Challenges in sensory and consumer analysis – a hands on stats perspective

Gemma Hodgson
Qi Statistics Ltd.

http://www.qistatistics.co.uk

Background

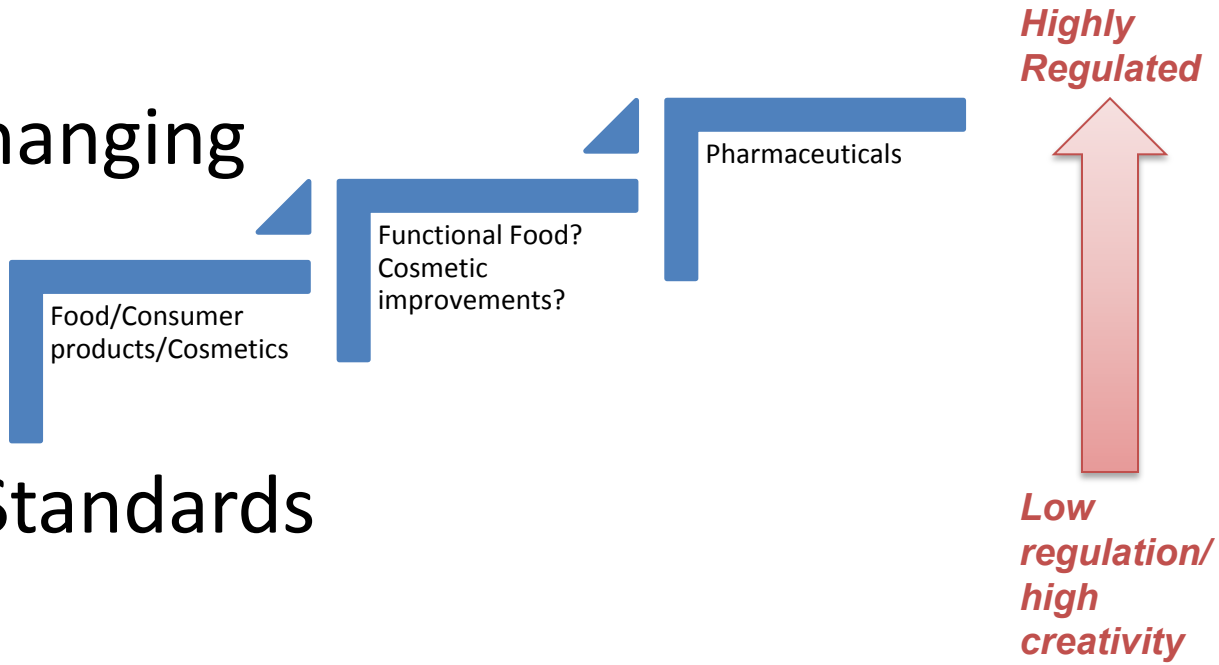
- BSc Maths with Statistics, Imperial College London
- MSc Medical Statistics, LSHTM, London
- 17 years in Pharmaceutical Industry: 12 years at Pfizer, 5 years at Takeda
- 4 years as Statistical Consultant & Trainer (All Industries) at Qi Statistics Ltd.

Topics

- Key challenges people regularly have
  - Not enough people
  - Ignoring/underestimating variation
  - Quest for a significant P-value
  - Using the wrong test (not knowing your objective)
  - Measuring the wrong thing
- Issues in global data collection
- Discussion?

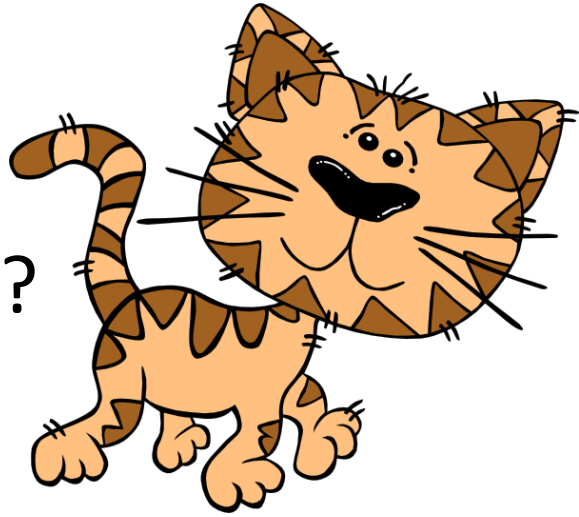
Why do we need statistics?

The world is changing



- Regulation/Standards
- Competitors
- Social media – you can't hide!
- A more aware public
- The world is smaller

Motivational Example



"8/10 cats like Whiskas"…

This brings some questions to mind?

- What if 6/10 cats liked Whiskas?
- Is that still good?
- What if 8/10 cats liked fresh fish but 6/10 cats liked Whiskas? Can we conclude Whiskas is as good as fresh fish?

