How to analyse my data

4 - 6 July 2018

Outlines

- Exploratory data analysis and visualising data
- Formulating research questions
- Data types and related statistical tests
- How to interpret statistical results

♦ **Explanation of common statistical tests**

♦ **Workbook with worked examples then hands on practice**

♦ **Use statistical software to create output (SPSS)**

♦ **SPSS software guide provided**

♦ **Focus on understanding, concepts and interpretation of results**

Instructors
Nic Croce, Fran Baker

Statistical Support Service

THE UNIVERSITY OF NEWCASTLE AUSTRALIA

Statistical Support Service

Statistics refresher seminar series

What is statistics about?

12-June-2018

Nic Croce

statsupport@newcastle.edu.au

Introduction: Key Ideas

- Variable types
- Distributions: centre, shape, spread
- Signal = information amidst noise

Introduction: Key Ideas

- Sampling error = noise

- Non-sampling error = bias = study design critical

Introduction: Key Ideas

Assessing differences

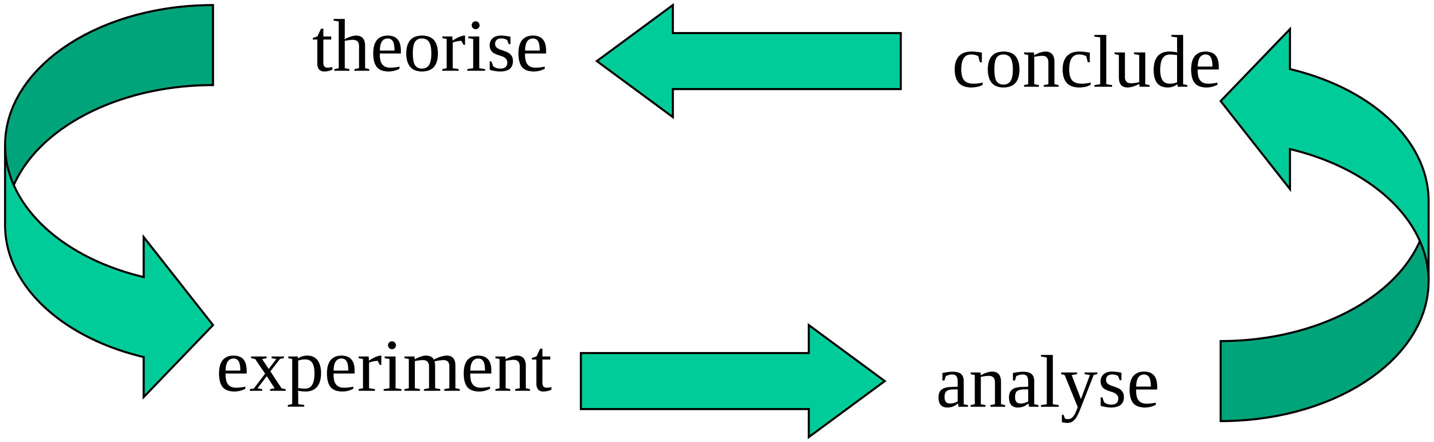
- Point Estimate vs Confidence intervals
- P values: α and Type I errors; to reject or not reject Ho

- Statistical tests
  - Single sample statistical testing of a mean
  - Statistical testing for differences of 2 samples
  - difference of 2 means, difference of 2 proportions

Statistics

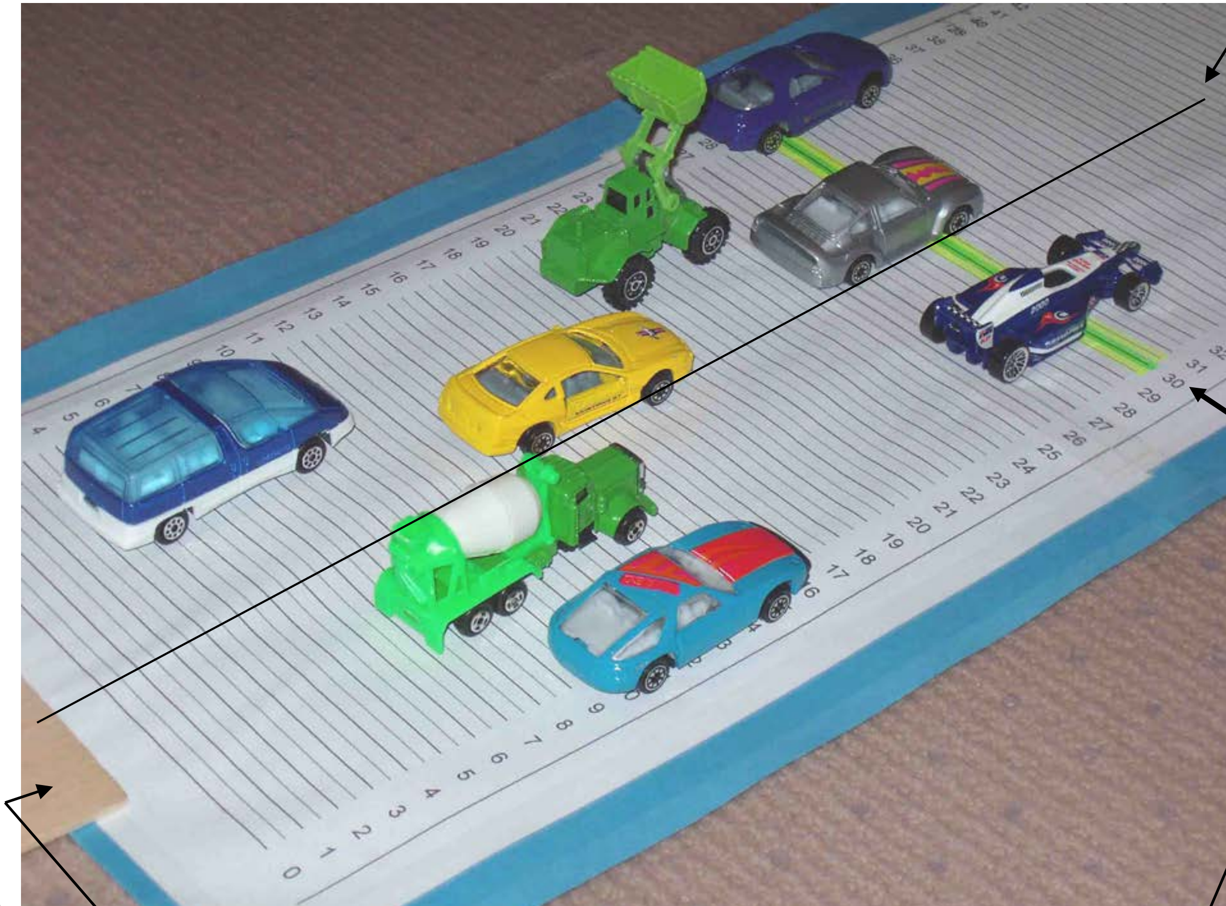
- Is the science of data.
- Data → Information (goal of research).
- Methods of collecting, analysing, interpreting, presenting data.

