Critical description of Statistical Inference: The case of "Cultural objects modulate reward circuitry".

Author

Carlos Utrilla Guerrero, MSc Manchester University, SRMS.

Supervisor assignment

Prof. Dr Johan Koskinen
Cathie Marsh Institute for Social Research
Manchester University
E-mail: Johan.Koskinen@manchester.ac.uk

Course

Statistical Foundations

Abstract:

The aim of this essay is to describe a neuroscience experimental research in statistical term. More specifically, the essay gives a several idea of (a) basics of statistical inference, (b) understanding the design approach focusing in inference discusses and (c) critical perspective of the methodology applied. Indeed, to successful understanding of the article, ANOVA and fMRI has been analysed as well as the sources of variance (precision), estimators, distribution, hypothesis testing, causal inference, modelling and practical consideration emerged on this research. The structure of the essay starts with a very brief resume of the experimental design. The next section draws the methodology applied and to locate the key points of the methods. Within this section, the conclusion and debate aims to arise the validity and reliability of each method. Subsequently, a summary of principal findings of the article will be the last part of the essay.

A (very) brief resume

Throughout this research, they investigate the reward-related activations on the human brain produced by car photography's, (associated as cultural artificial objects), signalling wealth and superfluity acting as strong social reinforces.

Based on evolutionary perspective, they have examined the differences in the brain regions activated showed 66 cars photographs randomized. Twenty males, participated in the experimental research, carrying out two different analyses: (1) Measure the attractivity rate by five point scale (ANOVA), and (2) fMRI results as baseline and dummy sport car. Given the complexity of the methods applied, I will attempt to synthesize how works this techniques, emphasizing which are the points of strength and weakness. But before this, I would like to highlight six conclusions that reach out theses authors on the article:

c1- *Region points in human brain associated with reward and reinforcement was significantly more activate for sports cars than other car types*.
c2- *The hypothesis of an activation of the reward circuitry by attractive sport cars was confirmed* with the expectation that ventral striatum and orbitofrontal cortex were activated more by sport cars than by small cars.
c3- The degree of attractiveness activates sports cars as first position, then limousines and finally, small cars. This classification would be expected from the intermediate attractivity score. *The mean signal difference in ventral striatum for activation elicited by limousines was lower than for sports cars but higher than for small cars.*
c4- A potentially rewarding stimulus leads to an increase of ventral striatum activation, as reliable predictors of potential social reward and social dominance.
c5- Car are processed in a similar way to face. Headlights of cars are associated as eyes of a human. Rest of the car part, would not produce activation in the ventral striatum.
c5.1- Also might be expected an increasing of fusiform face area activation on expert car brains. A open due is whether activation of fusiform face area is caused by face like appearance of the car or expertise in processing.
c6- *The shape of the car was most influenced factor to their judgments. Lateral occipital complex demonstrate the shape information process.*

The concluded that by comparing neural responses associated with the presentation of sport car vs small cars they can confirmed the hypothesis that sport car are strong social reinforces and would modulate the dopaminergic circuitry because ventral striatum as well as the orbitofrontal cortex were activated more by sports cars than by small cars. They confirmed as well that the degree of attractiveness activates the above mentioned structure, emerge the question of why **the reward circuitry is associated by the degree of attractiveness.** The fact

that one's could consider a potentially rewarding, ventral striatum is activated, make that this attractive car as a predictors of potential social reward is a highly reliable predictor for social dominance and high social rank

Discussion of Methodology & Principal Findings

This systematic review aims to synthesize the key point of methodology used on this research without going into detail upon the techniques.

ANOVA (Attractivity rating).

Behavioral results was interpreted by analysis of variance (ANOVA) elicited if there are any differences between groups of variables. This non parametric technique break up the group according type of car (sport, limousine, small) and then, testing if behavioral outcomes is different across these group of cars. Basically, they have done an analysis of differences between group means and their associated procedures (such as "variation" among and between groups. To clarify, ANOVA have produced a statistical test of whether or not the means of car groups are equal, generating a *t*-test. The result presented on this study split within two different behavioral outcomes: (1) mean attractivity rating and (2) mean reaction times. ANOVA design on this experiment presented the result below:

(1)	$F(2,33= 68.299)\ p < 0.0001$ and
(2)	$F(2,33)=0.69\ p = 0.508$

Further investigation in ANOVA:

In this case study, the objective to show type of car photography's was to measure: (1) attractiveness rating in five-point-scale and (2) reaction time of the cars by button press. Three different types of cars (sport, limousine, small car) were investigated. In this experiment, "Type of car" is the independent variable. Exploring ANOVA design, the term *factor* is used as a synonym of independent variable. Therefore, "Type of car" is the factor in this experiment. Since three types of cars were compared, the factor "Type of car" has three *levels*.

