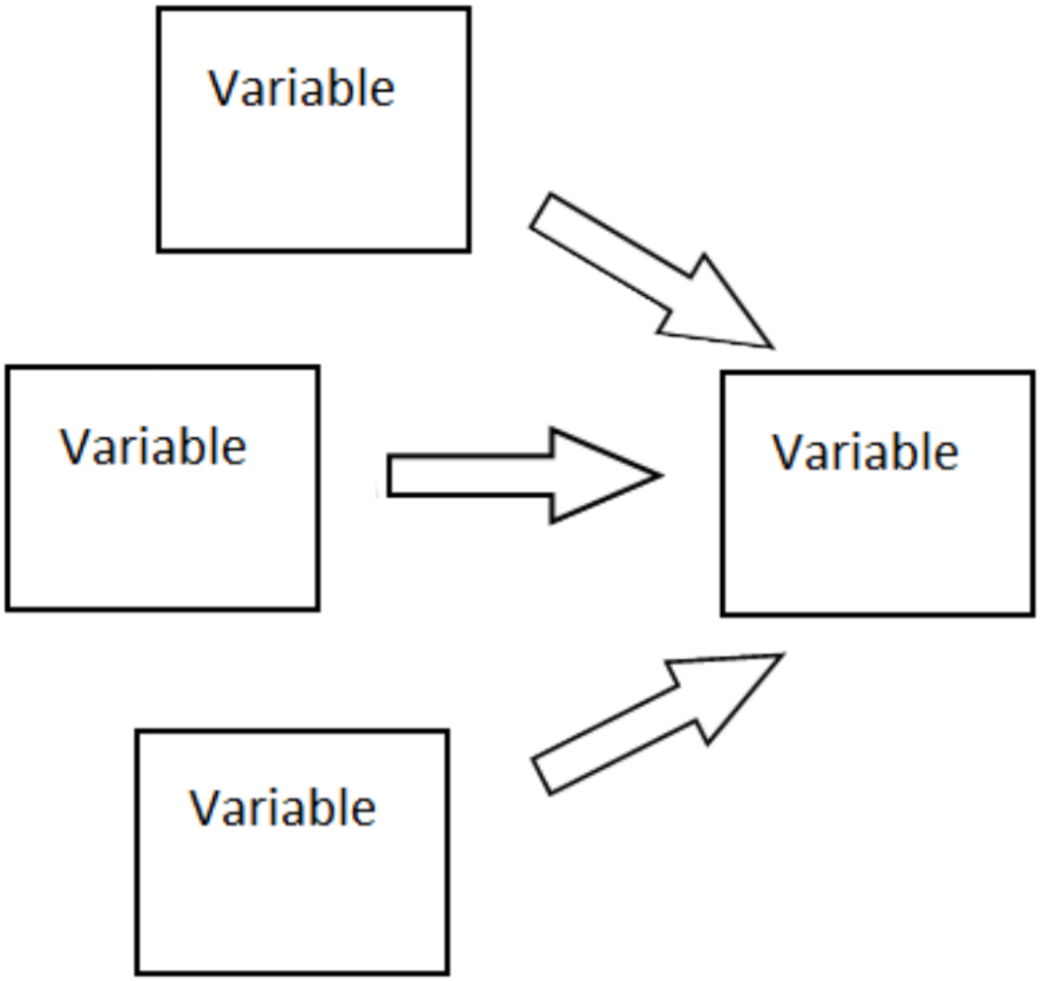
About a single concept



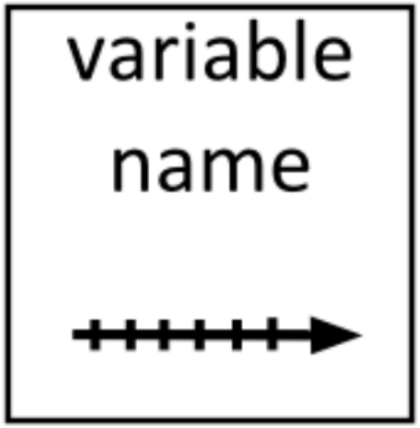
About the relationship between concepts

CONCEPTS become VARIABLES

when you measure them



VARIABLES come in two TYPES

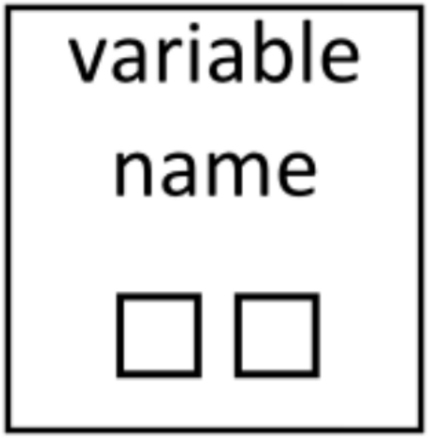


Numerical

also known as:

Quantitative, Interval, Scale

(numbers: how far apart has meaning)



Categorical

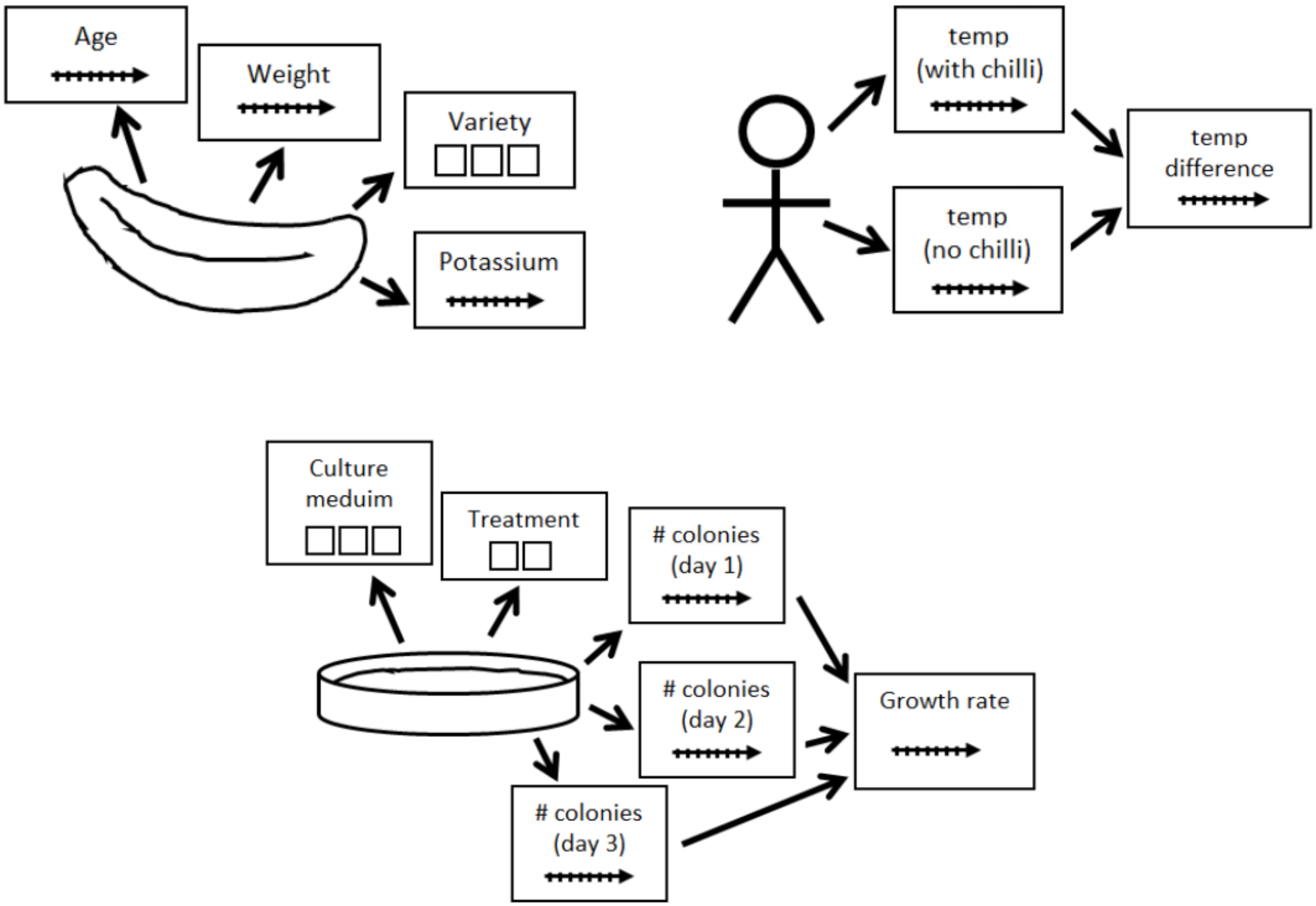
also known as:

Qualitative

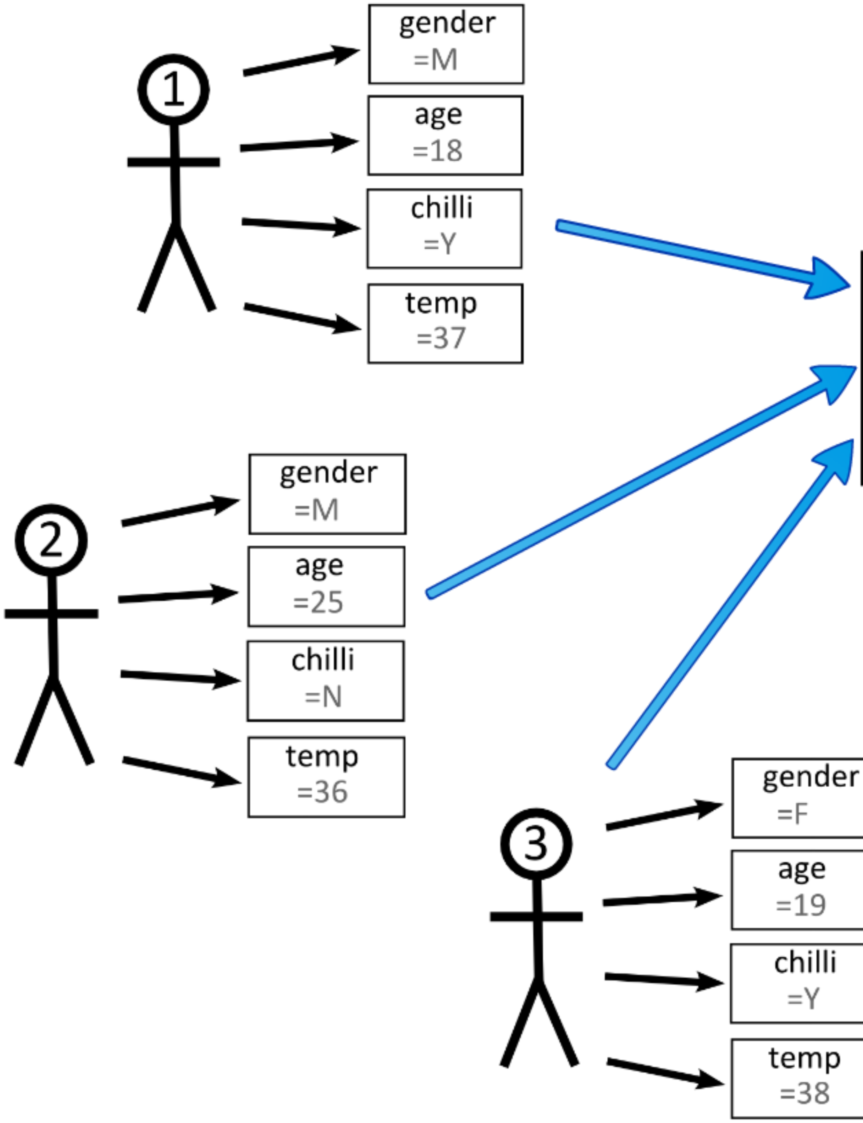
(words: how far apart has no meaning)

SUBJECTS are sources of VARIABLES

Variables can be used to calculate variables



INFO about subjects goes in ROWS



gender =M
age =18
chilli =Y
temp =37

gender =M
age =25
chilli =N
temp =36

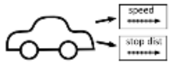
gender =F
age =19
chilli =Y
temp =38

	gender	age	chilli	temp
1	M	18	Y	37
2	M	25	N	36
3	F	19	Y	38

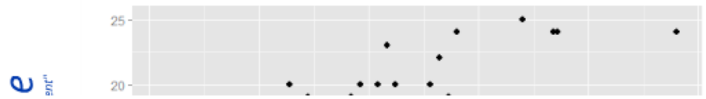
SCATTERPLOT

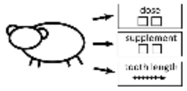
- Subjects/rows are points
- Info represented by what the point lines up with on the axes…

  … and by colour, shape etc
- Aim to see the variation, and see past the variation.



