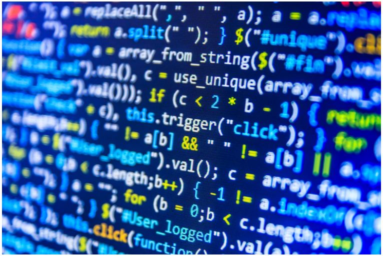
Using randomization in fMRI classification experiments to ensure generalizability

Charles Zheng

Stanford University

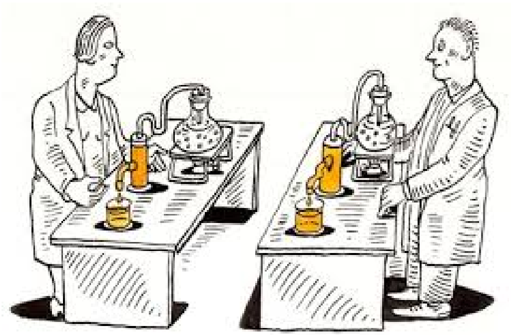
August 3, 2017

(Joint work with Yuval Benjamini.)



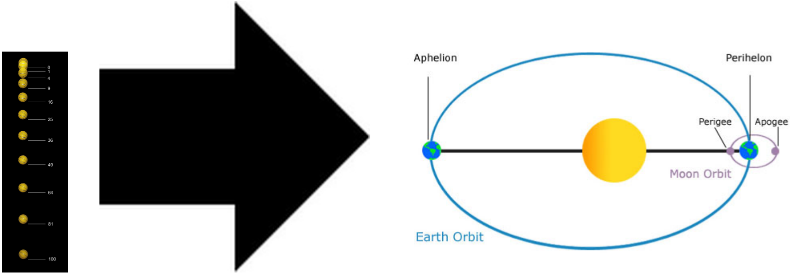
Transparency in sharing data, methods, code, etc.

Replicability



"The ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected"–National Science Foundation

Generalizability



Being able to predict results of new “experiments” or observations.

Problem of Induction

David Hume (1711-1776)

Why is it that "instances of which we have had no experience resemble those of which we have had experience"?

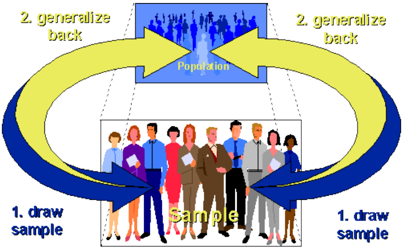
Peirceian Induction and Neyman-Pearson testing

C. S. Pierce

Deborah Mayo

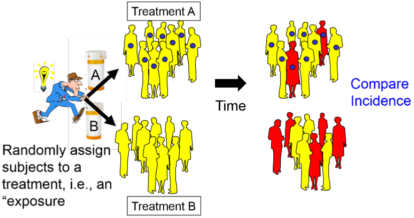
Theories can be confirmed inductively via *severe testing*. The Neyman-Pearson (classical statistical) framework provides one such mechanism.

Generalizing from samples to population



Thanks to key results in probability theory (law of large numbers, central limit theorem), sampling from a defined population is a well-understood form of induction.

Randomized Experiments enable Generalization

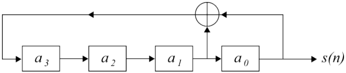


- *Design of Experiments* by R. A. Fisher introduced the concept of *randomization*
- *Randomized clinical trials* are the gold standard for inference of causal effects.
- Randomization + Law of Large Numbers implies quantitative replicability–a form of generalization to the population

Random vs deterministic design in fMRI

For designing event-related sequences for task fMRI...

- Buračas and Boynton (2001) showed that deterministic m-sequences are more efficient for estimating HRF than random designs by a large factor



- However, as Friston (1999) points out, random designs may have advanatages in terms of psychological effects
- Theoretically speaking, deterministic designs are fine as long as one can rule out higher-order dependencies between measurements
- However, when no principled approach exists to cancel out possible biases, randomization guarantees it (on average)

Generalizing beyond the population?

BEHAVIORAL AND BRAIN SCIENCES (2010), Page 1 of 75
doi:10.1017/S0140525X0999152X

The weirdest people in the world?

