Using randomization in fMRI classification experiments to ensure generalizability

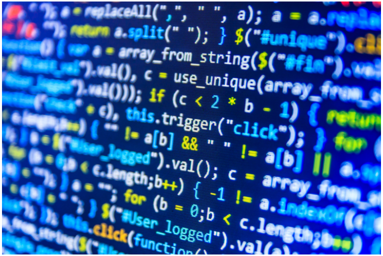
Charles Zheng

National Institute of Mental Health

August 4, 2017

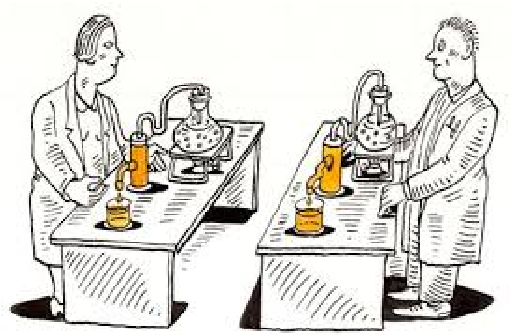
(Joint work with Yuval Benjamini.)

Reproducibility



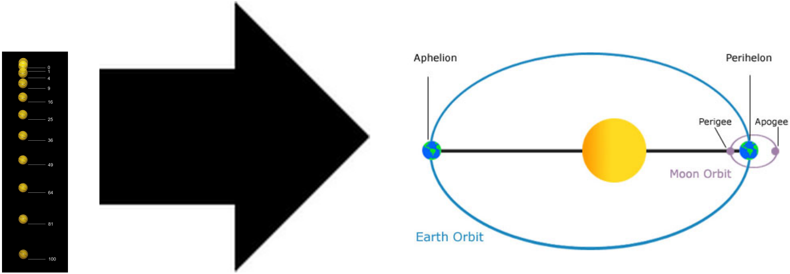
Transparency in sharing data, methods, code, etc.

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"?

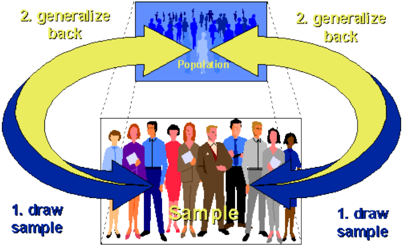
Peirceian Induction and Neyman-Pearson testing

C. S. Pierce

Deborah Mayo

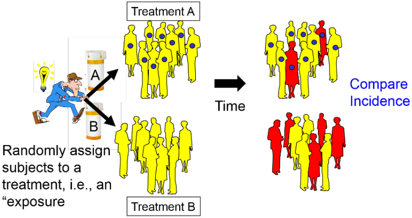
Theories can be confirmed inductively via *severe testing*. The Neyman-Pearson (classical statistical) framework provides one such mechanism.

Generalizing from samples to population



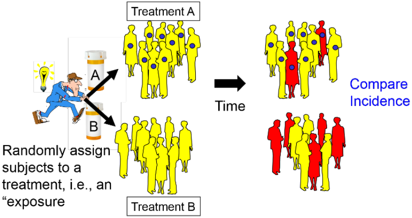
Thanks to key results in probability theory (law of large numbers, central limit theorem), sampling from a defined population is a well-understood form of induction.

Randomized Experiments enable Generalization



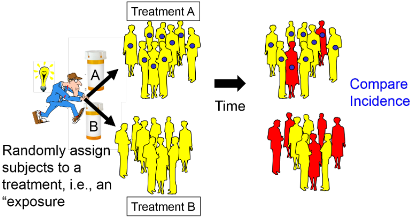
- *Design of Experiments* by R. A. Fisher introduced the concept of *randomization*

Randomized 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.

Randomized Experiments enable Generalization

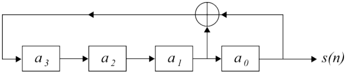


- *Design of Experiments* by R. A. Fisher introduced the concept of *randomization*
- *Randomized clinical trials* are the gold standard for inference of causal effects.
- Randomization + Law of Large Numbers implies quantitative replicability–a form of generalization to the population

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

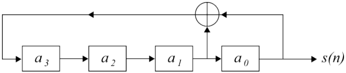


- However, as Friston (1999) points out, random designs may have advanatages in terms of psychological effects

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

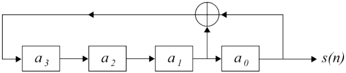


- However, as Friston (1999) points out, random designs may have advanatages in terms of psychological effects
- Theoretically speaking, deterministic designs are fine as long as one can rule out higher-order dependencies between measurements