Statistically speaking…

Proportion like whiskas	Proportion like fresh fish	Statistically significant difference(Fisher)
7/10	8/10	1.000
6/10	8/10	0.628
5/10	8/10	0.350
4/10	8/10	0.170
3/10	8/10	0.070
2/10	8/10	0.023

With 10 cats, there is only a statistically significant difference observed when the difference is pretty big…but maybe to market it, consumers might say anything less than 50% of cats not liking … *so perhaps 10 cats is not enough?*

- Consumer relevant difference?
- Chance of finding a difference?
- Statistically significant?....so many questions!

Trial in action…

- You each have the results from your panel of 8 people from a trial you ran today
- You are comparing your current product vs a new product – that is expensive to make
- Just look at ***your*** sheet!
- Is it worth launching the new product?
  - Go GREEN if you think YES it's worth the investment
  - Go RED if you think NO, stick with current

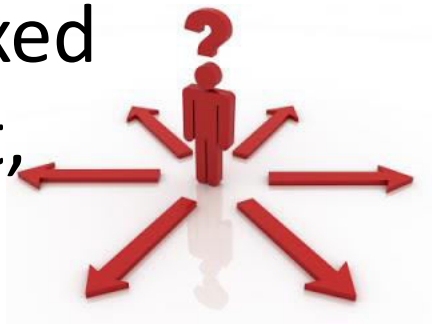
Trial in action…part 2

- You now have the results from your panel of **16** people from a trial you ran today
- Look at your updated results
- Is it worth launching the new product?
  - Go GREEN if you think YES it's worth the investment
  - Go RED if you think NO, stick with current

What do you conclude from this?

Small trials can be uninformative

- In the first set of trials, we had a mixed response to launch the new product, we were undecided

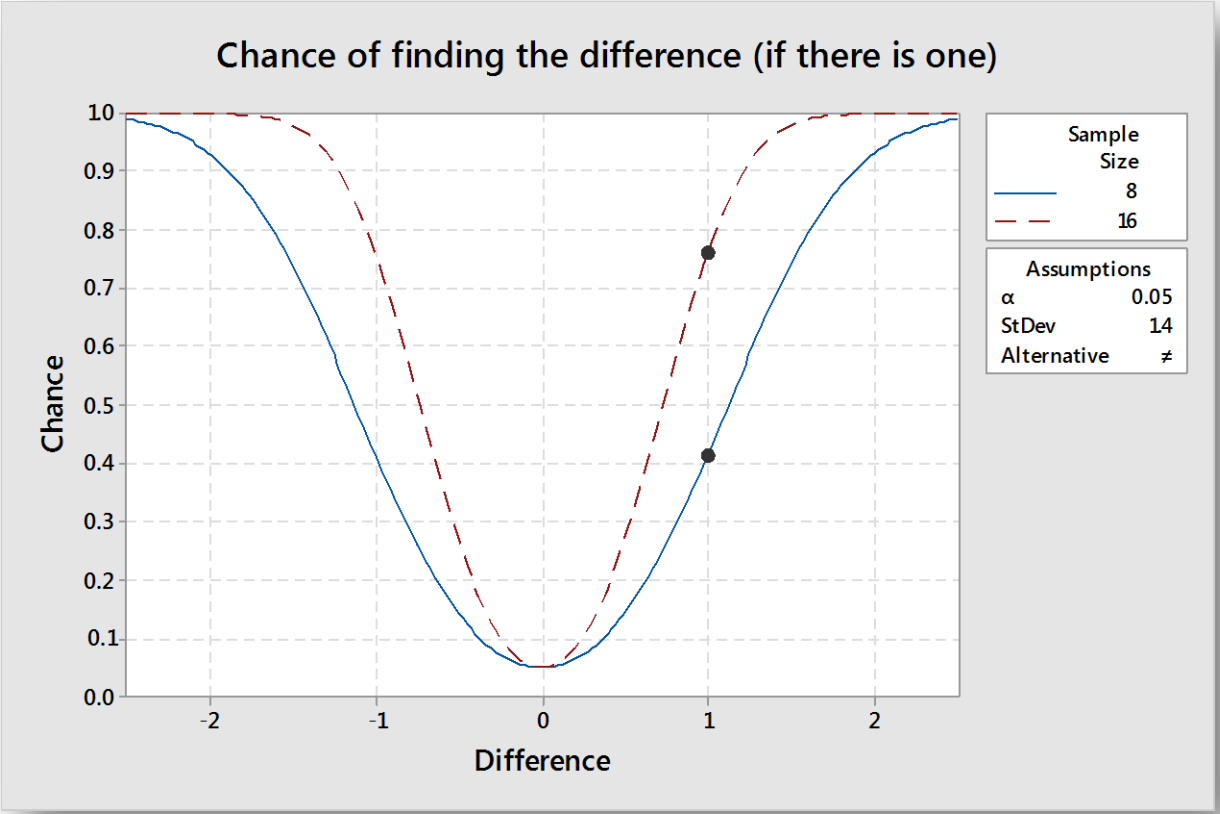


- In the second set, we were more sure.
- Here's the summary statistics from the full dataset :