. The null hypothesis tested by ANOVA is that the population (cars) means for all conditions are the same. The analysis of variance (ANOVA) is a F test and can be expressed as follow:

H_0: $\mu_{sport} = \mu_{limousine} = \mu_{smallcar}$

H_a : at least two of the population means are unequal

If the null hypothesis is rejected, then it can be concluded that at least one of the population means is different from at least one other population means.

We should mention the importance to deal with the assumption to consider by ANOVA:

- The populations have the same variance. (*homogeneity of variance).*
- The populations are *normally distributed*.
- Each value is sampled *independently* from each other value generating only one value.

(1) The *probability value* is 0.0001 and therefore the *null hypothesis* is rejected. Hence, the conclusion that at least one of the population means is different from at least one of the others is justified.

(2) Fisher distribution shape shows a small F = 0.69. Moreover, we accept the alternative hypothesis concluding that there are not significant differences among means.

Sample Size

66 car photographs should be the sample size, however ANOVA shows a F distribution shape of F(2,33) elicited from here that there are two degree of freedom (df) parameters for an estimate of variance: one for the numerator(MSB) and one for the dominator (MSE). Furthermore, we can deduce that k (parameters) is equal 3 and n (number of observation) is equal 36. N – k = df, n – 3 = 33 , n = 36?

One of the aims of analysis of variance method is to test the differences among means by analysing variance. The test is based on two estimates of the population variance (σ^2): (a) The Mean Square Error (MSE) based on differences among scores within the groups and secondly, (b) The mean square between-groups (MSB) based on differences among the sample means (See Appendix 1 for more detail of MSE and MSB compute).

Although there are errors of calculation; we might expect the follow deductions:

(1) The fact that F is large enough, *MSB is larger than MSE demonstrating that the population means are not equal*

(2) *Obviously, the interpretation of reaction time table is the opposite. Lower F demonstrate that* the population means are almost equal, being MSE and MSB similar. Even then, MSE will still estimate σ^2 because differences in population means do not affect variances. However, differences in population mean affect MSB since differences among population means are associated with differences among sample means.

Nonetheless, MSB could be larger than MSE by chance even if the population means are equal, MSB must be much larger than MSE in order to justify the conclusion that the population means differ(that's is true according my calculations, see appendix A).

In addition, according the data, the interpretation should be: Between-groups estimate is 68.299 times higher than within-groups estimate. Would this have been likely to happen if all the population means were equal? To clarify what exactly I mean, I should mention that what matter here is upon sample size. With a small sample size as we have, it would expect because small samples are unreliable. However, with a very large sample, the MSB and MSE are almost always about the same, and an F ratio of 3.465 or larger would be very unusual.

In conclusion, to point out some critiques:
- The lack of statistic output about the randomize observation values obtained in the experiment didn't allow us to analyze more on detail the sources of the variation (*sums of squares*, sum of squares error) in order to summarize the portioning of the variance and to figure out the innumerable other reasons why the scores differ. There are possibilities that were not under experimental investigation and therefore all of differences due to these possibilities are unexplained (see online book, sources of variation for discuss).
- n=36 observation of N=66 possibilities as total population. Almost 55% of the total population. Nevertheless, we have not prior information about the way to strata the sample. Simple random sampling with the same probability per photo? Apparently, they made use 12 photos per 3 types of cars. It is large enough to consider the viability and reliability of inference? 50% of N. Which ***Margin Error*** have they assumed precise to keep it with n=36? ***PE +- se x Cα***
- We don't have enough information of the X values in order to analyze the sources of variation.
• Relation F ANOVA $F(1, dfd) = t^2(df)$

fMRI (Funtional Magnetic Resonance Imaging) results:

The comprehensive and powerful tool of fMRI, make this technique heavily to explain but I will to synthesize and emphasizing the key points that are of interest to us. To investigate the function of the brain by fMRI, give high quality visualization of the localization of activity sensory stimulation on the brain.