Joseph Henrich
Department of Psychology and Department of Economics, University of British Columbia, Vancouver V6T 1Z4, Canada
joseph.henrich@gmail.com
http://www.psych.ubc.ca/~henrich/home.html

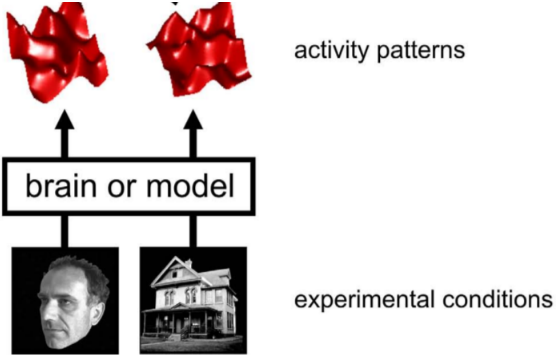
Steven J. Heine
Department of Psychology, University of British Columbia, Vancouver V6T 1Z4, Canada
heine@psych.ubc.ca

Ara Norenzayan
Department of Psychology, University of British Columbia, Vancouver V6T 1Z4, Canada
ara@psych.ubc.ca

Section 2

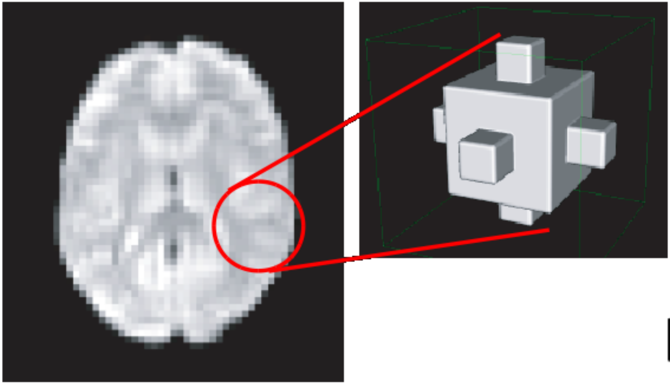
Classification experiments in fMRI

Studying the neural code



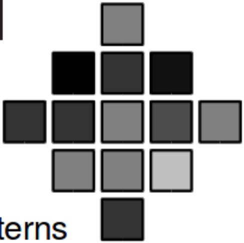
Present the subject with visual stimuli, pictures of faces and houses.
Record the subject's brain activity in the fMRI scanner.

Searchlight analysis



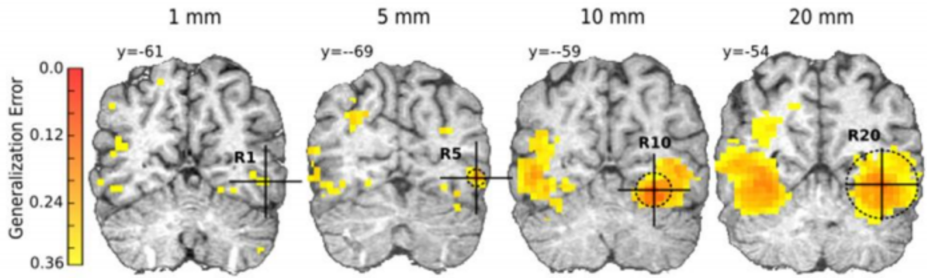
Pull out a local neighbourhood

BOLD image



Look at the patterns in that neighbourhood

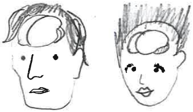
Searchlight analysis



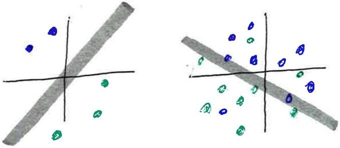
Produces a map of “informative” regions of the brain (as measured by generalization accuracy).