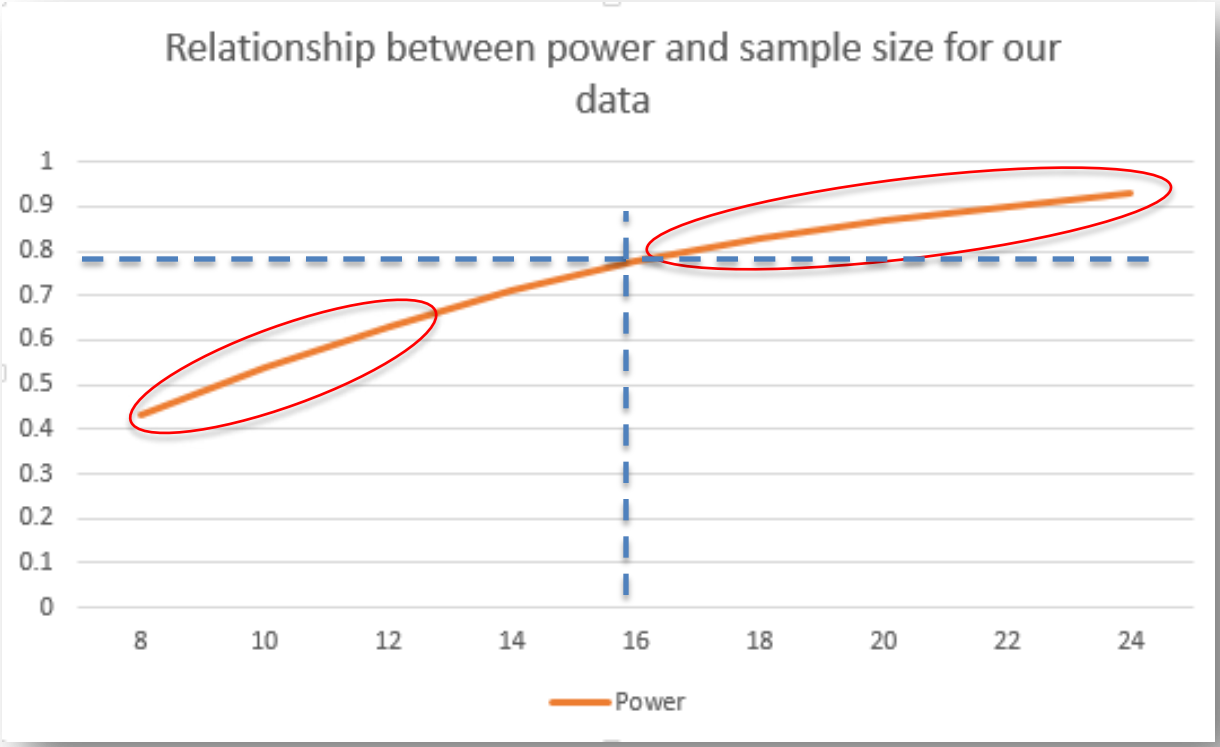
	current product	new product
N	300	300
Mean	6.0	6.9
Std Dev	1.4	1.3
Min	2	3
Max	9	9

Relationship

- What is the chance we will find a difference?



Relationship



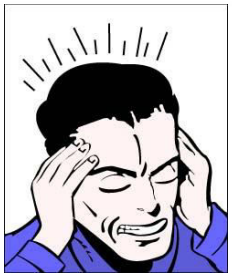
- With >16 people, the increase is not much more as we go up to 24.
- Difference at the lower end is quite big
- Depending on how certain you need to be, maximize your panel size!

Why is this?

- The chance of finding a difference in a trial, (if there really is a difference) ?
- Dependent on
  - size of the difference you're interested in
  - the variability you have in your panel/testers
  - how many testers you have…
- So one way to improve your chances is to maximize your panel size

- Another is to minimize the variability

The pain of variation…



- **Between and Within Person Variability**

Variability arises in many places…

Think of a golfer…

To consistently hit the shot exactly where he wants?

Affected by many things:

- Wind speed
- Stance
- Power he hits with
- Swing
- Noise behind him
- Length of grass
- Etc…

And they can all change!

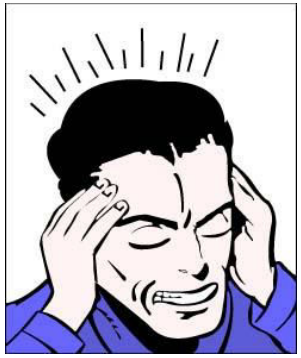
Some he can control, others he will have to learn to live with

In any sensory/consumer test or trial…

Variability arises in many places…

- Could be due to measurement error
  - equipment
  - technique
  - temperature of room/equipment
- Could be due to the panellist/consumer
  - psychological state
  - Physical effects/differences
  - time of day
- Could be due to other differences in people
  - Skin type
  - other physical attributes
  - *Etc.*

…we have to deal with it



- Statistical Tests measure any observed signal (effect) in the presence of all this noise…and give you a measure of how big your effect is ***relative*** to that noise

- If you ignore it, you may be actually reporting the product effect mixed up with the noise

…we have to deal with it

- So what can you do?
  - Report the estimate of effect with it's estimated variance (e.g. Mean + SD) and let people use their own judgement
  - Do a statistical test comparing observed effect relative to the observed noise (e.g. t-test) to see if it's a real effect
  - Use the variation you see to explain why and how your products differ (e.g. PCA)
  - Reduce the variation by training your panel
  - Cannot train consumers, hence larger sizes needed + careful thought into no. of questions
    - more questions -> poorer quality data?