Fundamental to the scientific method



Car Process

Centre-line Deviations less than ± 2cm required



Target line for positioning the front of the car at distance 30 cm

Ramp down which the car rolls

30 cm

Car Process

Research question – broad

- What factors affect the finishing position of a vehicle?

Car Process

Break down into more specific questions

- Sedan cars only.
- Effect of ramp surface roughness.
- Effect of ramp starting position.
- Are dark cars less reliable than light coloured?

Even more specific questions

- Will cars travel **further** on **plastic** compared to **wood**?
- Does a difference of 2 cm in starting position have the same effect if starting in different positions?
  eg 5 vs 7 = 10 vs 12 = 15 vs 17 cm
- Do dark cars veer more than ± 2 cm from the centre line more often than light cars?

Variable types

- **Numeric** - values that "mean numbers"

  – **Continuous**: temperature, weight, speed, distance
  Other terms scale, ratio are similar for our purpose.

  – **Discrete**: #defects, result of die toss, product count

Variable types

- **Categorical** – values based on categories

  – **Nominal**

  gender – male/female   colour - blue/green/yellow

  – **Ordinal**

  Grades - FF, P, C, D, HD,

  Temperature - Low, Medium, High

Car Process – variable types

Response	**Explanatory**	**Level**
Position ***Numeric***	Surface roughness ***Categorical***	**Wood** **Plastic**
Position ***Numeric***	Starting position ***Numeric***	**5, 10, 15**
± 2 cm of centre	Car colour **Both categorical**	**dark, light**

Variation is everywhere

- No matter what type of data is being collected, variation will be present
- Key to understanding a problem is often about **understanding the sources of variation.**
- The goal of research is to **find the meaningful sources of variation while not being distracted by meaningless variation.**

Dealing with variation
Deterministic and statistical approaches

- **Deterministic**

Most/all variation can be explained
using a suitable deterministic model
e.g equations of motion in physics.

Dealing with variation
Deterministic and statistical approaches

•**Statistical** **Can't explain all variation.**
Use statistical models for the unexplained variation.
e.g. travel time for a car through a city.

Travel time could be predicted perfectly from equations of motion given the exact information about speed and course, but the ***variability due to drivers, traffic conditions etc leads to an uncertainty component****.*

Statistical Analysis

The goal is to find the signal amongst the noise

data = pattern + random variation

↓ ↓

signal + noise

↓ ↓

Information

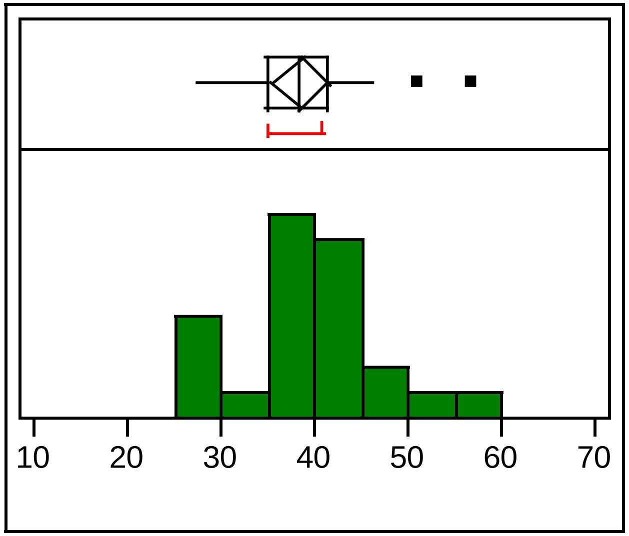
Estimate size so it isn't mistaken for signal!

Randomness

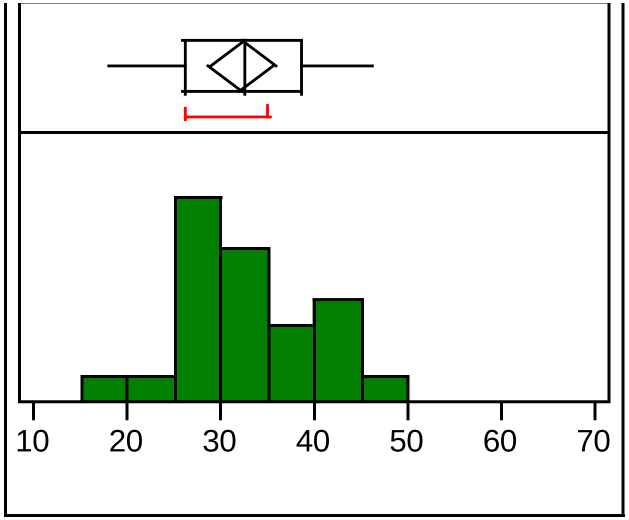
- SETI project
- Regularity in heartbeat of un-born babies

Results of study of roughness effect

What is the effect?



Plastic	
Mean	38.6
Std Dev	7.0
Std Err Mean	1.43
upper 95% Mean	41.6
lower 95% Mean	35.7
N	24



Balsa Wood	
Mean	32.4
Std Dev	7.6
Std Err Mean	1.55
upper 95% Mean	35.6
lower 95% Mean	29.2
N	24

Data Summary : using distributions - 3 features

Centre

- where is it? use mean or median- (point estimates)

Spread

- often measured using standard deviation (sd)
- small sd low variation,
- large sd large variation

Shape

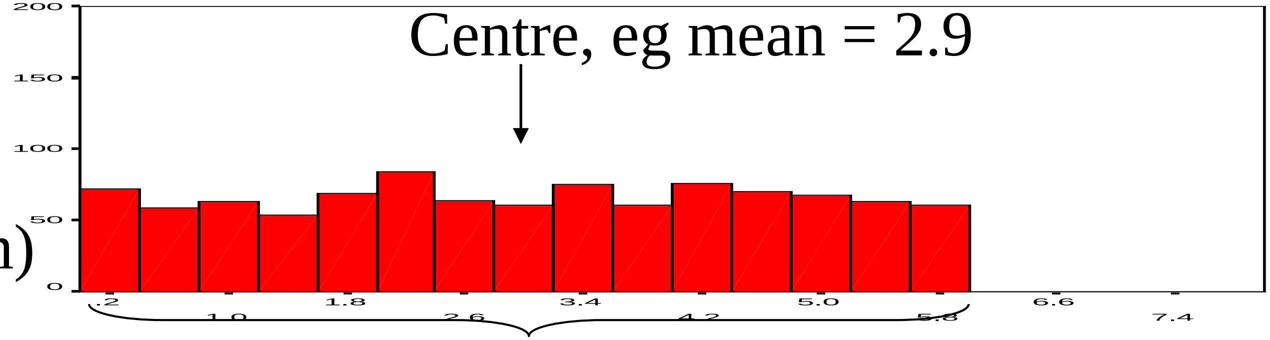
Symmetric, bell shaped, flat, skewed, truncated, bimodal

Histogram

Purpose: Present distribution of data showing centre, spread & shape.