ISSUES W/ TEST ACCURACY

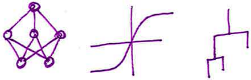
1. Subject dependence



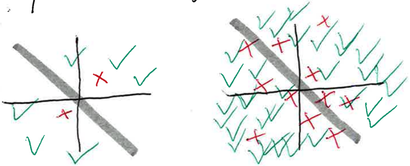
2. Dependence on Training Data



3. Dependence on Classifier

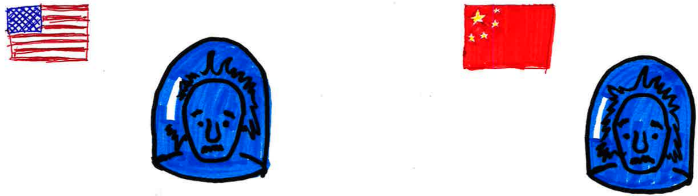


4. Variability due to finite Test Data

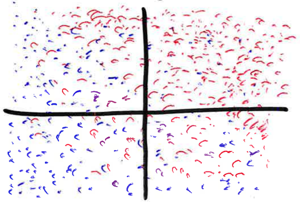


IDEAL WORLD

1. Every lab owns a clone of Einstein



2. Infinite training & test data (⇒ we can obtain Bayes accuracy)



Bayes accuracy

- Discrete $Y \in \{1, ..., k\}$, continuous or discrete X.
- A classifier is a function f mapping x to a label in $\{1, .., k\}$
- Generalization accuracy of the classifier:

$$\text{GA}(f) = \Pr[Y = f(x)]$$

- Bayes accuracy:

$$\text{BA} = \sup_f \Pr[Y = f(x)] = \Pr[Y = \text{argmax}_{i=1} p(X|Y = i)]$$

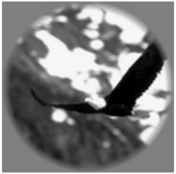
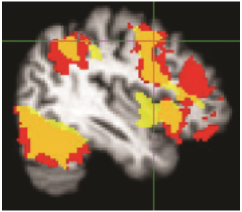
- Since random guessing is correct with probability $1/k$,

$$\text{BA} \in [1/k, 1]$$

(if Y is uniformly distributed)

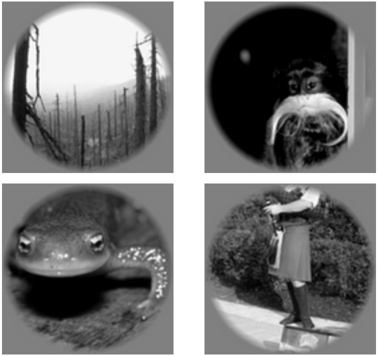
Fixed classification task

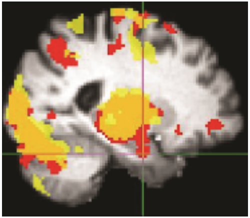


- Different stimuli sets lead to different *Bayes accuracy*.

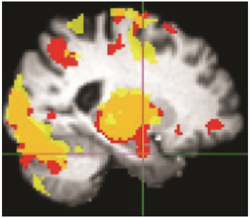
Fixed classification task





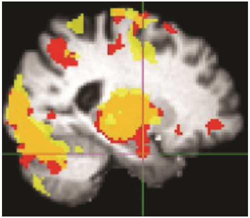
- Different stimuli sets lead to different *Bayes accuracy*.
- Results are incomparable, even in the large-sample limit.

Generalizing beyond the design



Scientists are not innately interested in the Bayes accuracy of a *particular* stimuli set, which is often chosen arbitrarily...

Generalizing beyond the design



But it would be more interesting to be able to make inferences from the data about a *larger* class of stimuli...

Section 3

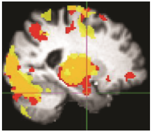
Randomized classification and Average Bayes accuracy

Randomized classification

1. Population of stimuli $p(x)$

2. Subsample k stimuli

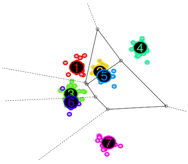
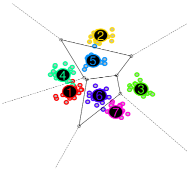
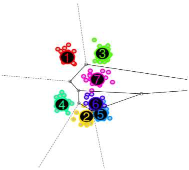
3. Data