Quest for a significant P-value

- Don't join the crowd hunting for significant p-values…

Statistical significance does not tell you that the difference is of *commercial/manufacturing/consumer* ***importance***

Significance is driven by sample size – the bigger your data set (higher N) the smaller the differences that will be detected as significant.

What does the significance measure?

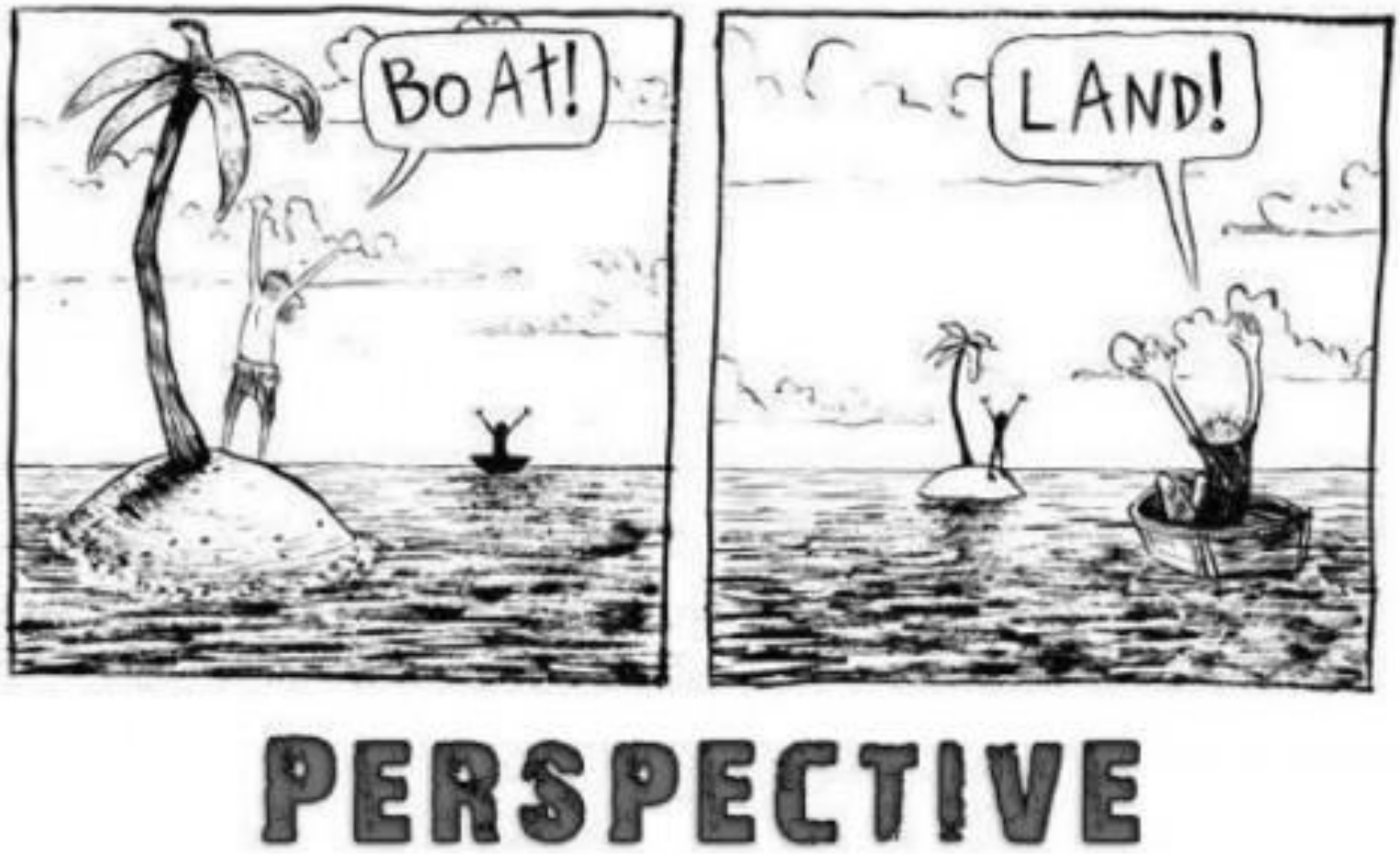
- Measures the risk that, if we report the samples as having different mean scores, we are in fact wrong and the signal we have seen is due to chance.

  i.e.