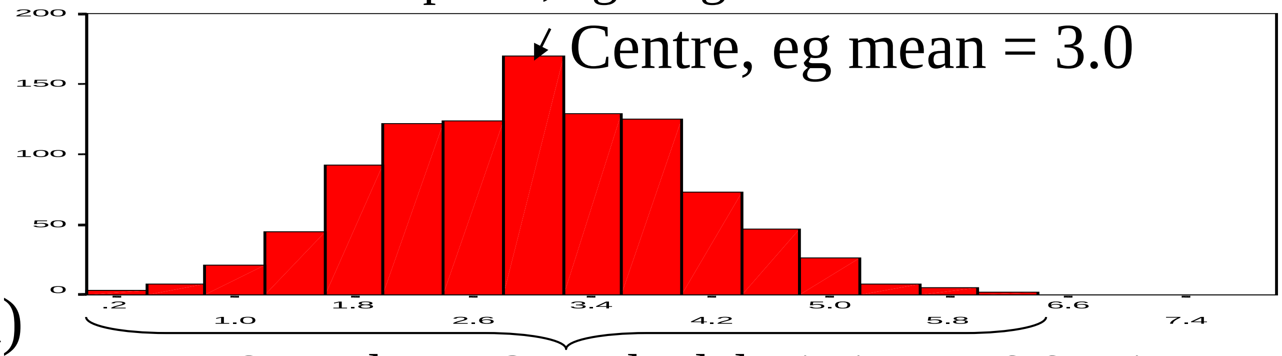
<u>Shape</u>

Flat (Uniform)



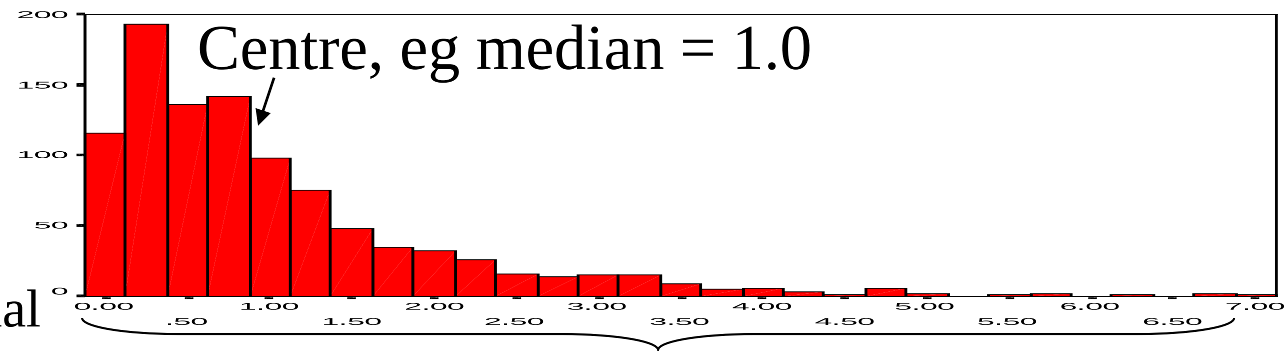
Spread, eg range = 5.8 units

Bell shaped (Normal)



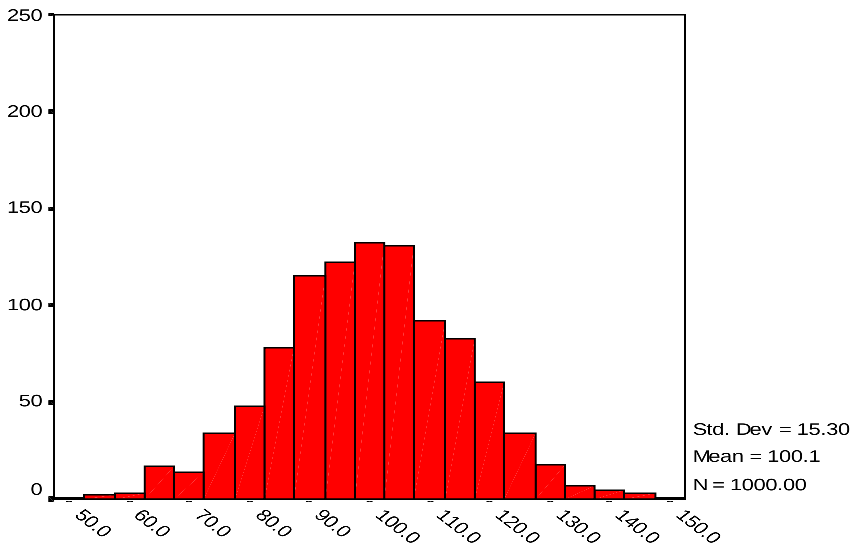
Spread, eg ±3standard deviations = 6.6 units

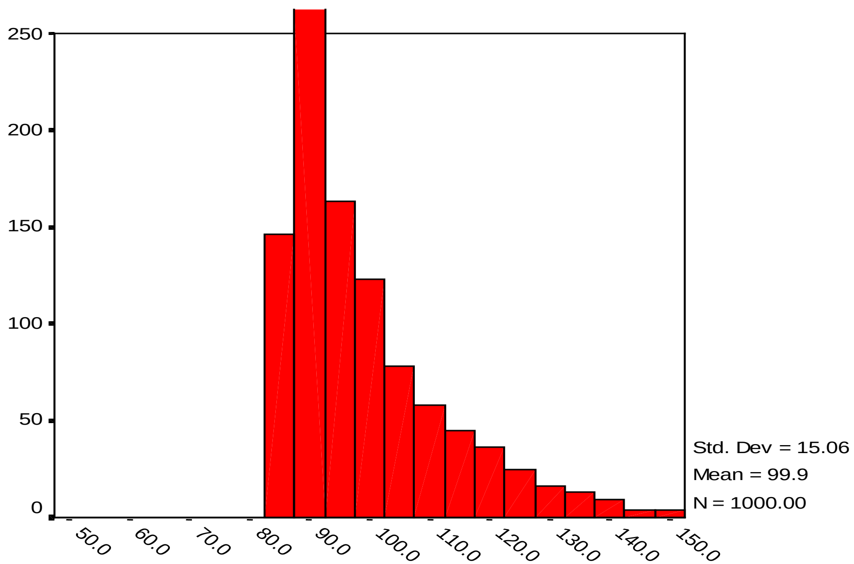
Left truncated & skewed to right (Exponential like)



Spread = 7 units

What is signal? What is noise?