Before get fMRI outcomes, data must be passed various analysis related with pre-processing, temporal linear modelling and activation thresholding.(Jezzard et al, 2001). It means that it will achieve to remove various kind of artefact in the data (noise), to maximize the sensibility of later statistical analysis and to increase the statistical validity (inference). Indeed, the aim of fMRI analysis is to identify in which vowel the signal of interest is significantly greater than the noise level.

To summarize, fMRI use GLM (General Linear Model) usually focus in functional specialization for every voxel to estimate the size of the experimental effect (Parameter estimates) on which one or more hypothesis (contrast) are tasted to make statistical inferences (p-values) correcting for a multiple comparisons across voxel (using Gaussian Field Theory).

After a pre-processing, statistical analysis is carried out to determine which voxels are activated by the stimulation. Moreover, it uses advanced modelling (hemodynamic response) for different pictures modelled in a single subject analysis for the different car classes. The main output from this step is a statistical map, indicating the points in the image where the brain has activated in the response to the stimulus. Also, they analyse by cluster-based thresholding to achieve final inference.

Specifications:

…about interindividual variance:

[….] By Random Effect Model (between-subject), it use individual statistical contrast, was performed […..] resulting $p < 0.001$ uncorrected for multiple comparisons.

We should interpret testing as expected probability or proportion of false positive for any number of voxels under investigation. In this case, 0,1% of voxels are active when they are not. To conceptualize, interindividual variance take into account between-subject variance can make inferences about the population an Z score, as critical value per voxel obtained in the Random effect analysis, are higher than Z score related with 99,9% confiance level, interpreted in the normal distribution function. In other words, we might assume that with 99,9% of probabilities, we reject the hypothesis null (Activation is Zero everywhere).

Summary statistic approach (SPM):

- 1st level design for all subjects must be the SAME
- Sample means brought forward to 2nd level

Though this technique, they are summarising the response of each subject by a single summary statistic – their effect size (Canonicall hemodynamic function). This random effect inference methods fight against Fixed effect inference methods, which is simplest model but, in contrast, assumes common variance over subject at each voxel. Nevertheless, all the statistical tests assume that the data points are independent samples of the underlying population. This however may not be the case, since the effect of the haemodynamic response function, and any temporal smoothing that is applied in the data, means that adjacent time points are no longer independent. This facts, has an effect on the effective number of degrees of freedom in the sample.

….about globally scaled (standardized size)

Random Field Theory is applied as defining theoretical result for smooth statistical maps and allowing a threshold in a set of data where is not easy to find the number of independent variables, to 'correct' p-values. (a cube of voxels of size FWHM in x, y and z. Depends on smoothness & total # of voxels N=100x100 voxels, Smoothness FWHM= 4.0 s FWHM, Gaussian kernel)

….about threshold corrected

[….] the result [….] in p > 0,05 at the cluster level

To reduce type I error (false positive), they use Correcting for Multiple Comparision

In addition, cluster level Inference, increasing sensitivity by trading off anatomical specificity. Given a voxel level threshold u, we can compute the likelihood (under the null hypothesis) of getting a cluster containing at least n voxels. In typical fMRI data, and supported by previous publications, we can conclude that p value at corrected level is quiet low.

In conclusion, to point out some critiques (fMRI):
• General point of view about data distribution, is that sport car outcomes are more spread out around the mean.
- Do they use principled correction in order to avoid the contamination of false positive proportion?
- Which is the true likelihood of false positives in the results? Did they choose accurate t statistic for inference?
- Interpretations shows of Table I, seems not clean and in accordance with the outcomes concerning Random Effect Analysis. Table I test whether one voxel is false active or not.
- There is confusion about number degree of freedom in order to make much clearer t statistic using for inferencial issues.
- How many voxel are in each clusters?

Finally, I attempt to make a holistic perspective of the article, highlighting the key point of interesting in term of probabilities, statistic, and research design.

Critiques of the Overall article
Nature of probability: Although the frequentistic approach is applied to the methodology, the research material evoke real life scenario. Thus, how they define the type of participant that should be in this experimental research? Why males? Why people that had participated at least once in a car purchase? Why highly interested in cars people? Maybe they have specific brain structure different from others? Subjective probabilistic approach.
How they generate the sampling design? However, they use as observable values different part of the human brains. If you have not specific knowledge about the subject, you are missing information about sample size. How many voxel, n sample, are in the experiment?
Multilevel modelling. From 20 participants to n voxel regions of brain. They didn't explain the causation and effect between them.
Criteria and limitation of classes. Why they made this classification of the cars? Is not more clear the different between luxury car, classic car and 4x4 car? It is just my interpretation thus, its same problem, subjective approach. Therefore, they could make much more differences of brain regions activation if they put extreme photo of cars. When they said, small car, should I consider that there are not sports small cars?
Statistic Outputs: they have showed variance information (not specific) of each method but it's still not clear the relationship between the conclusion and the way to statistic testing. Lack information of degree freedom (n in particular). There is no explanation about the result. Just a few lines on the report
- How many voxel are in each clusters?