  - Carry out t-test for difference in mean overall liking score between two products.
  - The results are significant at the 5% level (i.e. $p < 0.05$)
  - If we report to public that there is a difference in average liking score between the products – we only run a 5% (1/20) risk that we are wrong and the result has just occurred by chance

Significance gives you confidence in the *repeatability* of your result

The significance of 'significance' depends on your perspective…



http://www.jokeoverflow.com/picture-jokes/whatever/perspective

Scientists' Interpretation of p-values

P-value	Detection of difference from baseline assumption
Greater than 0.15 (>15%)	Risk in concluding a difference too high – conclude no significant difference detected
0.10-0.15 (10%-15%)	"Grey area" – Interpretation depends on context of test
0.05 -0.10 (5%-10%)	Difference cannot be altogether discounted (need more data to confirm)
Less than 0.05 (<5%)	Significant Difference detected

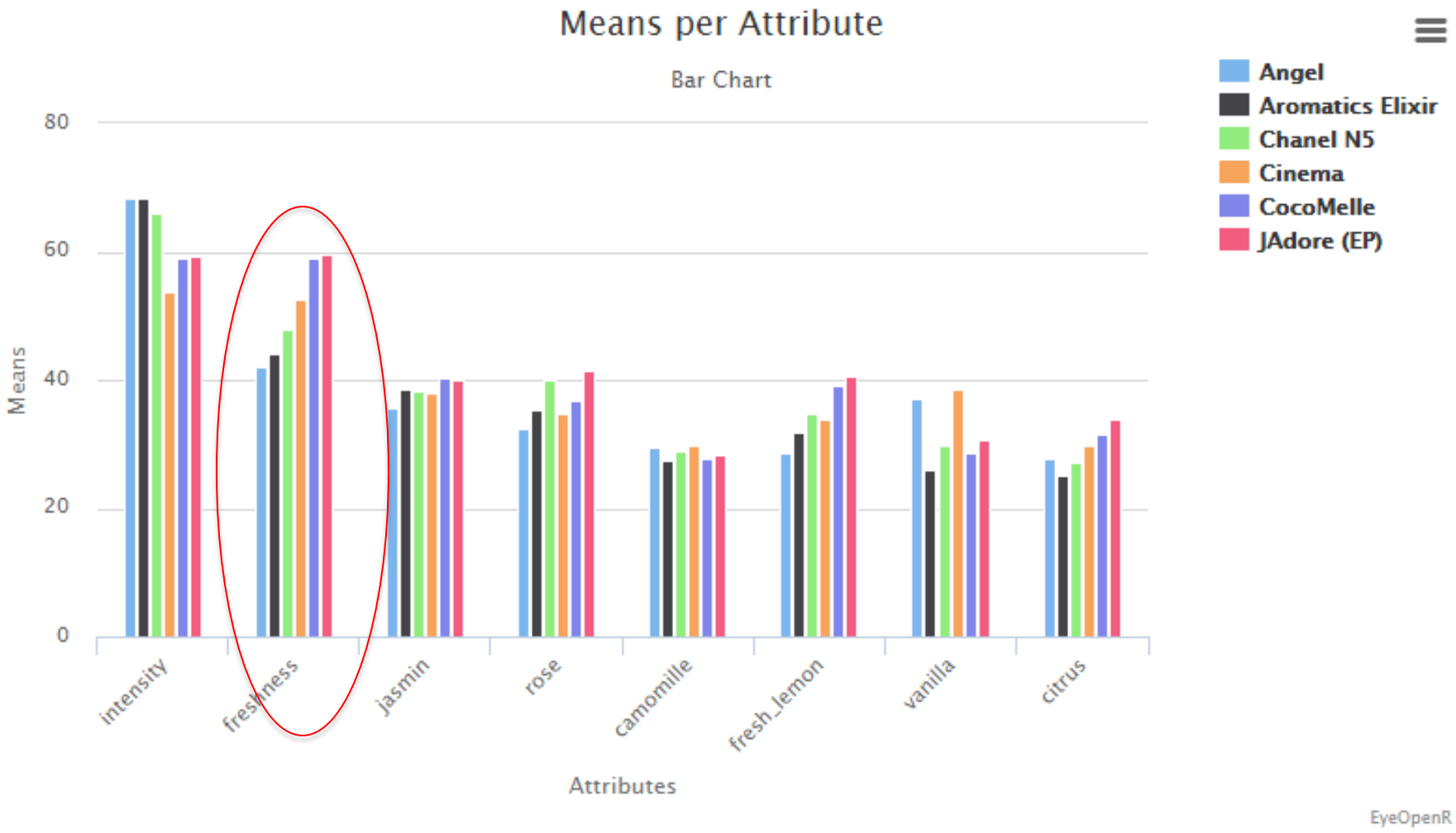
Confidence Interval for Difference

- More informative that just the significance test
- Tells you the likely size of the difference (worst case v best case)
- Also has the significance test embedded in it
- If confidence interval (95%) contains zero then t test will not be significant at p=5%
- Can be considered as a ‘range of options’ – to aid rapid understanding