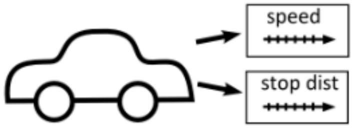
	speed	dist
1	4	2

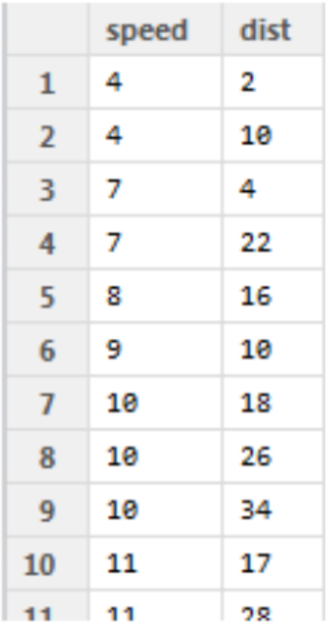




	len	supp	dose
1	4.2	VC	low

Jittered scatterplot

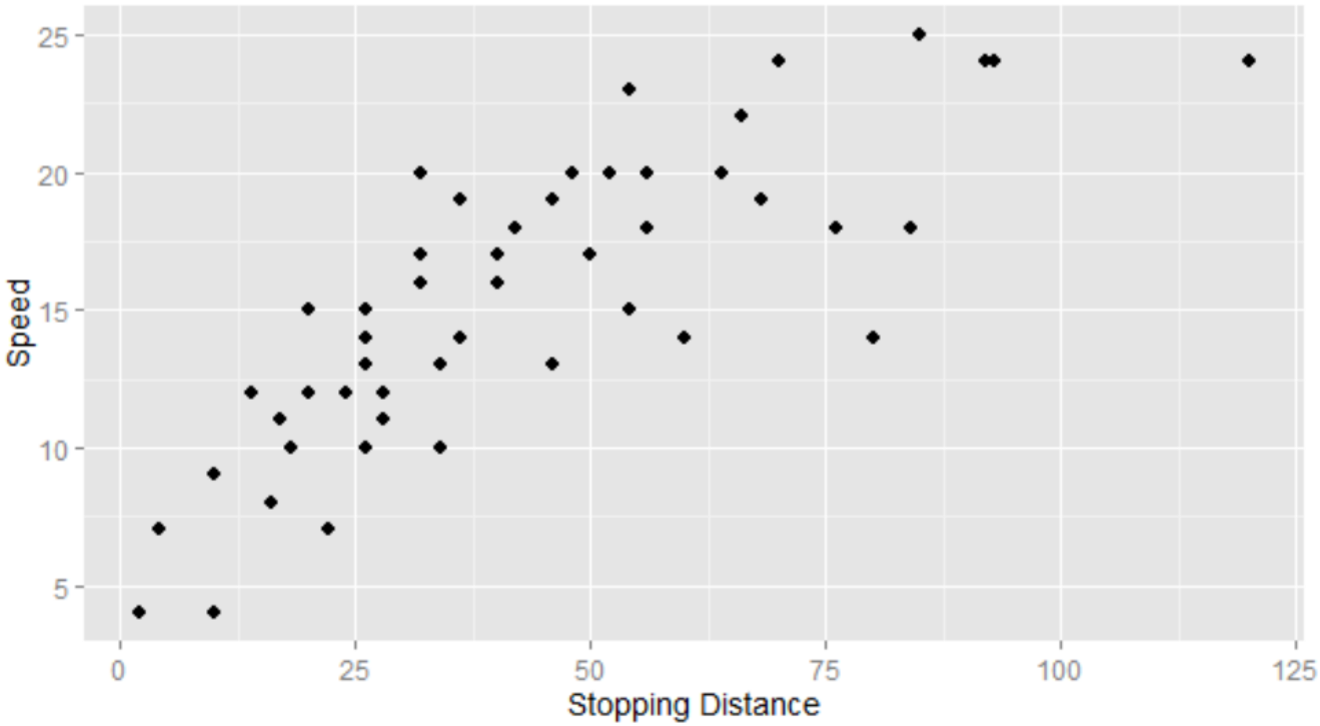




	speed	dist
1	4	2
2	4	10
3	7	4
4	7	22
5	8	16
6	9	10
7	10	18
8	10	26
9	10	34
10	11	17
11	11	28

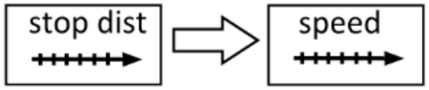
outcome

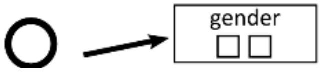
also known as "dependent" or "response"



explanatory

also known as "independent"



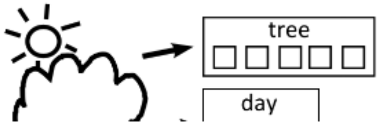


e

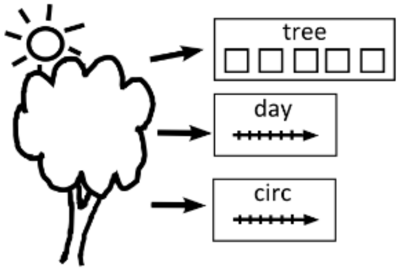
Jittered scatterplot

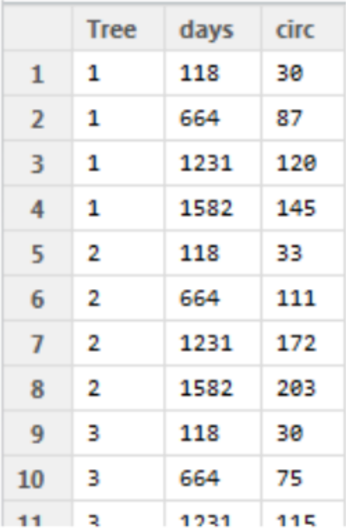
LINE GRAPH

- 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

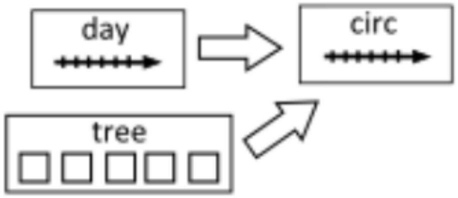


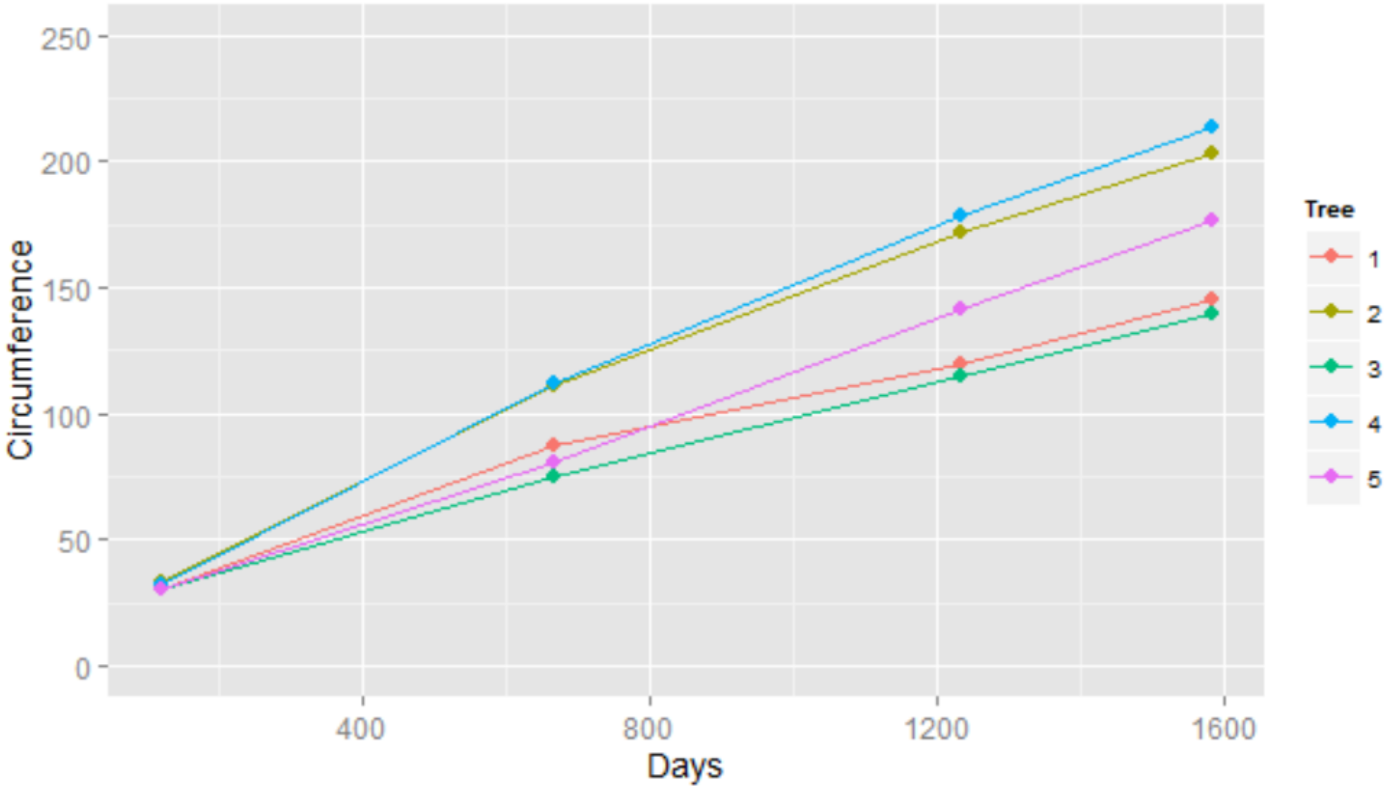
connected by lines



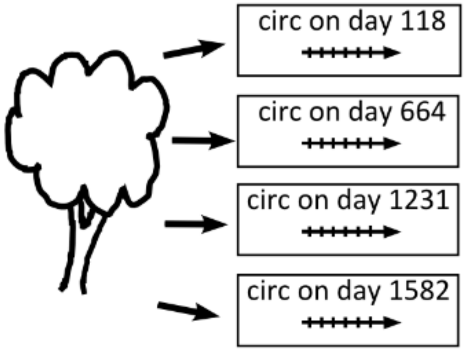


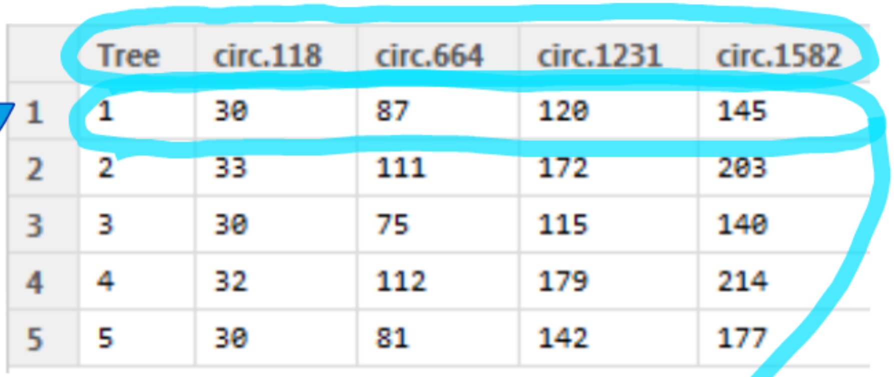
	Tree	days	circ
1	1	118	30
2	1	664	87
3	1	1231	120
4	1	1582	145
5	2	118	33
6	2	664	111
7	2	1231	172
8	2	1582	203
9	3	118	30
10	3	664	75
11	3	1231	115