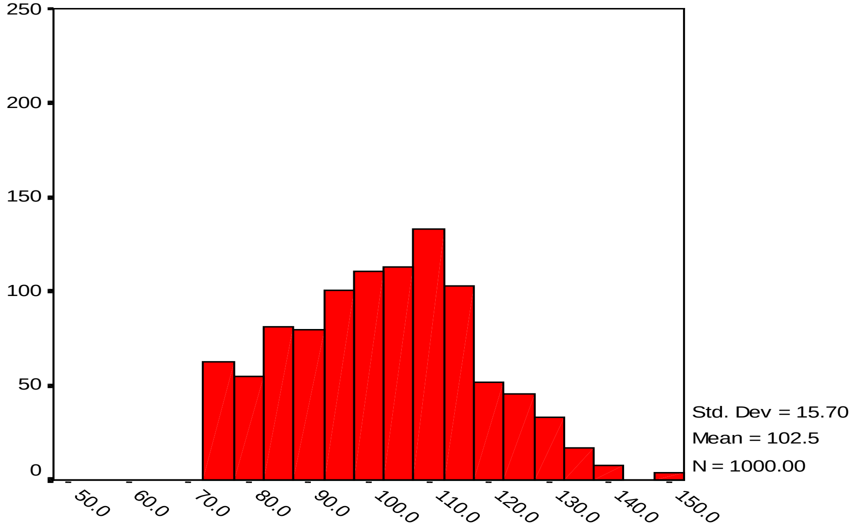
Signal = centre,
size of spread
shape

Noise = body of distribution

Comment re shape

E*xponential shape perhaps relates to physical processes, eg time between arrivals in a queue at counter in a shop.*

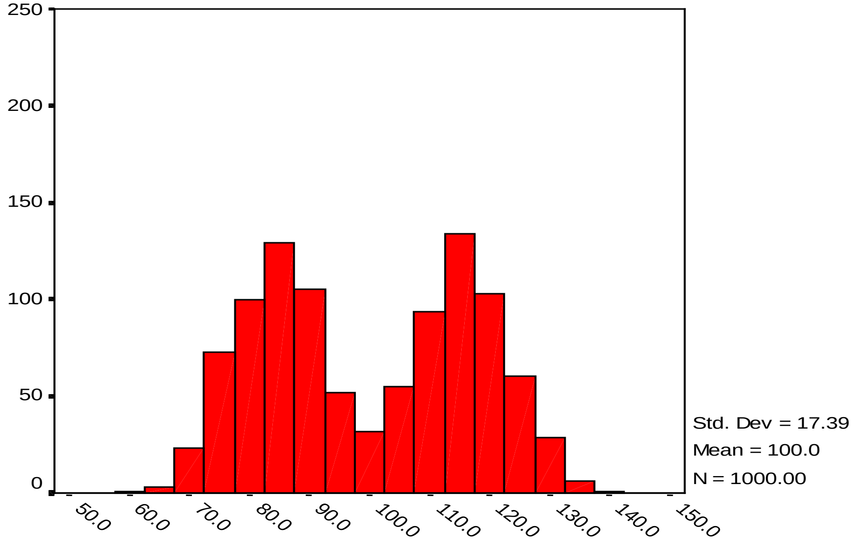
Interpretation



Two distributions, similar centres, similar spreads

What about shapes?

Left truncation suggests a limit of some sort.



Bimodal shape
i.e. two peaks,

suggests two different sources of variation

Refining the question further

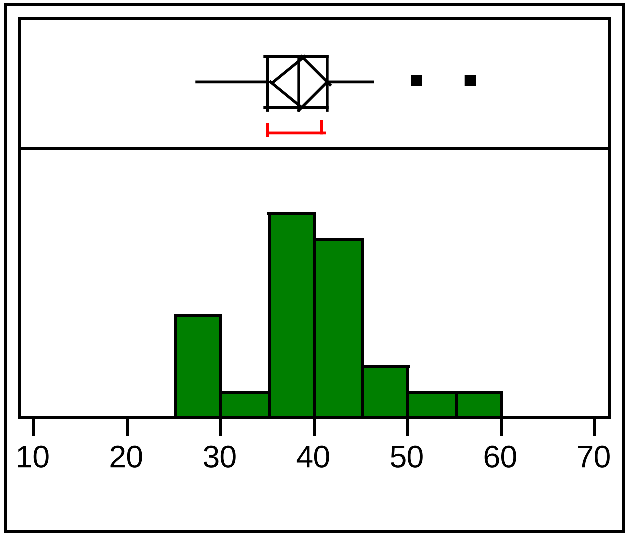
Will cars travel **further** on **plastic** compared to **Balsa wood**?

Not specific enough yet

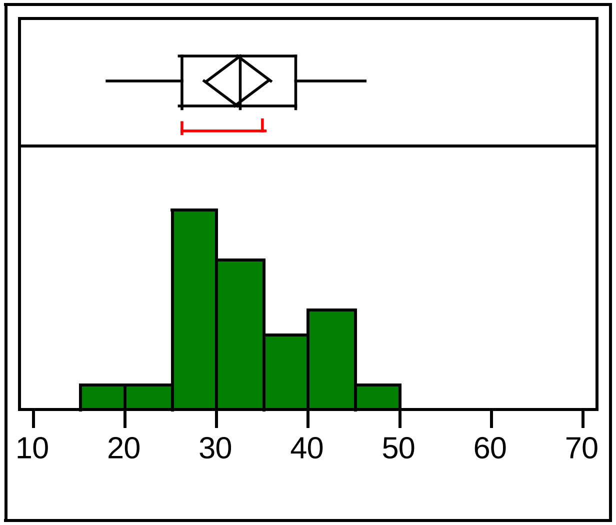
Could imply individual cars.

Is the ***mean*** distance travelled by cars **further** on **plastic** compared to **wood**?

Signal & noise – for individual cars



Plastic	
Mean	38.6
Std Dev	7.0
Std Err Mean	1.43
upper 95% Mean	41.6
lower 95% Mean	35.7
N	24



Balsa Wood	
Mean	32.4
Std Dev	7.6
Std Err Mean	1.55
upper 95% Mean	35.6
lower 95% Mean	29.2
N	24

Shapes are similar ?