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

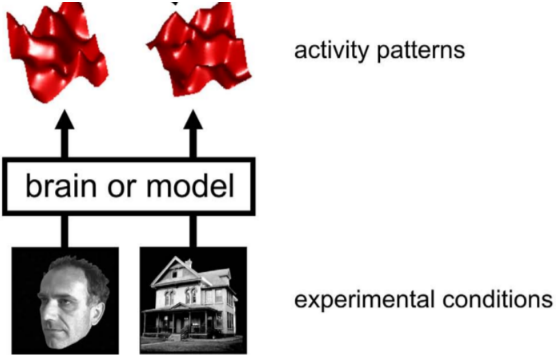
Steven J. Heine
Department of Psychology, University of British Columbia, Vancouver V6T 1Z4, Canada
heine@psych.ubc.ca

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

Section 2

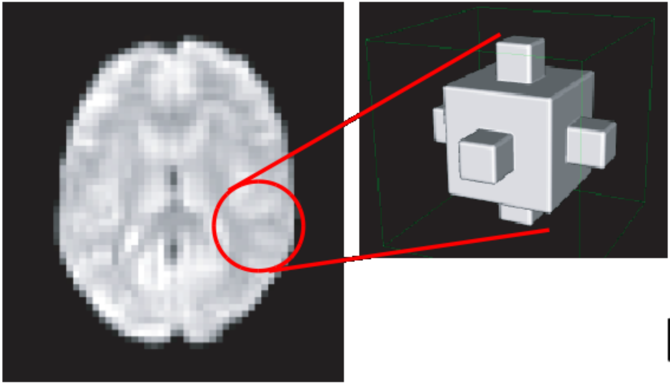
Classification experiments in fMRI

Studying the neural code



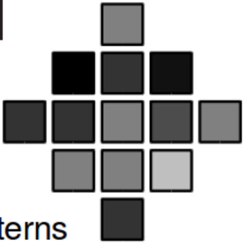
Present the subject with visual stimuli, pictures of faces and houses.
Record the subject's brain activity in the fMRI scanner.

Searchlight analysis

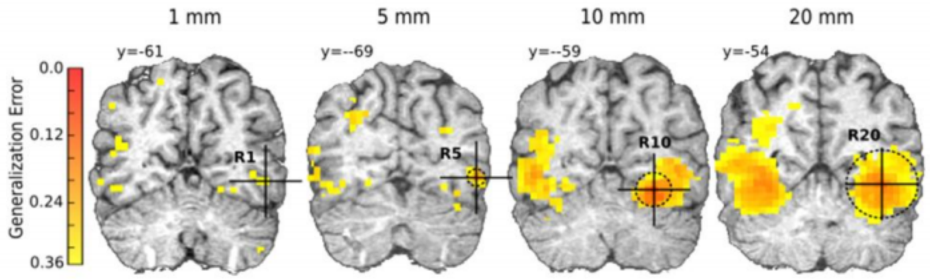


Pull out a local neighbourhood

BOLD image



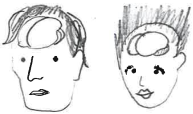
Look at the patterns in that neighbourhood



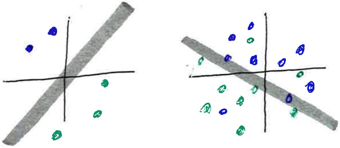
Produces a map of "informative" regions of the brain (as measured by generalization accuracy).

ISSUES W/ TEST ACCURACY

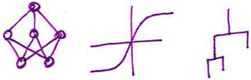
1. Subject dependence



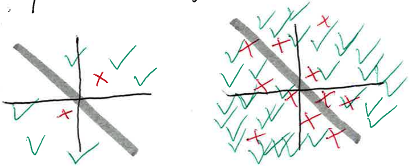
2. Dependence on Training Data



3. Dependence on Classifier



4. Variability due to finite Test Data



Bayes accuracy

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

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

- 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:

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

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:

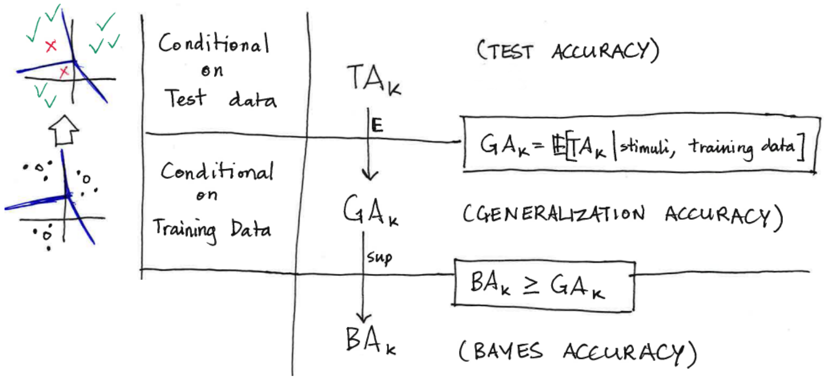
$$\text{BA} = \sup_f \Pr[Y = f(x)] = \Pr[Y = \text{argmax}_{i=1} p(X|Y = i)]$$