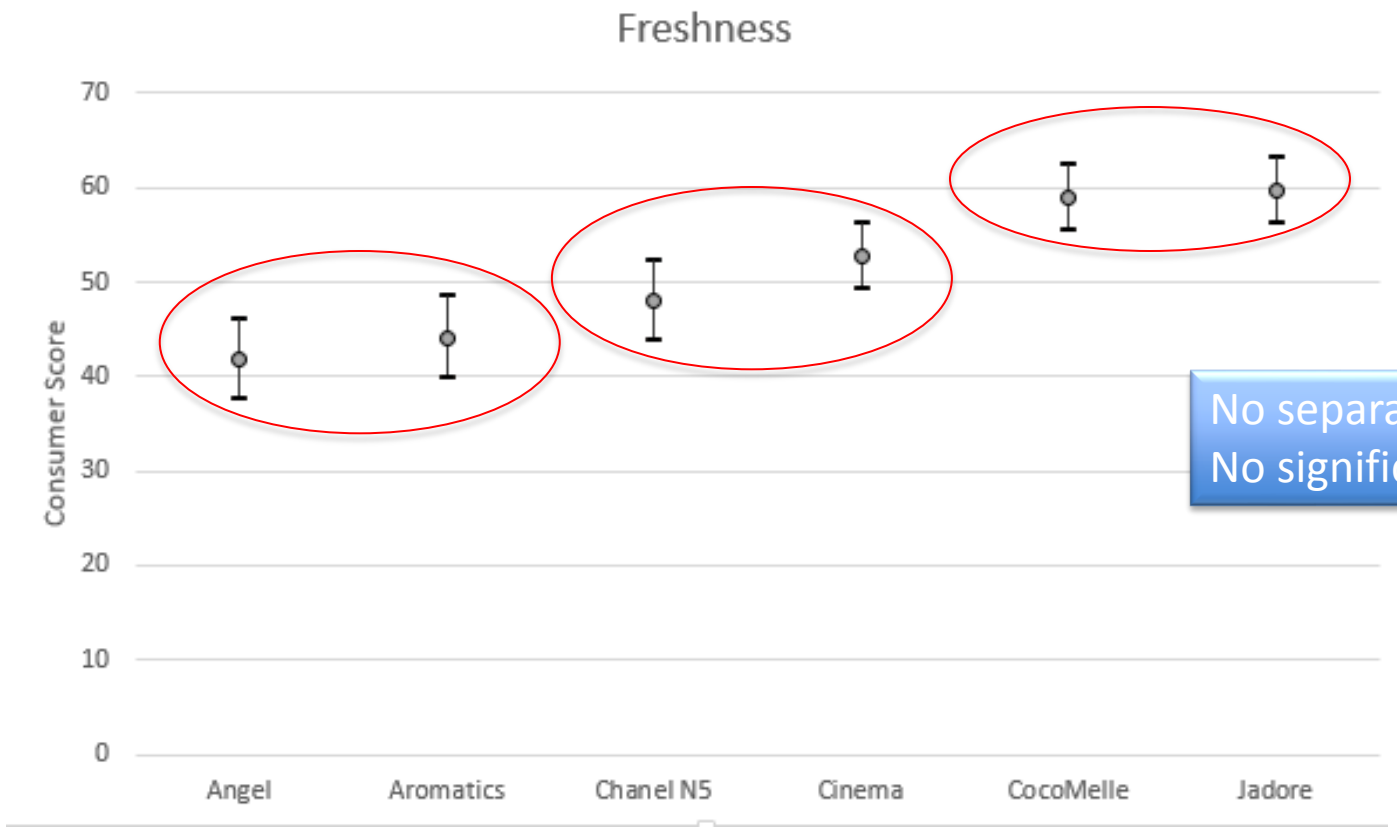
Example from a consumer trial

- 103 consumers
- 5 perfumes
- Judged on intensity and type of smell
- Scored from 1-100