Sampling

- Population vs sample
- Population parameters vs sample statistics
- μ, P, σ, vs $\bar{x}$, p, s

- Measure a part - **the sample**, but make conclusions, inferences, about the whole - **the population**

Sampling

Statistical inference is **only valid if sample is representative** of the population of interest

Random sampling is the gold standard

Sampling Error - the noise

Error in my sample estimate of the true but unknown population parameter which is solely related to the measurements being made on a sample

Will always occur (unless e.g. census)

Sampling Error - the noise

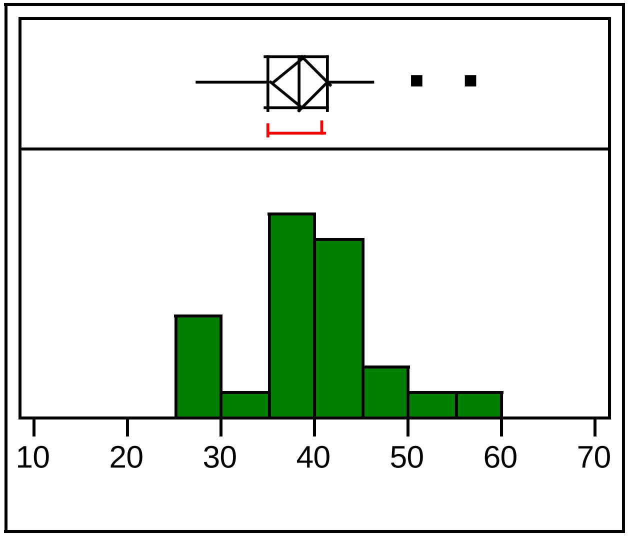
Decreases with increasing sample size

Likely sampling error can be estimated using statistics.

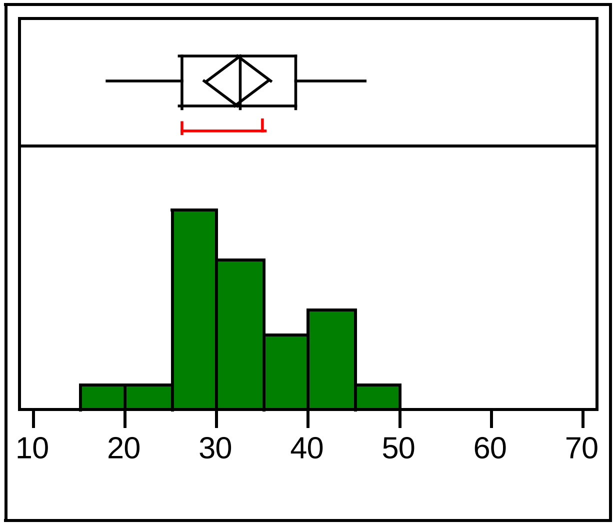
Common population parameter is mean

Noise for mean of population of cars

Standard error of mean



Plastic	
Mean	38.6
Std Dev	7.0
Std Err Mean	1.43
upper 95% Mean	41.6
lower 95% Mean	35.7
N	24



Balsa Wood	
Mean	32.4
Std Dev	7.6
Std Err Mean	1.55
upper 95% Mean	35.6
lower 95% Mean	29.2
N	24

Noise for means
=> standard error of mean

The standard error of the mean is the measure of **variability of means, not individual observations.**

This is based on the concept of repeatedly taking random samples of 24 cars and finding the mean of the 24 individual position measurements.

Noise for means
=> standard error of mean

This creates a new population – not of individual position measurements, but of means of 24 items.

The standard deviation of the collection of the means (of 24 items) is the standard error of the mean.

Non-Sampling error → Bias

- All other errors are classified as non-sampling error (even if related to aspects of the sampling process).

- **Cannot be reduced by increasing the sample size**.

Non-Sampling error → Bias

- Statistical methods <u>cannot compensate for poor data or poor study design</u>.

- Consider television phone in surveys.

Bias & Precision

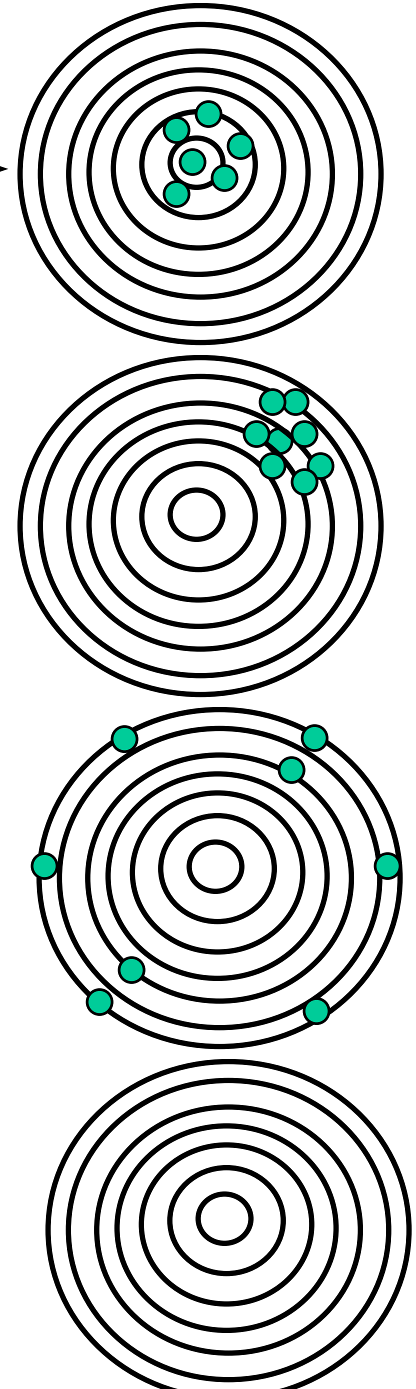
Accuracy = unbiased and precise

Accurate = correct result - often

Inaccurate = wrong result - often

- Bias - off target,
- Poor precision - too much variation.

Surveys and other studies can get the wrong answer, ie a bias, if non-sampling error is not controlled.



Unbiased and precise

Precise but biased

Unbiased but not precise

Biased and not precise

Example of bias in a survey

- Poll of readers of Literary Digest magazine in USA correctly predicted presidential election outcome from 1916.
- In 1936 they were wrong, they predicted the current president Roosevelt would only obtain 44% of vote. He won by a landslide with 62% of vote.
- Why was the prediction wrong this year when it was correct all the others? Bias in sample.

Example of bias in a survey

- Magazine mailed to people from addresses in telephone directory. In those years wealthier people tended to have telephones. Selection procedure was biased against the poor.
- In previous years wealthy and poor voted similarly.
- But this time, due to the effect of the great depression and unemployment, the way the poor voted was different to that of wealthier people.
- Modern survey methods since 1950's rely on random samples.