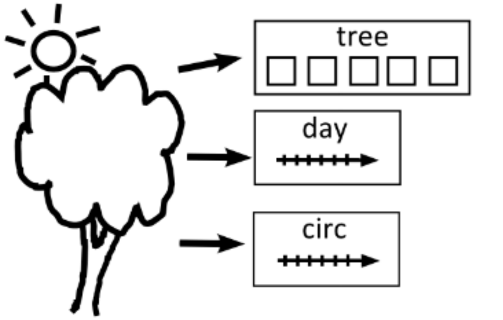


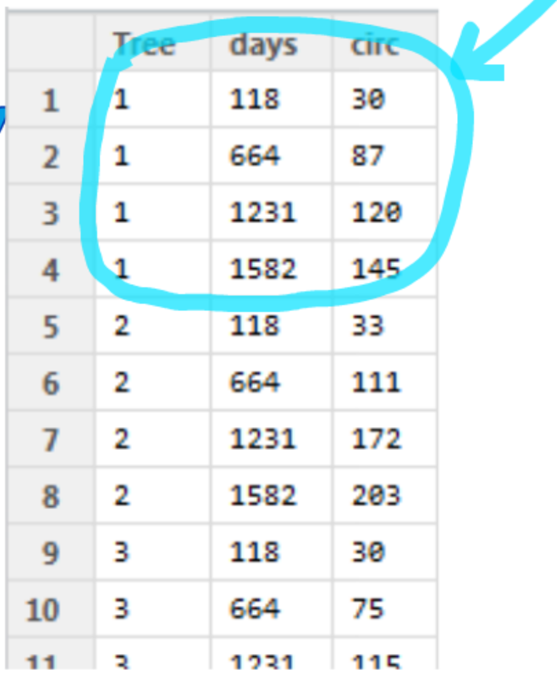
Repeated measurements: rethink your "subjects"





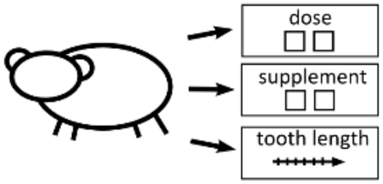
	Tree	circ.118	circ.664	circ.1231	circ.1582
1	1	30	87	120	145
2	2	33	111	172	203
3	3	30	75	115	140
4	4	32	112	179	214
5	5	30	81	142	177

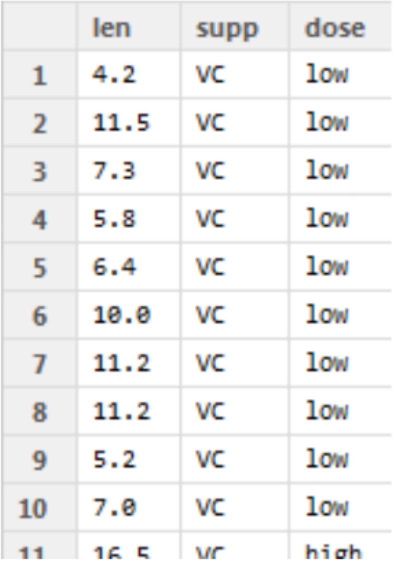




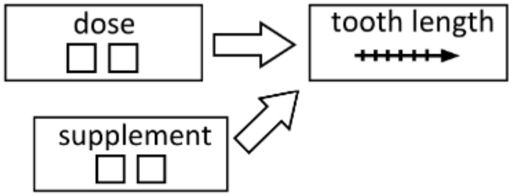
	Tree	days	circ
1	1	118	30
2	1	664	87
3	1	1231	120
4	1	1582	145
5	2	118	33
6	2	664	111
7	2	1231	172
8	2	1582	203
9	3	118	30
10	3	664	75
11	3	1231	115

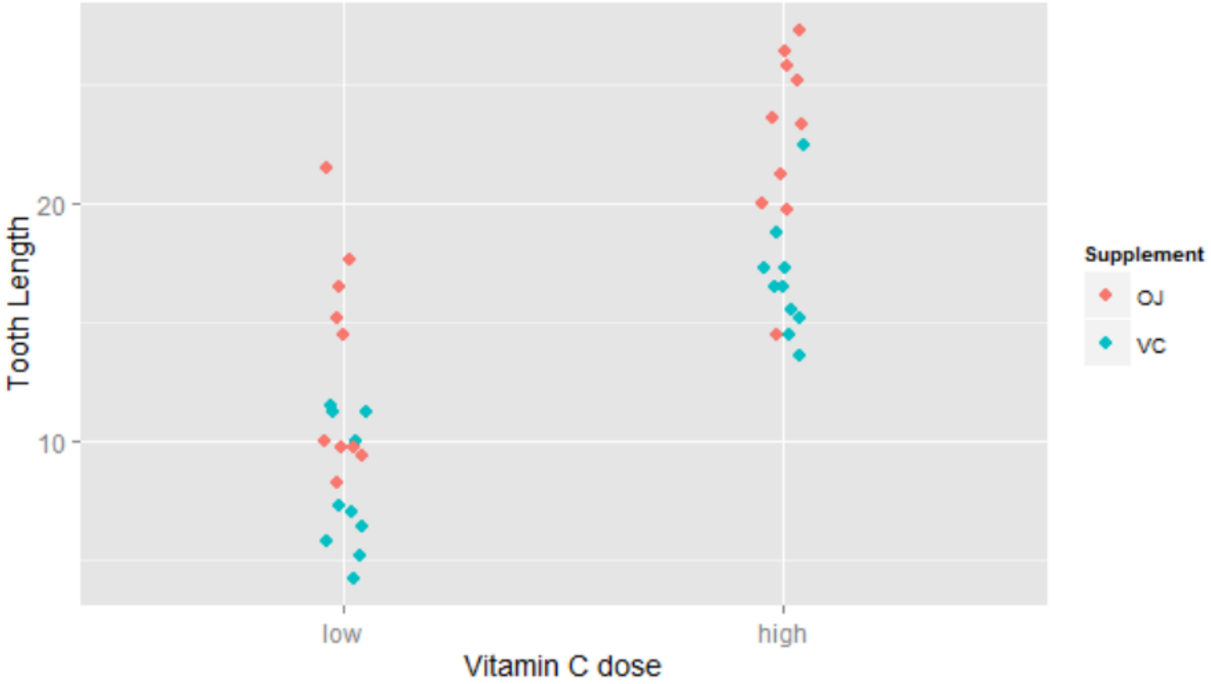
Jittered scatterplot



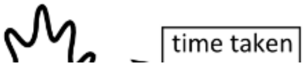


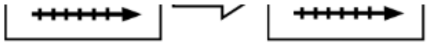
	len	supp	dose
1	4.2	VC	low
2	11.5	VC	low
3	7.3	VC	low
4	5.8	VC	low
5	6.4	VC	low
6	10.0	VC	low
7	11.2	VC	low
8	11.2	VC	low
9	5.2	VC	low
10	7.0	VC	low
11	16.5	VC	high