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

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

(if Y is uniformly distributed)

Inferring Bayes accuracy

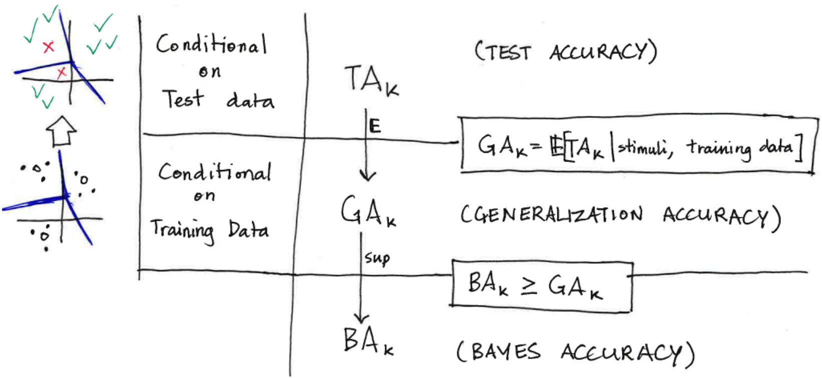


- Given m test observations,

$$\underline{\text{GA}}_\alpha(\hat{f}) = \text{TA} - z_\alpha \sqrt{\frac{\text{TA}(1-\text{TA})}{m}}$$

is a an $(1-\alpha)$ lower confidence bound for BA.

Inferring Bayes accuracy

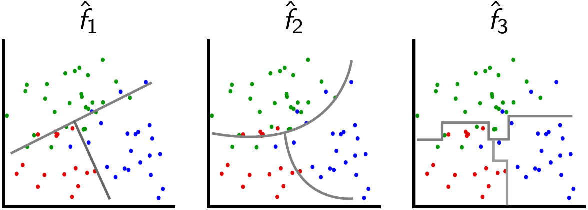


- Since BA $\geq$ GA by definition,

$$\underline{\mathrm{BA}}_\alpha = \underline{\mathrm{GA}}(\hat{f})$$

is an $(1-\alpha)$ lower confidence bound for BA.

Inferring Bayes accuracy under model selection



- Or, if $\hat{f}_1, \ldots, \hat{f}_d$ result from d different procedures,

$$\underline{\mathrm{BA}}_\alpha = \min_{i=1}^{d} \underline{\mathrm{GA}}_{\frac{\alpha}{d}}(\hat{f}_i)$$

  is also an $(1-\alpha)$ lower confidence bound for BA (using Bonferroni's inequality).

Can we get an *upper bound* for Bayes accuracy?

- Mathematically speaking, no, since for all we know there could be a *super-complicated* classification rule (that is impossible to learn from data) that gets 100 percent accuracy.

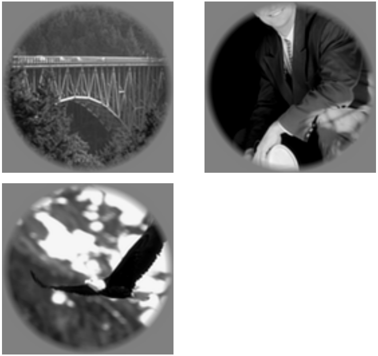
Can we get an *upper bound* for Bayes accuracy?

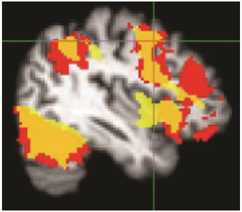
- Mathematically speaking, no, since for all we know there could be a *super-complicated* classification rule (that is impossible to learn from data) that gets 100 percent accuracy.
- However, if we can make some kind of smoothness assumption on the Bayes boundary, it might be possible

Can we get an *upper bound* for Bayes accuracy?