Confidence interval for a population parameter

= sample estimate ± some margin of error

eg mean or difference of 2 means

Confidence interval

= sample estimate ± multiplier × std error uncertain knowledge + knowledge about that uncertainty = something we can use

Example assumption for means

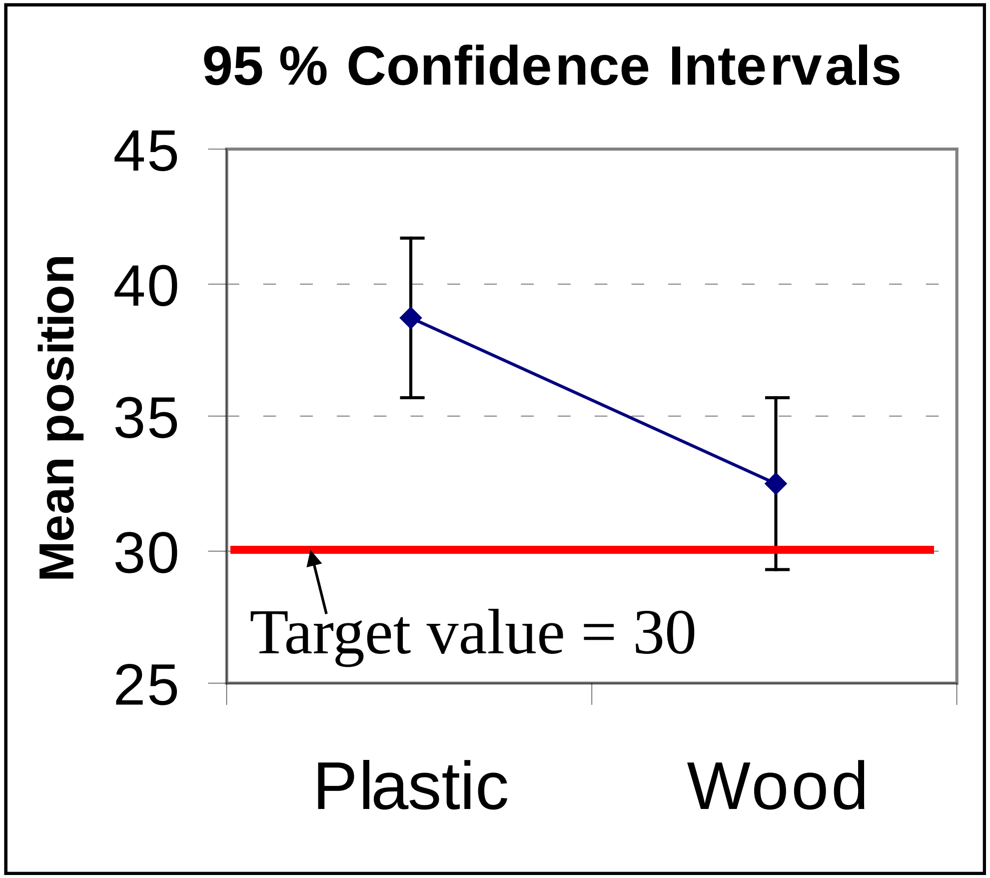
- Data is a random sample from the population of interest (otherwise bias)
- Distribution of sample means is normal

Answering questions for a single sample of data

Different to a reference value?

Reference Values and Confidence Intervals

			95% CI		
	Mean	**se**	**LCL**	**UCL**	**Uncertainty**
Plastic	38.6	1.43	35.7	41.6	± 3.0
Wood	32.4	1.55	29.2	35.6	± 3.2



Q: Are either of the two populations not consistent with the target value of 30?

A: Yes – plastic.

The target value 30 is outside the CI range for plastic.

Interpretation of confidence interval

We are concerned with what is the true (population) **mean** distance that would be travelled by **ALL** sedan cars.

We have an **estimate** of the **mean from a sample of 24 cars**

AND

an estimate of uncertainty for that mean.

Interpretation of confidence interval

For wood we have estimated that the true mean distance travelled by all sedan cars is 32.4 cm

However, the true mean could be as low as 29.2 cm or as high as 35.6 cm

How sure are we of this? 95% confident.

Alternative approach: Hypothesis Testing

Uses the same information used for confidence intervals, but in a different way (1 sample t test).

Testing is done relative to a reference value that is meaningful to the investigator.
In this case the target value = 30

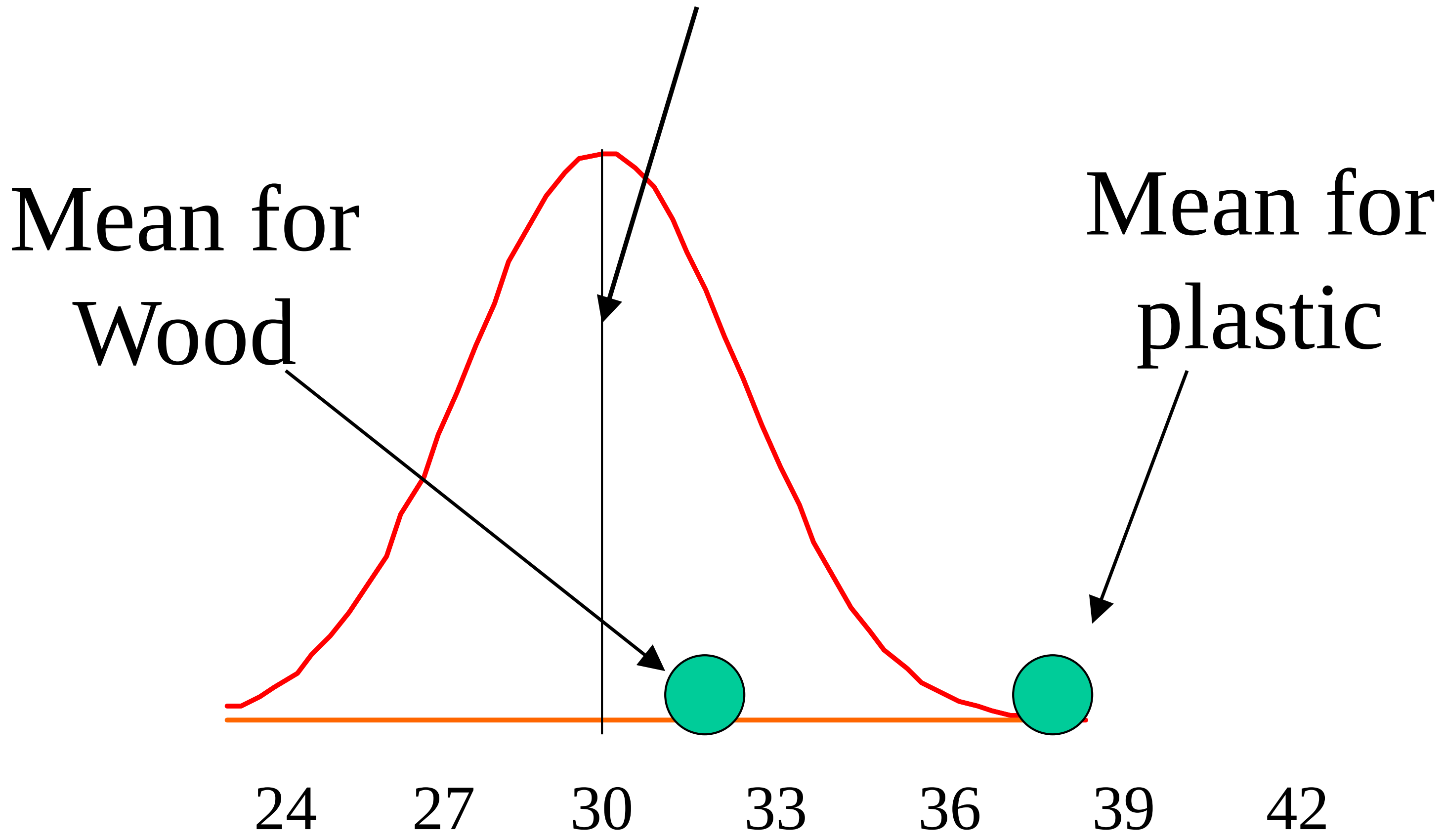
Alternative approach: Hypothesis Testing