BIBLIOGRAPHY:

Fumiko Hoeft (2004): Voxel--Wise Neuroimaging Analysis: Statistics. Stanford University.
http://www.stanford.edu/~fumiko/psyc250_2009.files/Class3_fMRI_statistics_hoeft.pdf

Jezzard P, Matthews P, Smith S, editors. Functional MRI: an introduction to methods. Oxford: OUP, 2001.

Simonoff, Jeffrey S. (1998): Smoothing Methods in Statistics, 2nd edition, Springer

S.M Smith, MA, DPhil (2004): Overview of fMRI analysis. In British Journal of Radiology
http://bjr.birjournals.org/content/77/suppl_2/S167.long

William Penny, Karl Friston et ad. (2007): Statistical Parametric Mapping: The Analysis of Functional Brain Images
http://www.amazon.co.uk/Statistical-Parametric-Mapping-Analysis-Functional/dp/0123725607

Website Resources:

fMRI Methodology:

- J.Ashburner, KFriston and W Penny: Human Brain Function 2nd Edition

http://www.fil.ion.ucl.ac.uk/spm/doc/books/hbf2/

- Statistical Parametric Map: SPM Course: Group Analysis Edinburgh (2010)

http://www.sbirc.ed.ac.uk/cyril/SPM-course/Talks/2010/Daunizeau_GroupAnalysis.pdf

- Functional MRI of the Brain (***FMRIB***)
  http://www.fmrib.ox.ac.uk/

  Inference:

  http://users.fmrib.ox.ac.uk/~stuart/thesis/chapter_6/section6_4.html

- MRC CBSU Wiki (Cognition and Brain Science Unit)

  Principles statistics:

  http://imaging.mrc-cbu.cam.ac.uk/imaging/PrinciplesStatistics#t_statistics_and_contrasts

Unthresholded effect:

http://imaging.mrc-cbu.cam.ac.uk/imaging/UnthresholdedEffectMaps

ANOVA Methodology:

- Online Statistics: An Interactive Multimedia Course of Study

  http://onlinestatbook.com/version_1.html

  http://onlinestatbook.com/lms/analysis_of_variance/one-way.html
- Distributibe. Org

  http://www.distributome.org/js/DistributomeNavigator.html
- Interpreting test statistics, p-values, and significance:

  http://geography.uoregon.edu/geogr/topics/interpstats.htm

APPENDIX A

Computing MSE

This variance, σ^2, is the quantity estimated by MSE and is computed as the mean of the sample variances.

Computing MSB

The formula for **MSB** is based on the fact that the variance of the *sampling distribution* of the mean is

$$\sigma_M^2 = \frac{\sigma^2}{n}$$

where n is the sample size. Rearranging this formula we have

$$\sigma^2 = n\sigma_M^2$$

Although Fisher's original formulation took a slightly different form, the standard method for determining the probability is based on the ratio of MSB to MSE. This ratio is named after Fisher and is called the F ratio.

Since makes the interpretation of attractivity rating is five-point-scale, the shape of the distribution is a one-tailed probability since the probability is the area in the right-hand tail of the distribution.

Sources of Variation

Why do scores in an experiment differ from one another?

An obvious possible reason that the scores could differ is that the subjects were treated differently (they were in different conditions and saw different stimuli). A second reason is that the two subjects may have differed with regard to their tendency to judge cars attractiveness. A third is that, perhaps, one of the subjects was in a bad mood after receiving a low grade on a test. You can imagine that there are innumerable other reasons why the scores of the two

subjects could differ. All of these reasons except the first (subjects were treated differently) are possibilities that were not under experimental investigation and therefore all of differences (variation) due to these possibilities are unexplained. It is traditional to call unexplained variance *error* even though there is no implication that an error was made. Therefore, the variation in this experiment can be thought of as being either variation due to the condition the subject was in or due to error (the sum total of all reasons subjects's scores could differ that were not measured).