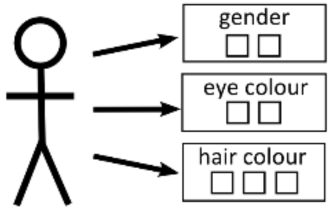
CATEGORICAL explanatory

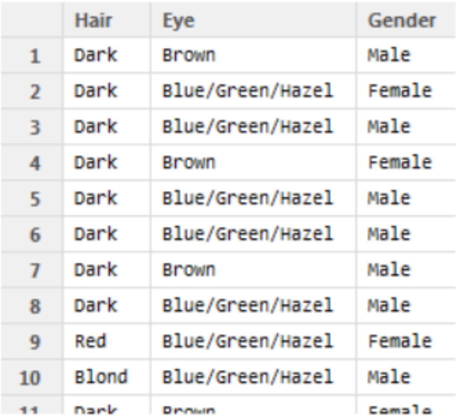




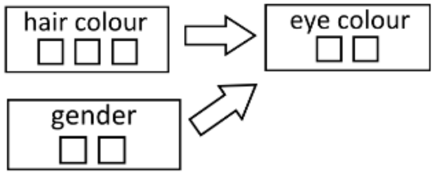
explanatory

also known as "independent"



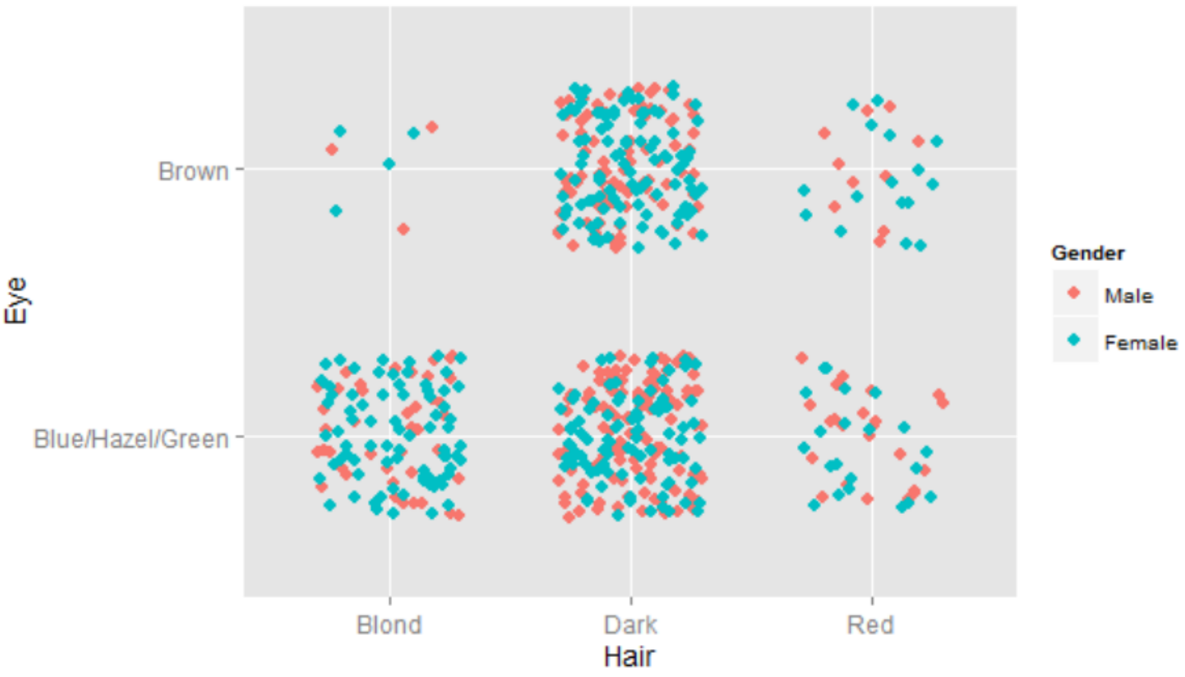


	Hair	Eye	Gender
1	Dark	Brown	Male
2	Dark	Blue/Green/Hazel	Female
3	Dark	Blue/Green/Hazel	Male
4	Dark	Brown	Female
5	Dark	Blue/Green/Hazel	Male
6	Dark	Blue/Green/Hazel	Male
7	Dark	Brown	Male
8	Dark	Blue/Green/Hazel	Male
9	Red	Blue/Green/Hazel	Female
10	Blond	Blue/Green/Hazel	Male
11	Dark	Brown	Female



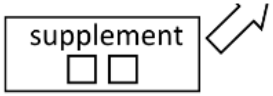
CATEGORICAL outcome

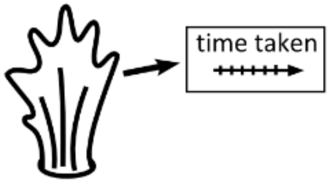
Jittered scatterplot

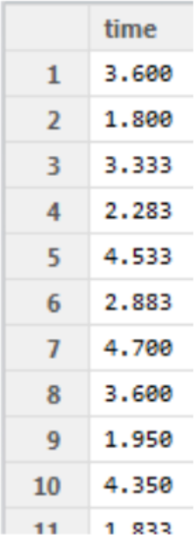


CATEGORIGAL explanatory

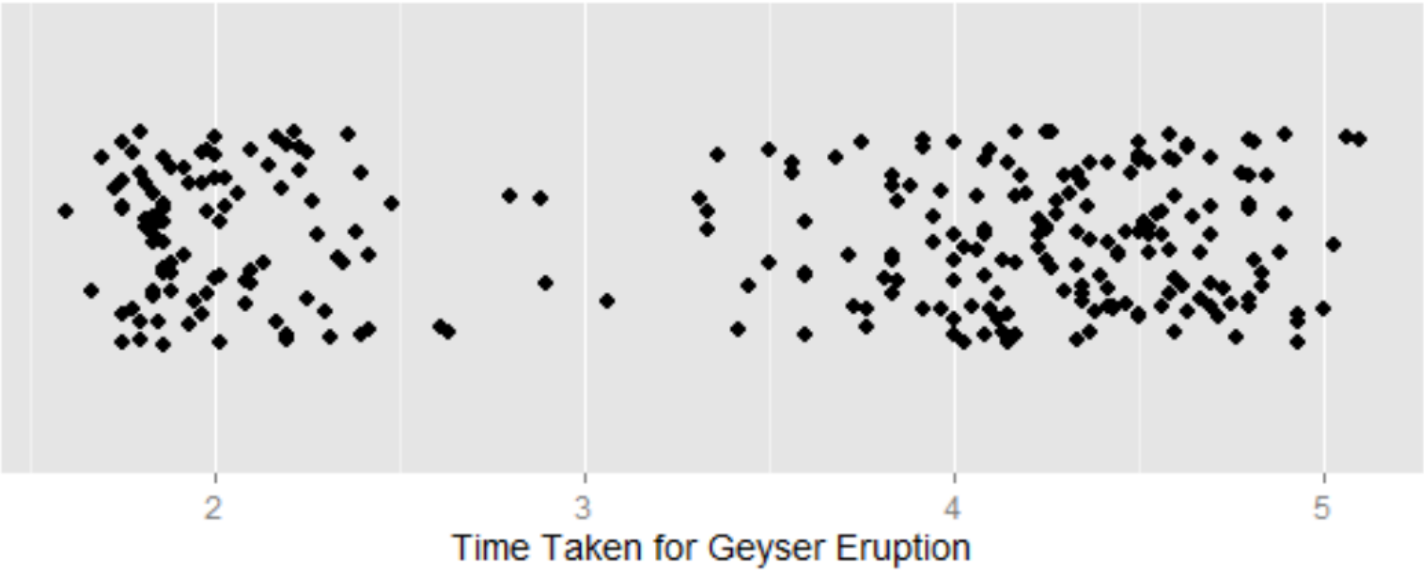
CATEGORICAL explanatory

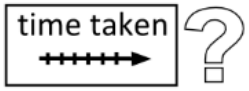






	time
1	3.600
2	1.800
3	3.333
4	2.283
5	4.533
6	2.883
7	4.700
8	3.600
9	1.950
10	4.350
11	1.833