- Mathematically speaking, no, since for all we know there could be a *super-complicated* classification rule (that is impossible to learn from data) that gets 100 percent accuracy.
- However, if we can make some kind of smoothness assumption on the Bayes boundary, it might be possible
- Some relevant work (Cortes et al 1994) but this is a wide-open problem in machine learning

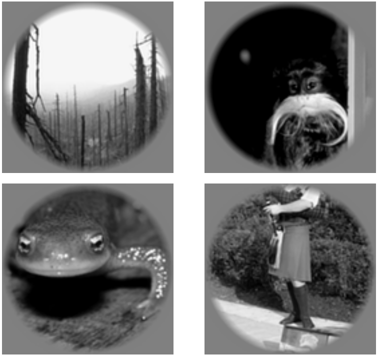
Problem with Bayes accuracy

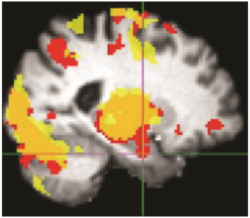




- Different stimuli sets lead to different *Bayes accuracy*.

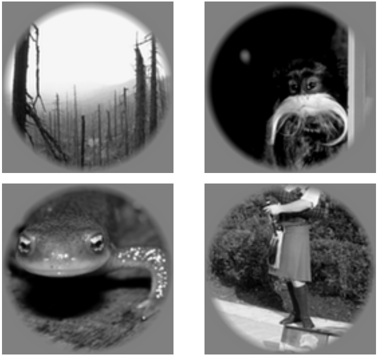
Problem with Bayes accuracy

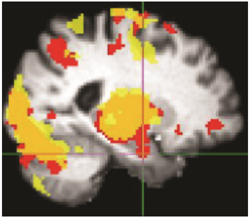




- Different stimuli sets lead to different *Bayes accuracy*.

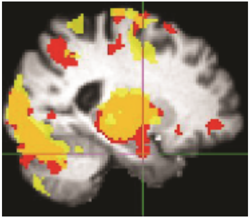
Problem with Bayes accuracy





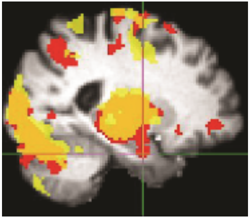
- Different stimuli sets lead to different *Bayes accuracy*.
- Results are incomparable, even in the large-sample limit.

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

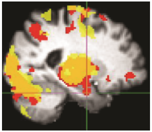
Randomized classification and Average Bayes accuracy

Randomized classification

1. Population of stimuli $p(x)$

2. Subsample k stimuli

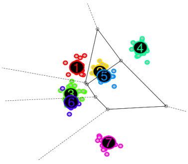
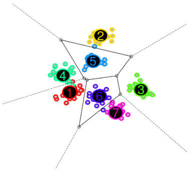
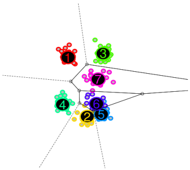
3. Data



4. Train a classifier

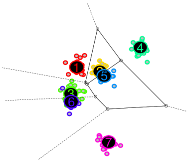
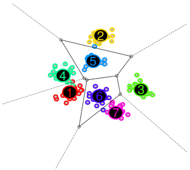
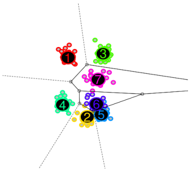
5. Estimate generalization accuracy (which is lower bound for the *random* Bayes accuracy BA_k)

Average Bayes accuracy

	Experiment 1	Experiment 2	Experiment 3
			
Bayes accuracy	0.55	0.65	0.52

- Bayes accuracy depends on the stimuli drawn.

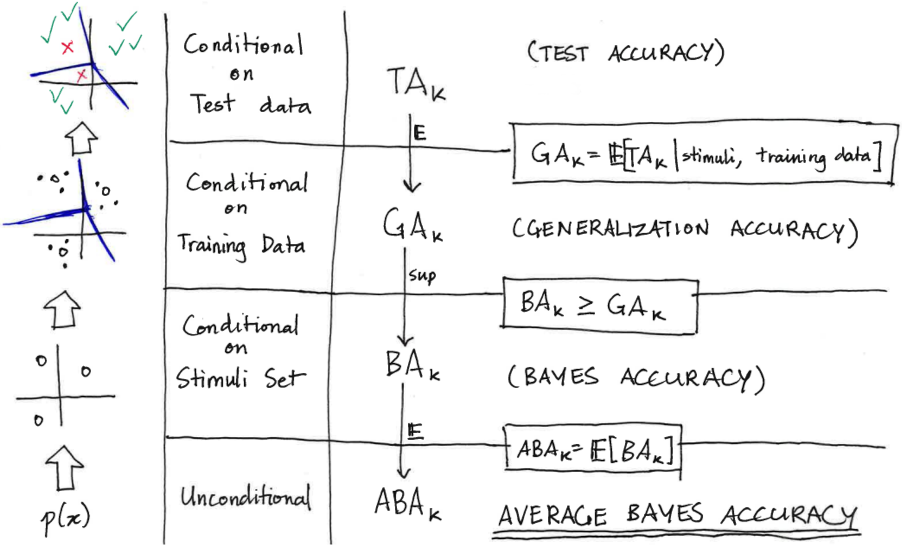
Average Bayes accuracy

	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 $X_1, .., X_k \stackrel{iid}{\sim} p(x)$.