Results 1

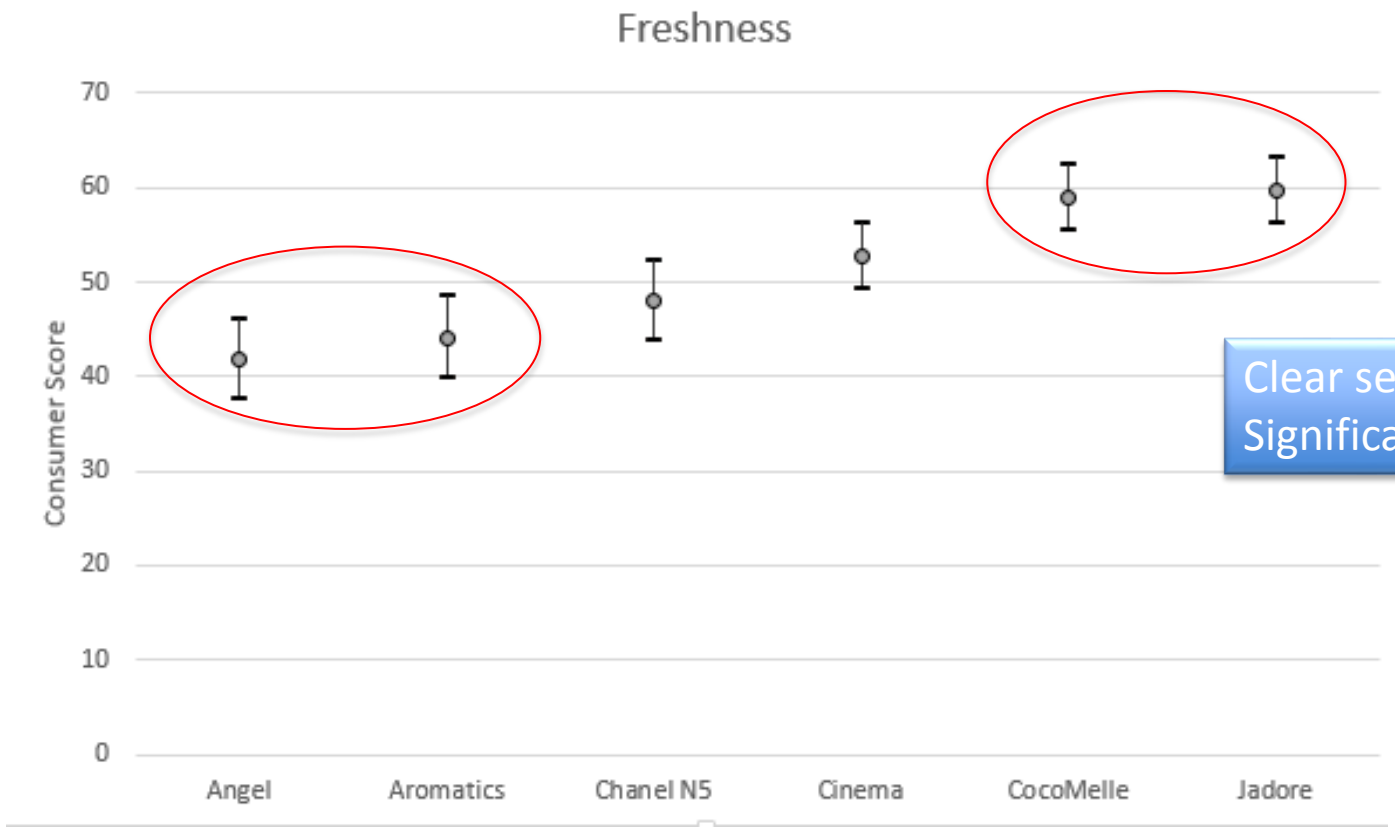


Results 2 - Freshness



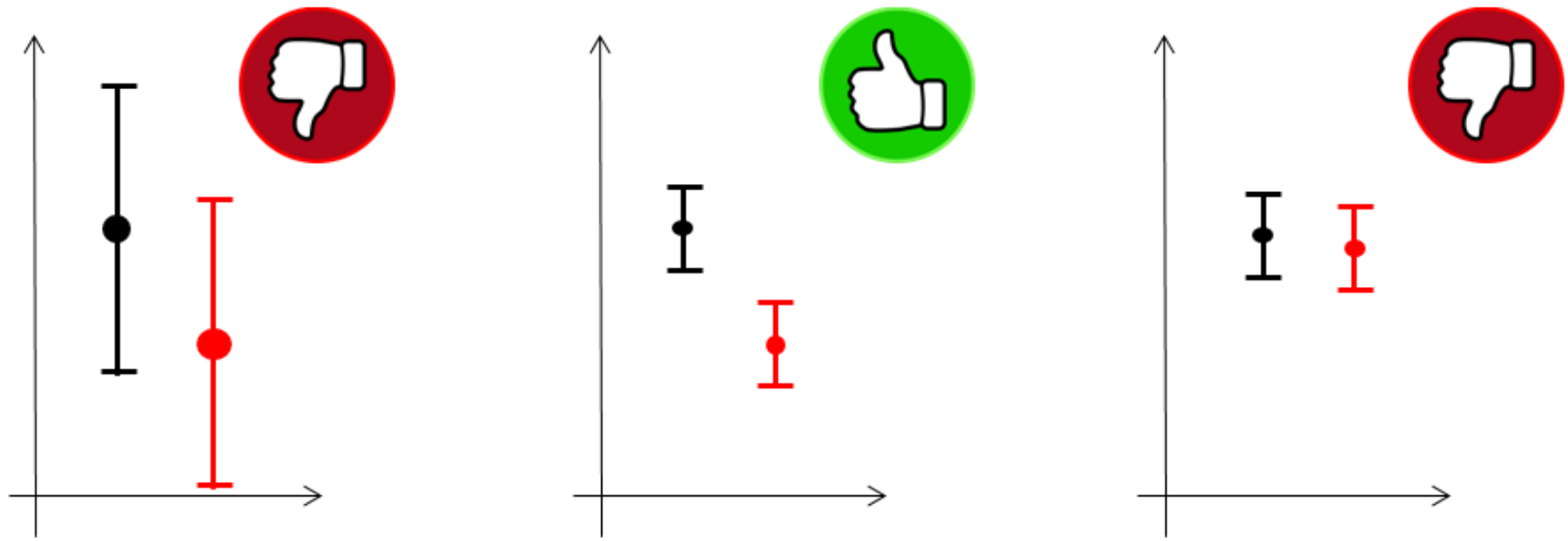
No separation
No significant difference

Results 2 - Freshness



Clear separation
Significant difference

In general…



Good separation, high variation

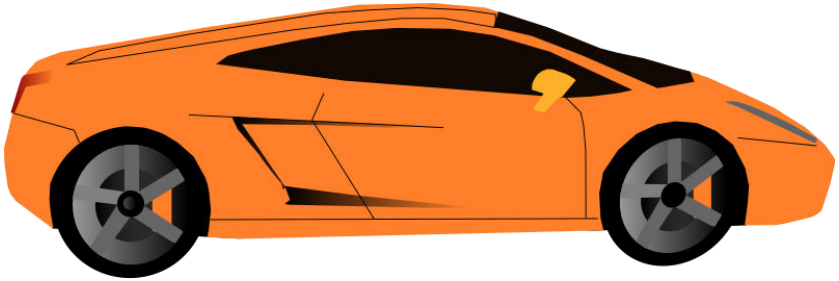
Good separation, low variation

Poor separation, low variation

Detectable product differences depend on background variation

Know your objective

- If I look for a difference between 2 products and don't find one, can I conclude they are the same?
- *If I stand on the street and look for a orange car and don't find one, does it mean orange cars don't exist?*



Know your objective

- NO! *Not unless I stand there for ever…*
- It could be just that you don't have enough data to show it

Absence of evidence ≠ Evidence of absence

If you are looking to prove equivalence/parity then you should know this up front and use the correct statistics to show this (needs higher N)

What should I measure?

- Need to also think about what to measure…

1. Continuous Measurements
2. Grading score (sensory expert?)
3. Consumer score
4. Binary question
   (e.g. Is it creamier? Yes or no)

Objective, sensitive to a difference

Hard to detect a difference
Loss of information
Can be subjective (high variability)

- Consumer tests will have many

Maximise the info from trials…

- Think about the sample size and what you measured – is it sensitive enough to detect a difference?

- Use a comparator product – blinded where possible

- Randomise order

- Report the results with estimates of variability
- Use statistical tests to report evidence of a significant difference

Data collection challenges 1

Language/Culture

- Different countries/cultures use language in different ways
- Many papers in FQ&P:

- *The roles of culture and language in designing emotion lists: Comparing the same language in different English and Spanish speaking countries.* Hannelize van Zyl Herbert L. Meiselman, Food Quality and Preference, Volume 41, April 2015, Pages 201–213
  - More similarities among the four English countries than between Spain and Mexico.
  - In English countries positive emotion terms were more discriminating.
  - In Spanish countries positive and negative emotions were more equally discriminating