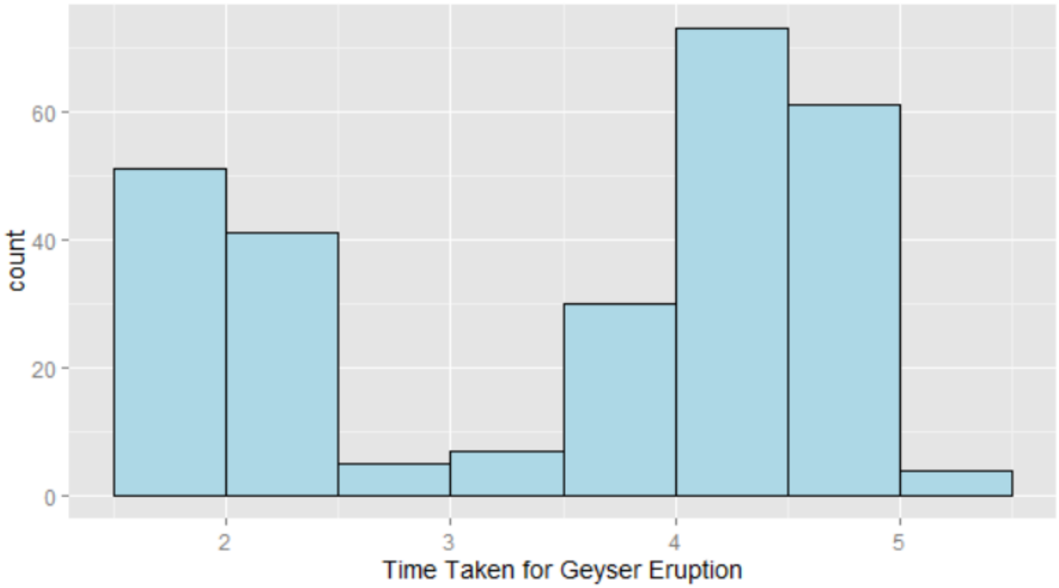




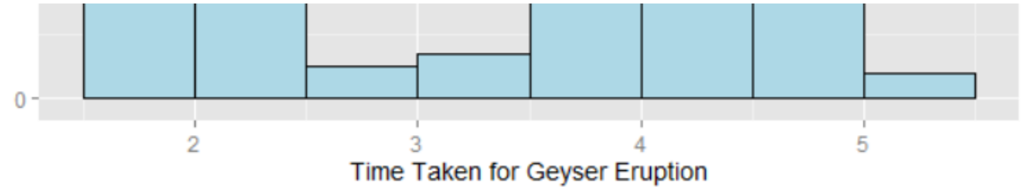
Only one variable

HISTOGRAM

Shows how much of the data is in each zone.