4. Train a classifier

5. Estimate generalization accuracy (which is lower bound for the *random* Bayes accuracy BA_k)

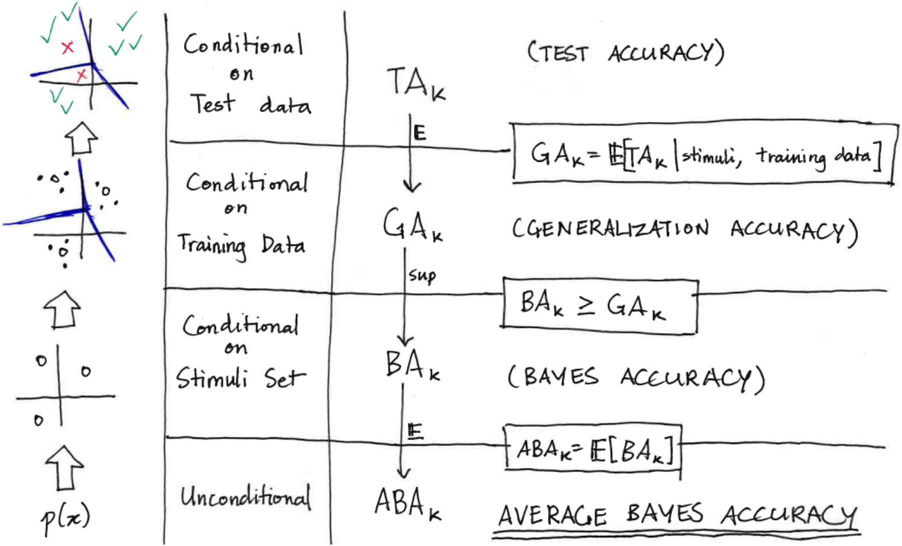
Average Bayes accuracy

	Experiment 1	Experiment 2	Experiment 3
			
Bayes accuracy	0.55	0.65	0.52

- Bayes accuracy depends on the stimuli drawn.
- Therefore, define k-class *average Bayes accuracy* as the expected Bayes accuracy for $X_1, .., X_k \overset{iid}{\sim} p(x)$.

$$\text{ABA}_k = \mathbf{E}[BA(X_1, ..., X_k)]$$

Average Bayes accuracy



Inferring average Bayes accuracy

- $\text{BA}_k \stackrel{def}{=} \text{BA}(X_1, .., X_k)$ is unbiased estimate of

$$\text{ABA}_k = \mathbf{E}[BA_k]$$

  by definition.
- But what is the variance?

$$\text{Var}[\text{BA}(X_1, ..., X_k)]$$

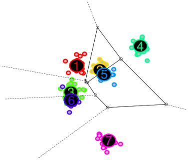
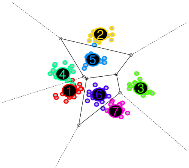
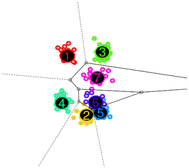
- *Theoretical result.* Maximal variability is of order $1/k$.
- Therefore, it is feasbile to get a good idea of ABA_k by choosing a sufficiently large sample size k.

Two intuitions for variability result

Why does variability decrease with k?

- 1. Bayes accuracy behaves like an average of k i.i.d random variables. (Also gives correct $1/k$ rate.)
- 2. Bayes accuracy behaves like a max of k i.i.d. random variables.

Intuition 1: averaging

	Experiment 1	Experiment 2	Experiment 3
			
Bayes accuracy	0.55	0.65	0.52

Average of k gaussian probability integrals... (which are asympt. uncorrelated.)

Intuition 2: An identity

- It is a well-known result from Bayesian inference that the optimal classifier f is defined as

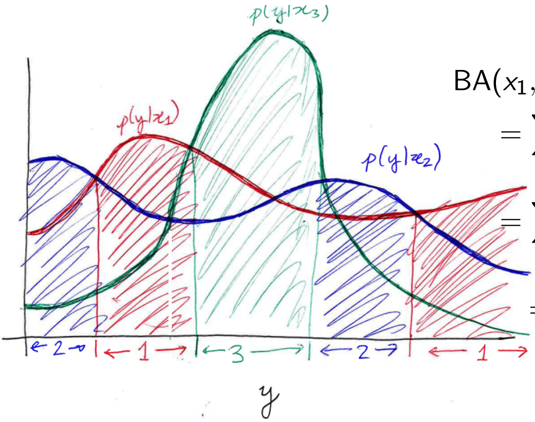
$$f(y) = \text{argmax}_{i=1}^{k} p(y|x_i),$$

since the prior class probabilities are uniform.