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

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.

Inferring average Bayes accuracy

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

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

  by definition.

- But what is the variance?

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

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 $1/k$.

Inferring average Bayes accuracy

- $\text{BA}_k \overset{def}{=} \text{BA}(X_1, .., X_k)$ 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.

Variability of Bayes accuracy

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

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

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

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).

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.

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

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

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.

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*

$$\underline{\text{ABA}_k} = \underline{\text{BA}_k} - \frac{1}{2\sqrt{\alpha k}}$$

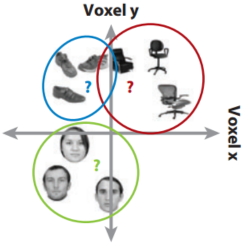
   is a lower confidence bound for ABA_k.

Back to fMRI experimental design...

How should one select the tasks for an experiment?

Design strategy	**Pros**	**Cons**
Arbitrary	Convenient Could be more engaging for subject (e.g. using a movie)	Could be biased
Systematic	Efficient Could be standardized (and enable inter-subject comparison)	Might not be representative of ``typical" performance Could be biased Needs special theory to prevent bias
Random	Generalizes to population Controls bias Facilitates inference	Need to decide what the population is Need sufficient number of random samples

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

Conclusions

- Classification accuracy is being used for a variety of downstream inferences and interpretations in fMRI

Conclusions

- Classification accuracy is being used for a variety of downstream inferences and interpretations in fMRI
- Test accuracy is hard to interpret for a variety of reasons

Conclusions

- Classification accuracy is being used for a variety of downstream inferences and interpretations in fMRI
- Test accuracy is hard to interpret for a variety of reasons
- Using test accuracy as a means of *lower-bounding* Bayes accuracy, we can make rigorous inferential statements, and this is more honest about what classification really tells us

Conclusions

- Classification accuracy is being used for a variety of downstream inferences and interpretations in fMRI
- Test accuracy is hard to interpret for a variety of reasons
- Using test accuracy as a means of *lower-bounding* Bayes accuracy, we can make rigorous inferential statements, and this is more honest about what classification really tells us
- It would be nice if we could also upper-bound Bayes accuracy, but more theory is needed.

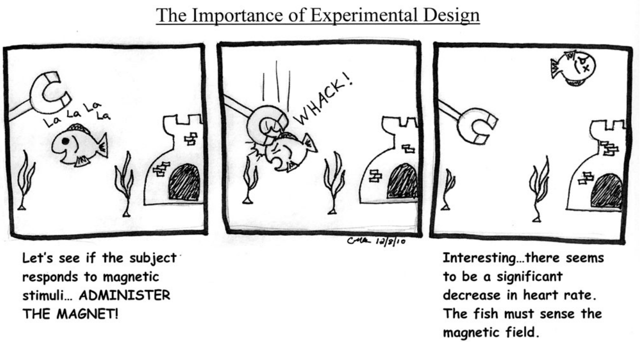
Conclusions

- Classification accuracy is being used for a variety of downstream inferences and interpretations in fMRI
- Test accuracy is hard to interpret for a variety of reasons
- Using test accuracy as a means of *lower-bounding* Bayes accuracy, we can make rigorous inferential statements, and this is more honest about what classification really tells us
- It would be nice if we could also upper-bound Bayes accuracy, but more theory is needed.
- Bayes accuracy, however, does not necessarily generalize beyond an arbitrary stimulus set.

Conclusions

- Classification accuracy is being used for a variety of downstream inferences and interpretations in fMRI
- Test accuracy is hard to interpret for a variety of reasons
- Using test accuracy as a means of *lower-bounding* Bayes accuracy, we can make rigorous inferential statements, and this is more honest about what classification really tells us
- It would be nice if we could also upper-bound Bayes accuracy, but more theory is needed.
- Bayes accuracy, however, does not necessarily generalize beyond an arbitrary stimulus set.
- One way to make sure it generalizes to a population is to use a sufficiently large number of random samples, and our theory tells us how many are needed for a given level of replicability

The end



(credit C. Ambrosino)