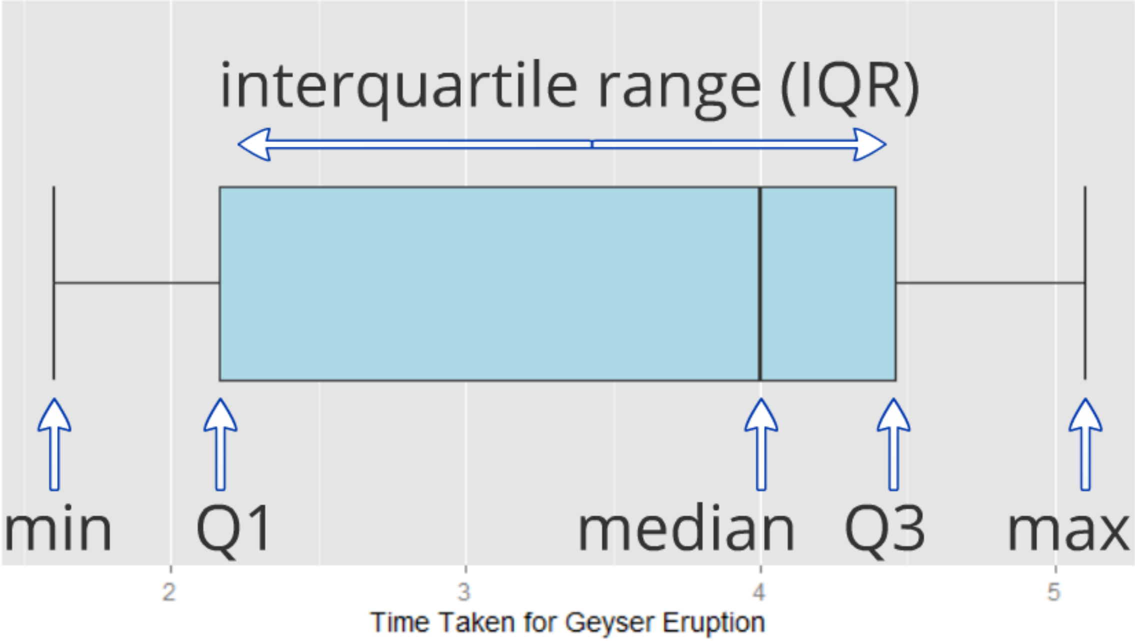


BOXPLOT



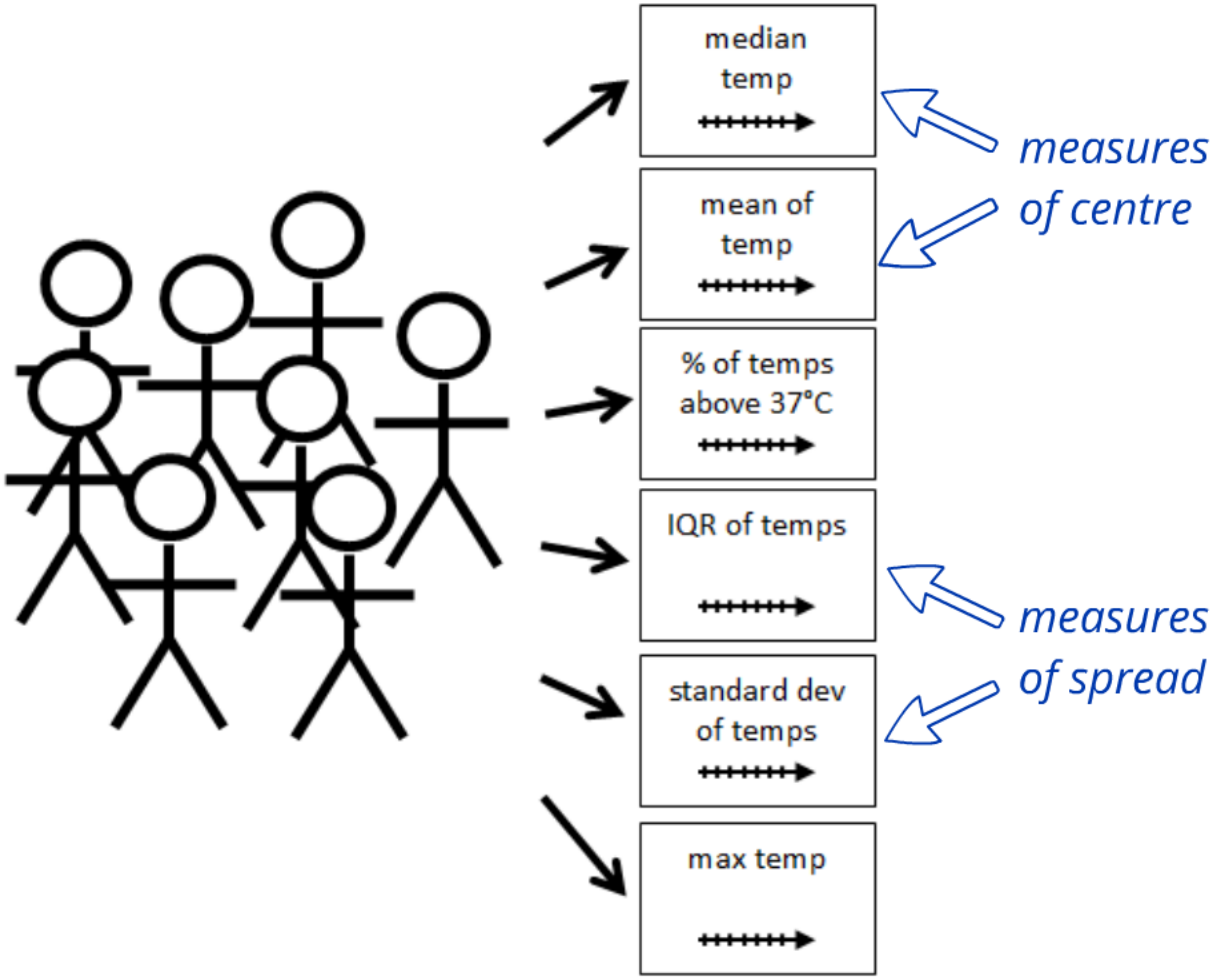
BOXPLOT

Shows how spread out each quarter of the data is.



The individual subjects are not here!

VARIABLES can be recorded about a GROUP



median temp

mean of temp

% of temps above 37°C

IQR of temps

standard dev of temps

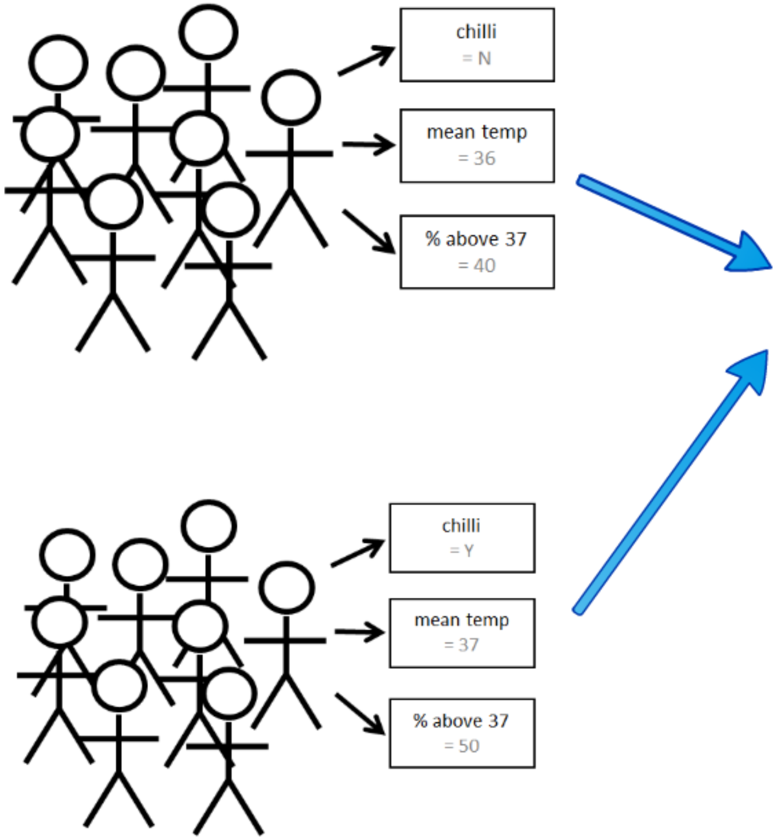
max temp

measures of centre

measures of spread

INFO about GROUPS goes in ROWS in a NEW DATASET

AGGREGATE DATA



chilli = N

mean temp = 36

% above 37 = 40

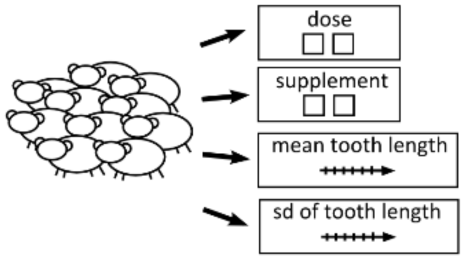
chilli = Y

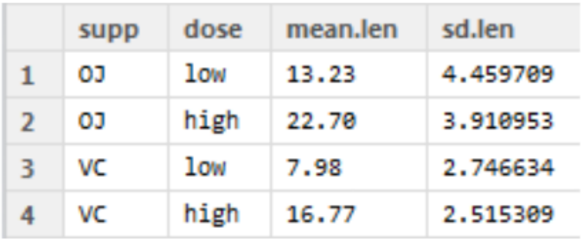
mean temp = 37

% above 37 = 50

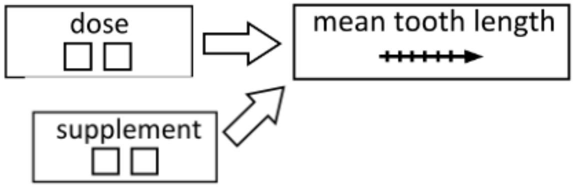
	chilli	mean temp	% above 37
1	Y	37	50
2	N	36	40

GRAPHS OF GROUP STATS

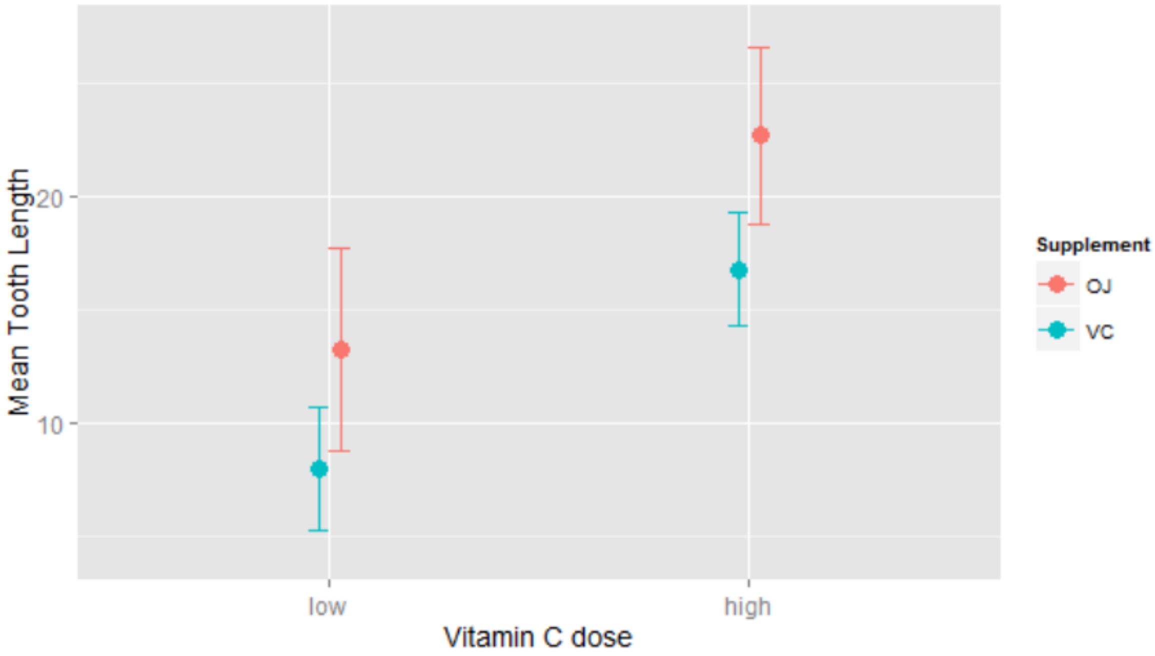




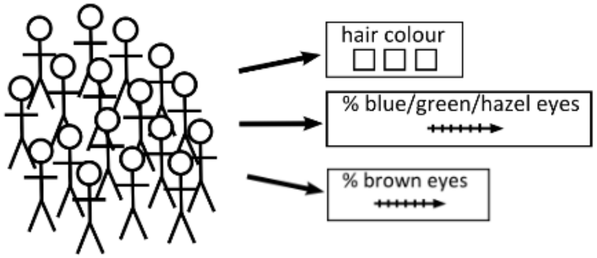
	supp	dose	mean.len	sd.len
1	OJ	low	13.23	4.459709
2	OJ	high	22.70	3.910953
3	VC	low	7.98	2.746634
4	VC	high	16.77	2.515309