- *Do we all perceive food-related wellbeing in the same way? Results from an exploratory cross-cultural study,* Gastón Aresa, Ana Giméneza, Leticia Vidal. Food Quality and Preference, Volume 52, September 2016, Pages 62–73
  - Cross-cultural differences in how participants evaluated food-related wellbeing were identified.
  - Participants in the seven countries mostly agreed on their evaluation of physical and intellectual aspects.
  - The largest differences among countries were found for items related to social, spiritual and emotional wellbeing.

Data collection challenges 1

Be careful…2 examples…

ENGLISH

CREAMY

Bulgarian

мек

ENGLISH

soft

ENGLISH

SOUR

Turkish

ekşi

ENGLISH

minus, hard, sharp

Data collection challenges 1

What can you do??

- The essential difference may get lost in translation
  - Use back translation where translation is used!
  - Use of statistical analysis to assess this and see impact
    - Add 'country effect' into statistical models
    - More use of techniques that rely on consumers using own vocabulary and then grouping it statistically (MFA/GPA)
- Also use of scale varies in different 'groups'
  - Review the literature first – research your popln
  - Consider adjusting for scale use in stats analysis

Data collection challenges 2

- Know the regulatory view on product type
- Recent example:
  - Study comparing 2 consumer products – own plus competitor
  - Data collection carried out by a third party (as usual)
  - Company were inspected by regulators as their product had 'therapeutic' properties
  - 'Usual' data collection methods +stats no longer appropriate as not used sufficient standards, although *as per other trials*
  - Company not allowed to use the data without re-databasing it all and re-doing all stats analysis…HUGE costs

…Seek advice/approval BEFORE beginning study

Summary

- Key challenges people re
  - Not enough p
  - Ignorin
  - Quest fo
  - Using the (not kno ve)
  - Measuring the wro
- Issues
- Discussion!

Statistical 'Design' Crucial
- seek advice
- plan to succeed

Statistical Analysis can help
- adequate research first
- statistical analysis may 'solve'

Thank you for listening!

Qi Statistics Ltd.
Penhales House,
Ruscombe Lane
Ruscombe, Reading
RG10 9JN
UK

Phone Numbers

Tel: +44 (0) 118 934 5722
Mob: +44 (0) 7708 700503

E-mail contact

gemma@qistatistics.co.uk