- Therefore,

$$\begin{aligned} \text{BA}(x_1, ..., x_k) &= \Pr[\text{argmax}_{i=1}^{k} p(y|x_i) = Z | x_1, ..., x_k] \\ &= \frac{1}{k} \int \max_{i=1}^{k} p(y|x_i) dy. \end{aligned}$$

Intuition behind identity



$$
\begin{aligned}
&\mathrm{BA}(x_1, x_2, x_3)\\
&= \sum_i \Pr[x_i] \Pr_{Y \sim p(y|x_i)}[Y \in \text{zone } i]\\
&= \sum_i \frac{1}{k} \text{Area under curve } i \text{ in zone } i\\
&= \frac{1}{k} \text{Area under } \max_{i=1}^{k} p(y|x_i)
\end{aligned}
$$

Variability of Bayes accuracy

Theoretical result. In the max formulation of BA_k, we can apply Efron-Stein inequality to get

$$\mathrm{sd}[\mathrm{BA}_k] \leq \frac{1}{2\sqrt{k}}$$

Empirical results. (searching for worst-case stimuli).

k	2	3	4	5	6	7	8
$\frac{1}{2\sqrt{k}}$	0.353	0.289	0.250	0.223	0.204	0.189	0.177
Worst-case sd	0.25	0.194	0.167	0.150	0.136	0.126	0.118

Improving the variance bound?

- All of the worst-case distributions take the form

$$\mathcal{Y} = \mathcal{X} = \{1, ..., d\} \text{ for some } d$$

$$p(y|x) = \frac{1}{d} I\{x = y\}$$

- Sampling k items from d with replacement; BA_k is the number of unique items divided by k.
- According to Birthday paradox,

$$\text{ABA}_k \approx (1 - e^{-d/k})$$

and

$$\text{Var}(\text{BA}_k) \approx \frac{1}{d} e^{-d/k} (1 - e^{-d/k})$$

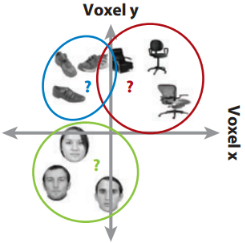
- "Discreteness" of the distribution seems to maximize variance?
- If we could prove that this is indeed the worst case, then we have a better constant for variance bound.

Inferring average Bayes error

For now, return to the world of finite data...

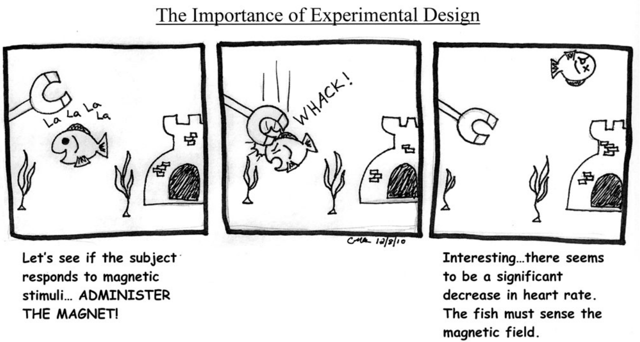
1. *Experimental design*: draw k stimuli $X_1, ..., X_k$ iid from $p(x)$. Then collect data (X_i, Y_i^j).
2. *Supervised learning*: train a classifier and obtain a test accuracy TA_k.
3. *Generalization accuracy*: if n_{test} is the size of the test set,
$$\underline{\text{GA}_k} = \text{TA}_k - \frac{z_{\alpha/2}\sqrt{\text{TA}_k(1-\text{TA}_k)}}{\sqrt{n_{test}}}$$
is a lower confidence bound for GA_k
4. *Bayes accuracy*:
$$\underline{\text{BA}_k} = \underline{\text{GA}_k}$$
is a lower confidence bound for BA_k
5. *Average Bayes accuracy*
$$\underline{\text{ABA}_k} = \underline{\text{BA}_k} - \frac{1}{2\sqrt{\alpha k}}$$
is a lower confidence bound for ABA_k.

Future work



- Theory can be extended to handle discrimination between a fixed number of categories
- Category-based classification is equivalent to a cost function $C(y, y')$ which is equal to 0 if y and y' are from the same category, and 1 otherwise.
- Sampling of random exemplars is stratified by category, but amounts to a minor adjustment to the variance bounds

The end



(credit C. Ambrosino)