Call this the null hypothesis.

Then the mean of the data is compared against this.

We need an estimate of the noise.
Where can we get it? Standard error of mean.
(we will assume a common (pooled) standard error for simplicity)

Probability distribution of uncertainty about hypothesised value = 30



Wood is **consistent** with hypothesised value

Plastic is **not consistent** with hypothesised value

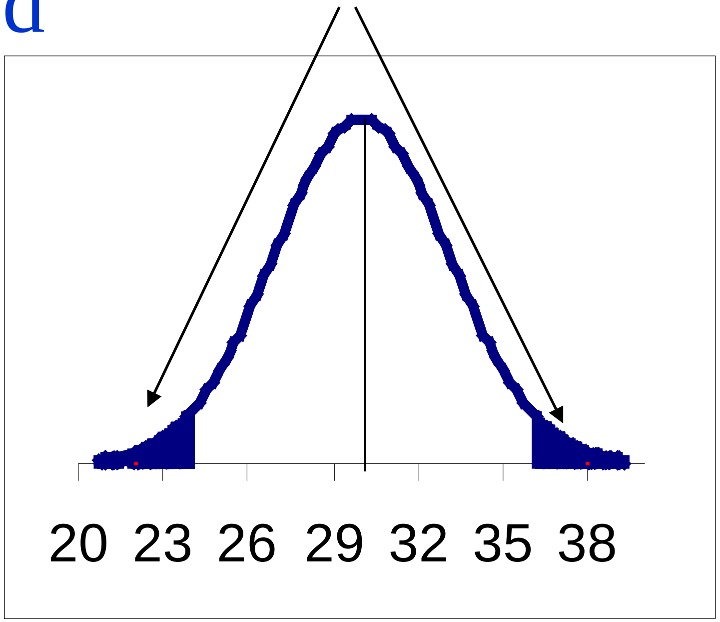
Reasoning and conclusions

- The mean value for Wood 32.4 is **not unusual** relative to the uncertainty so the mean for Wood is not significantly different to 30.
- The mean value for Plastic 38.6 is **unusual** compared to the expected variation associated with the hypothesised value, so we conclude it is **not consistent** with it. Therefore the mean 38.6 is significantly different to 30.
- How would the conclusions change if the spread of the uncertainty was twice as large?

P values quantify how unusual the data is compared to the hypothesis

- P value is the probability a mean as extreme, or more, than that observed could be obtained by chance, **IF THE NULL HYPOTHESIS WAS TRUE**.
- **The smaller the P value the stronger the evidence provided by the data against the hypothesised value.**
- **Why?**

P value is equal to the area of the tails of the distribution



Making decisions

First choose a significance level, often referred to as alpha α.

A common choice is $\alpha = 0.05$, so let's use it.

If P value < .05 we conclude there is a significant difference between the hypothesised value and the data.

If P value $\geq$ .05 we conclude no difference.

Making decisions

In our case for Plastic P = < 0.001 is less than 0.05 so we have a significant difference from the target value.

For Wood P = 0.10 which is NOT less than 0.05 so we do not have sufficient evidence to declare we have a significant difference from the target value.

P Values and Alpha

The P value generated from a statistical test also represents the probability of rejecting Ho when in fact Ho is true.

This error is known as alpha, α, the chance of making a Type I error, ie, rejecting Ho when Ho is true.

By convention, a number of disciplines set α at 0.05.

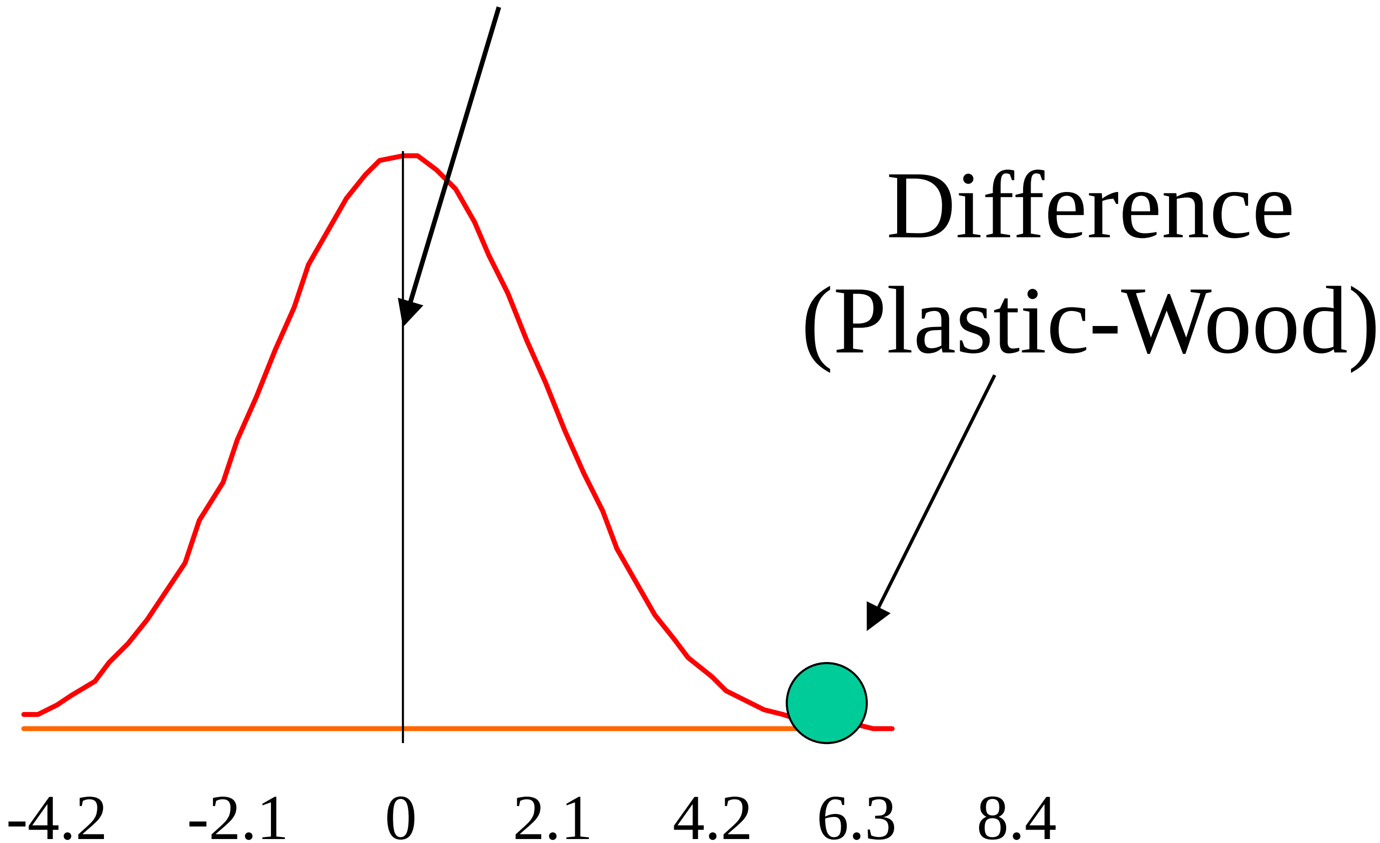
This means the researcher is willing to accept a 5% chance of wrongly rejecting Ho, the null hypothesis.