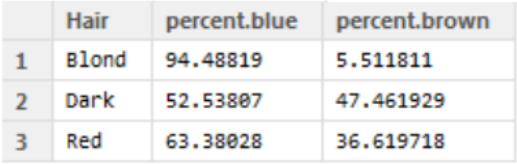


graph with errorbars

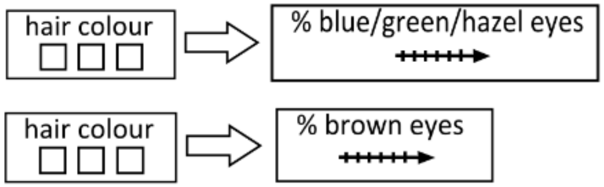


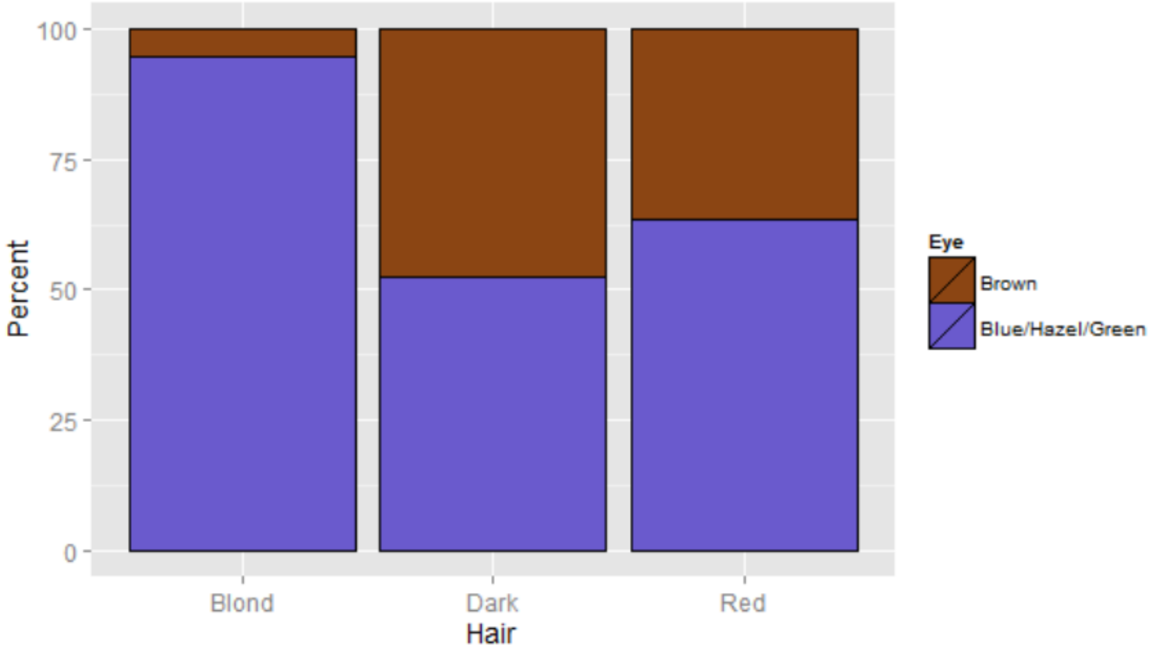
stacked bar chart





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





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
E HAD.
F

o.

*HAVE BEEN
UTION.*

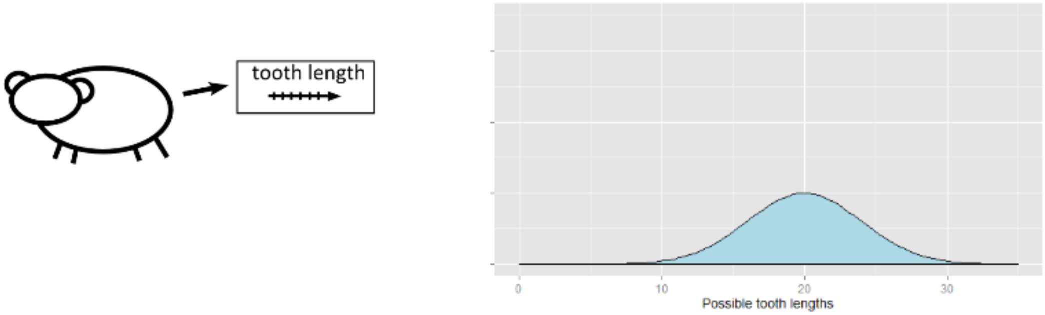
UTION

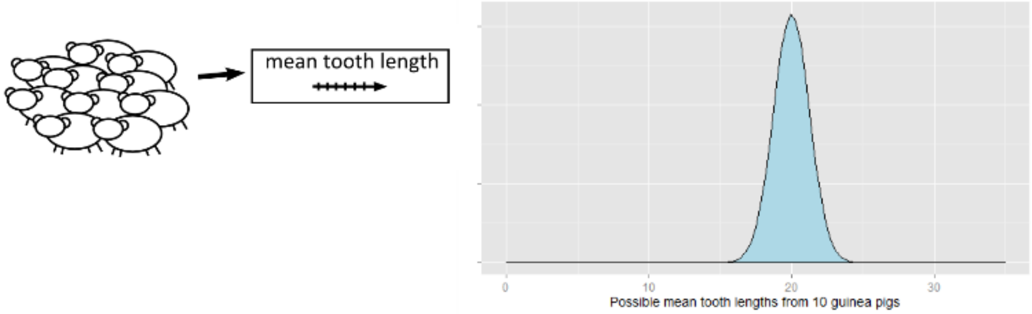
ight
l.

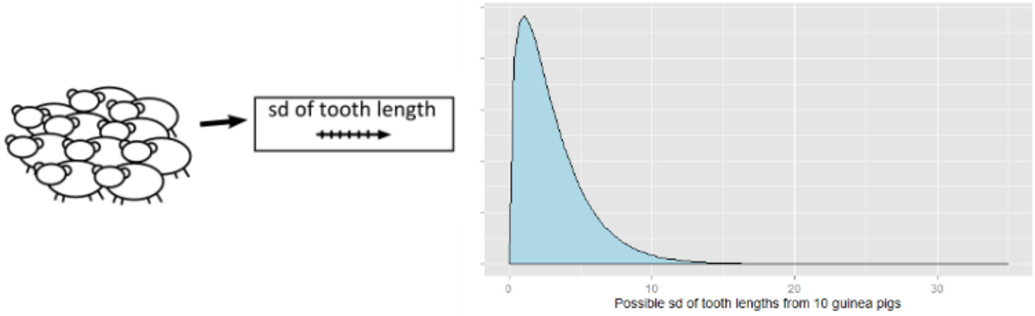
GRAPHS OF DISTRIBUTIONS

Tend to look like smooth "histograms"

PROBABILITY DENSITY FUNCTIONS


tooth length
Possible tooth lengths


mean tooth length
Possible mean tooth lengths from 10 guinea pigs


sd of tooth length
Possible sd of tooth lengths from 10 guinea pigs