Comparing two samples of data

Calculate difference of two means

- (Plastic – Wood) = 38.6 – 32.4 = 6.2
- Standard error (se) of difference = 2.1 (individual se's = 1.43, 1.55)
- Why is se of the difference larger than the individual se's?
- Variability/error is additive when combining numbers.
- So now we proceed in a similar way to the one sample situation.

Probability distribution of uncertainty about hypothesised value = 0; Why zero?



The difference observed, 6.2, is not **consistent** with the hypothesised difference of zero. Conclusion: we have a significant difference.

P value for difference of two means

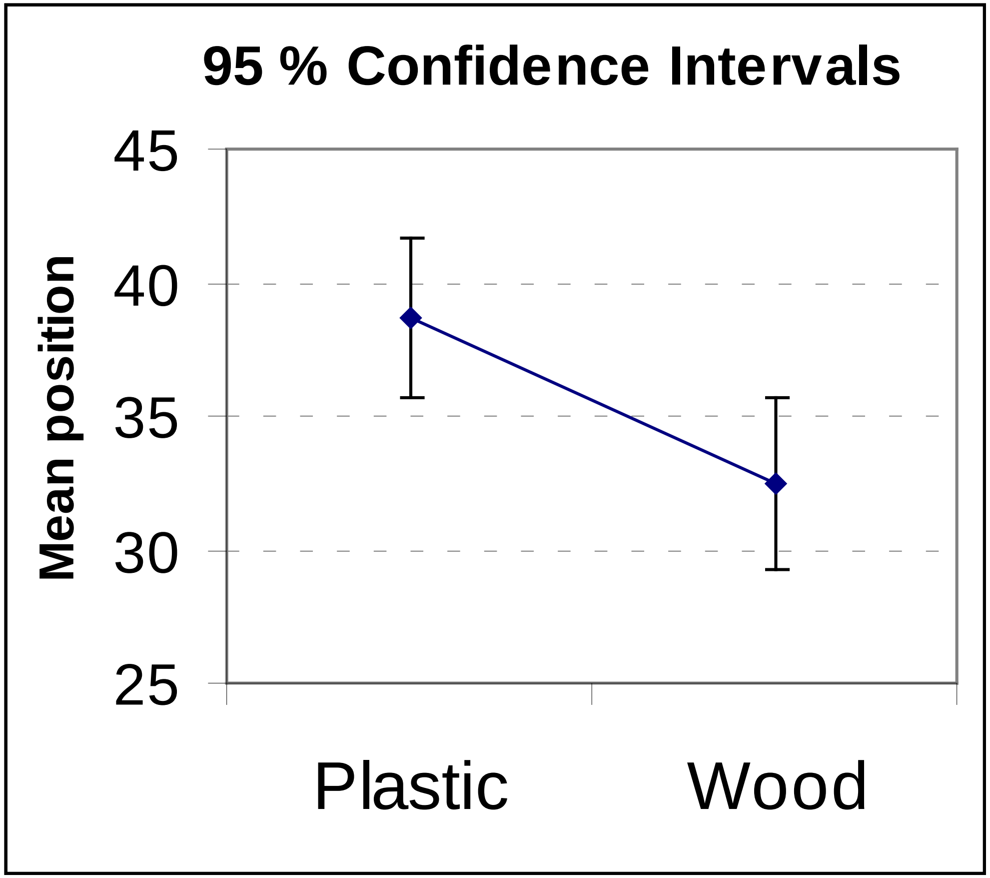
- The p value quantifies what we see visually in the previous slide (using the 2 sample t test).
- P = .005
- What does this mean?

P value for difference of two means

The probability we could see a difference as large as 6.2 or more (or –6.2 or more) is 5 in 1000, **if there was really no difference**.

So we choose to believe there was a difference, rather than believe there was no difference and what we see came about through a very unusual chance event.

A visual assessment using CI's for a single mean



The CI's do not overlap at all hence we would also conclude the true, but unknown population means are different.

What is the logic?

Conclusion: Cars run further on plastic.

What if there was a lot of overlap?

Confidence interval for difference of 2 means can be calculated too.

This could then be used to assess the significance of the differences as we did with p value method.

Categorical variables

Recall our earlier questions about reliability of cars.

Do dark cars veer more than ± 2 cm from the centre line more often than light cars?

Response	**Explanatory**	**Level**
± 2 cm of centre	Car colour	dark, light

Given what you know about the natural world what do you think is a reasonable answer is to this question?

Analysis – hypothesis test approach

Count	No	Yes	Total
Dark	18	26	44
Light	18	34	52
Total	36	60	96

Row %	No	Yes
Dark	40.9	59.1
Light	34.6	65.4

Is 59.1% different to 65.4%?

Significance Test

Test	ChiSquare	Prob>ChiSq
Pearson	0.403	0.53

← P value

Conclusion **No significant differences between %'s for light and dark**

What was the hypothesised value?

- The significance test in the table was based on a hypothesis that the percentages for the two groups were the same.

- That is

Null hypothesis

% = 60/96 = 62.5%

What was the hypothesised value?

The P value = 0.53 is the chance the two %'s could be 59.1% and 65.4% if the true % was 62.5.

Conclusion The data did not differ enough between the two groups so there was no evidence to support the conclusion the two groups were different.